CLUSTER-ROBUST BOOTSTRAP INFERENCE IN QUANTILE REGRESSION MODELS

Andreas Hagemann*

July 3, 2015

In this paper I develop a wild bootstrap procedure for cluster-robust inference in linear quantile regression models. I show that the bootstrap leads to asymptotically valid inference on the entire quantile regression process in a setting with a large number of small, heterogeneous clusters and provides consistent estimates of the asymptotic covariance function of that process. The proposed bootstrap procedure is easy to implement and performs well even when the number of clusters is much smaller than the sample size. An application to Project STAR data is provided.

JEL classification: C01, C15, C21
Keywords: quantile regression, cluster-robust standard errors, bootstrap

1. Introduction

It is common practice in economics and statistics to conduct inference that is robust to within-cluster dependence. Examples of such clusters are households, classrooms, firms, cities, or counties. We have to expect that units within these clusters influence one another or are influenced by the same sociological, technical, political, or environmental shocks. To account for the presence of data clusters, the literature frequently recommends inference using cluster-robust versions of the bootstrap; see, among many others, Bertrand, Duflo, and Mullainathan (2004) and Cameron, Gelbach, and Miller (2008) for an overview in the context of linear regression models estimated by least squares. In this paper I develop a bootstrap method for cluster-robust inference in linear quantile regression (QR) models. The method, which I refer to as *wild gradient bootstrap*, is an extension of a wild bootstrap procedure proposed by Chen, Wei, and Parzen (2003). Despite the fact that

*Department of Economics, University of Michigan, 611 Tappan Ave, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 764-2355. Fax: +1 (734) 764-2769. E-mail: hagem@umich.edu. I would like to thank Roger Koenker for his continued support and encouragement. I would also like to thank the co-editor, an associate editor, two referees, Matias Cattaneo, Sílvia Gonçalves, Carlos Lamarche, Sarah Miller, Stephen Portnoy, João Santos Silva, Ke-Li Xu, and Zhou Zhou for comments and discussions. All errors are my own.

it involves resampling the QR gradient process, the wild gradient bootstrap is fast and easy to implement because it does not involve finding zeros of the gradient during each bootstrap iteration. I show that the wild gradient bootstrap allows for the construction of asymptotically valid bootstrap standard errors, hypothesis tests both at individual quantiles or over ranges of quantiles, and confidence bands for the QR coefficient function.

Quantile regression, introduced by Koenker and Bassett (1978), has become an important empirical tool because it enables the researcher to quantify the effect of a set of covariates on the entire conditional distribution of the outcome of interest. This is in sharp contrast to conventional mean regression methods, where only the conditional mean can be considered. A disadvantage of QR in comparison to least squares methods is that the asymptotic variance of the QR coefficient function is notoriously difficult to estimate due to its dependence on the unknown conditional density of the response variable. An analytical estimate of the asymptotic variance therefore requires a user-chosen kernel and bandwidth. Hence, two researchers working with the same data can arrive at different conclusions simply because they used different kernels or bandwidths. Furthermore, a common concern in applied work is that analytical estimates of asymptotic variances perform poorly in the cluster context when the number of clusters is small or the within-cluster correlation is high; see, e.g., Bertrand et al. (2004) and Webb (2014) for extensive Monte Carlo evidence in the least squares case. Their overall finding is that true null hypotheses are rejected far too often. Simulations in MacKinnon and Webb (2015) suggest that similar problems also occur when the cluster sizes differ substantially.

I show that the wild gradient bootstrap is robust to each of these concerns: It performs well even when the number of clusters is small, the within-cluster dependence is high, and the cluster sizes are heterogenous. The wild gradient bootstrap consistently estimates the asymptotic distribution and covariance functions of the QR coefficients without relying on kernels, bandwidths, or other user-chosen parameters. As such, this paper complements recent work by Parente and Santos Silva (2015), who provide analytical, kernel-based covariance matrix estimates for QR with cluster data. Their estimates have the advantage that they are simpler to compute than the bootstrap procedures presented here. However, as the Monte Carlo study in this paper shows, a drawback is that tests based on their covariance matrix estimates can suffer from severe size distortions in the same situations as those described for the mean regression case above. In addition, Parente and Santos Silva's method does not allow for uniform inference across quantiles because the limiting QR process generally has an analytically intractable distribution. In contrast, the bootstrap approximations of the distribution and covariance functions developed here can be combined to perform uniform Wald-type inference about the QR coefficient function.

The wild bootstrap procedure discussed in this paper was first introduced by Chen et al. (2003) as a way to construct confidence intervals for QR coefficients at a single quantile. However, they only provide heuristic arguments for the consistency of the bootstrap approximation and note that "as far as [they] know, there is no analytical proof that the bootstrap method is valid for the general quantile regression model with correlated

observations." I considerably extend the scope of their method under explicit regularity conditions to allow for inference on the entire QR process and uniformly consistent covariance matrix estimates of that process. Some parts of the proofs below rely on a recent result by Kato (2011) regarding the convergence of bootstrap moments. In turn, his results build on a technique developed in Alexander (1985) and tail bounds for the empirical process given in van der Vaart and Wellner (1996). I also utilize empirical process results of Pollard (1990) and Kosorok (2003) to address some nonstandard issues arising from the fact that I allow clusters to be heterogeneous both in terms of their size and their distributions.

Other types of wild bootstrap procedures for QR are given in Feng, He, and Hu (2011) and Davidson (2012). They do not deal with cluster data but their methods are likely to generalize in this direction. Although typically only convergence of the bootstrap distribution is shown, these and other bootstrap methods have been suggested in the literature as ways to construct bootstrap standard errors or, more generally, bootstrap covariance matrix estimates. Hahn (1995) and Gonçalves and White (2005) explicitly caution against such conclusions because convergence in distribution does not imply convergence of moments without uniform integrability conditions. This paper establishes these conditions for QR estimates in the cluster context. As I show in my Monte Carlo study, the availability of a bootstrap covariance matrix estimate is crucial for bootstrap tests to have good size and power in many empirically relevant situations.

Cluster-robust inference in linear regression has a long history in economics; see Cameron and Miller (2015) for a recent survey. Kloek (1981) is an early reference. Moulton (1990) highlights the importance of correcting inference for within-cluster correlation when covariates do not vary within clusters. However, apart from Chen et al. (2003) and Parente and Santos Silva (2015), cluster inference in QR models has not received much attention. Notable exceptions are Wang and He (2007) and Wang (2009), who develop methods for cluster-robust quantile rank score inference, Chetverikov, Larsen, and Palmer (2013), who introduce a method for instrumental variables estimation in a QR model with cluster-level treatment, and Yoon and Galvao (2013), who discuss QR in a panel model where clusters arise from correlation of individual units over time.

The paper is organized as follows: Section 2 states and discusses several assumptions that are then used to establish the large sample distribution of the QR estimator with cluster data. Section 3 introduces the wild gradient bootstrap procedure and shows how it can be applied to conduct bootstrap inference on the QR process. Section 4 illustrates the finite-sample behavior of the wild gradient bootstrap in three Monte Carlo experiments. Section 5 contains a brief application of the proposed bootstrap procedure to Project STAR data. Section 6 concludes. The appendix contains auxiliary results and proofs.

I use the following notation throughout the paper: $|\cdot|$ is Euclidean norm and $1\{\cdot\}$ is the indicator function. Limits are as $n \to \infty$ unless otherwise noted and convergence in distribution is indicated by $\rightsquigarrow$.

2. Quantile Regression with Clustered Data

This section discusses linear QR with clustered data and outlines the basic assumptions used to justify asymptotic inference. Then I establish weak convergence of the QR estimator in the cluster context.

Regression quantiles in a framework with cluster data express the conditional quantiles of a response variable Y_{ik} in terms of an observed covariate vector X_{ik}. Here i indexes clusters and k indexes observations within that cluster. There are n clusters and cluster i has c_i observations. The cluster sizes c_i need not be identical across i, but will be taken to be small relative to n for the asymptotic theory. Because there is typically no natural ordering of observations in the same cluster (unless k indexes time), I allow for arbitrary within-cluster dependence of the data.

2.1 Assumption. *For all $i, j \geq 1$ with $i \neq j$ and all $1 \leq k \leq c_i$, $1 \leq l \leq c_j$, the random vectors $(Y_{ik}, X_{ik}^\top)^\top$ and $(Y_{jl}, X_{jl}^\top)^\top$ are independent. The cluster sizes c_i are bounded by some $c_{\max} < \infty$ uniformly in $i \geq 1$.*

The τth quantile function of Y_{ik} conditional on $X_{ik} = x$ is given by $Q_{ik}(\tau \mid x) := \inf\{y : \mathrm{P}(Y_{ik} \leq y \mid X_{ik} = x) \geq \tau\}$, where $\tau \in (0,1)$. I assume that the linear QR framework is an appropriate model for the data.

2.2 Assumption. *For $\{Y_{ik} : i \geq 1, 1 \leq k \leq c_i\}$, the τth quantile function satisfies*

$$Q_{ik}(\tau \mid x) = x^\top \beta(\tau), \qquad x \in \mathcal{X}_{ik} \subset \mathbb{R}^d,\ \tau \in T,$$

where $\mathcal{X}_{ik}$ is the support of X_{ik} and T is a closed subset of $(0,1)$. For all $\tau \in T$, $\beta(\tau)$ is contained in the interior of a compact and convex set $B \subset \mathbb{R}^d$.

Remarks. (i) Because $\tau \mapsto \beta(\tau)$ does not depend on i, this assumption implicitly rules out cluster-level "fixed effects" as they would lead to incidental parameter problems; see Koenker (2004). It does *not* rule out covariates that vary at the cluster level and, more importantly, fixed effects for levels above the cluster level. For example, the application in Section 5 has classroom-level clusters and school-level fixed effects. There are several ways to address the incidental parameters problem when more is known about the dependence structure in the data; see Yoon and Galvao (2013) and the references therein.

(ii) The assumption of compactness of B has no impact on the estimation of the QR model in practice because B can always be viewed as large. Compactness is, however, essential for the validity of bootstrap moment estimates in the QR context; see the discussion below Theorem 3.3 in the next section.

Estimates of the unknown QR coefficient function $\tau \mapsto \beta(\tau)$ can be computed with the help of the Koenker and Bassett (1978) check function $\rho_\tau(z) = (\tau - 1\{z < 0\})z$. For clustered data, the QR problem minimizes

$$\mathbb{M}_n(\beta, \tau) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \rho_\tau(Y_{ik} - X_{ik}^\top \beta)$$

with respect to β for a given τ so that $\tau \mapsto \beta(\tau)$ is estimated by

$$\tau \mapsto \hat{\beta}_n(\tau) := \arg\min_{\beta \in B} \mathbb{M}_n(\beta, \tau), \qquad \tau \in T.$$

The main goal of this paper is to provide a method for cluster-robust bootstrap inference about the QR coefficient function that is valid uniformly on the entire set T and leads to cluster-robust bootstrap covariance matrix estimates for $\tau \mapsto \hat{\beta}_n(\tau)$. The validity of this method relies on the asymptotic normal approximation to the distribution of $\sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau))$, which I turn to next.

Asymptotic normality of the QR estimates in the cluster context requires straightforward extensions of smoothness and moment conditions familiar from the iid case. The following assumption allows for arbitrary heterogeneity of the clusters as long as the observations within these clusters satisfy mild restrictions on the smoothness of their conditional distributions and the tail behavior of the covariates. This assumption is needed to ensure identification of the QR coefficient function and to justify an approximation argument following immediately below.

2.3 Assumption. (i) $\mathrm{E}\,|X_{ik}|^q < \infty$ *uniformly in* $i \geq 1$ *and* $1 \leq k \leq c_i$ *for some* $q > 2$,
(ii) $n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \mathrm{E}\, X_{ik} X_{ik}^\top$ *is positive definite, uniformly in* n,
(iii) *the conditional density* $f_{ik}(y \mid X_{ik} = x)$ *of* Y_{ik} *and its derivative in* y *are bounded above uniformly in* y *and* $x \in \mathcal{X}_{ik}$, *uniformly in* $i \geq 1$ *and* $1 \leq k \leq c_i$, *and*
(iv) $f_{ik}(x^\top \beta \mid X_{ik} = x)$ *is bounded away from zero uniformly in* $\beta \in B$ *and* $x \in \mathcal{X}_{ik}$, *uniformly in* $i \geq 1$ *and* $1 \leq k \leq c_i$.

To establish distributional convergence of the QR estimator, I consider the recentered population objective function $\beta \mapsto M_n(\beta, \tau) := \mathrm{E}(\mathbb{M}_n(\beta, \tau) - \mathbb{M}_n(\beta(\tau), \tau))$. The recentering ensures that M_n is well defined without moment conditions on the response variable. Provided Assumptions 2.2 and 2.3 hold, the map $\beta \mapsto M_n(\beta, \tau)$ is differentiable with derivative $M_n'(\beta, \tau) := \partial M_n(\beta, \tau)/\partial \beta^\top$ and achieves its minimum at $\beta(\tau)$ by convexity. I show in the appendix that under Assumptions 2.1-2.3 we can write

$$0 = \sqrt{n} M_n'(\beta(\tau), \tau) = \sqrt{n} M_n'(\hat{\beta}_n(\tau), \tau) - J_n(\tau) \sqrt{n}\big(\hat{\beta}_n(\tau) - \beta(\tau)\big) + o_{\mathrm{P}}(1) \tag{2.1}$$

uniformly in $\tau \in T$ by a mean value expansion about $\hat{\beta}_n(\tau)$. Here $J_n(\tau)$ is the Jacobian matrix of the expansion evaluated at $\beta(\tau)$,

$$J_n(\tau) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \mathrm{E}\, f_{ik}\big(X_{ik}^\top \beta(\tau) \mid X_{ik}\big) X_{ik} X_{ik}^\top.$$

After rearranging (2.1), a stochastic equicontinuity argument (see the appendix for details) can be used to show that $\sqrt{n} M_n'(\hat{\beta}_n(\tau), \tau)$ is, uniformly in $\tau \in T$, within $o_{\mathrm{P}}(1)$ of the first term on the right of

$$J_n(\tau) \sqrt{n}\big(\hat{\beta}_n(\tau) - \beta(\tau)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \psi_\tau\big(Y_{ik} - X_{ik}^\top \beta(\tau)\big) X_{ik} + o_{\mathrm{P}}(1), \tag{2.2}$$

where $\psi_\tau(z) = \tau - 1\{z < 0\}$. The outer sum on the right-hand side can be viewed as an empirical process evaluated at functions of the form $\sum_{k=1}^{c_i} \psi_\tau(Y_{ik} - X_{ik}^\top \beta(\tau))X_{ik}$ indexed by $\tau \in T$.[1] This empirical process has covariances

$$\Sigma_n(\tau, \tau') := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \sum_{l=1}^{c_i} \mathrm{E}\, \psi_\tau\big(Y_{ik} - X_{ik}^\top \beta(\tau)\big) \psi_{\tau'}\big(Y_{il} - X_{il}^\top \beta(\tau')\big) X_{ik} X_{il}^\top, \qquad \tau, \tau' \in T.$$

As a referee points out, similar covariances arise in the generalized estimating equations framework of Liang and Zeger (1986).

In the absence of clusters (i.e., $c_i \equiv 1$), $\Sigma_n(\tau, \tau')$ reduces to the familiar form of $n^{-1} \sum_{i=1}^n \mathrm{E}\, X_{i1} X_{i1}^\top$ times the covariance function $(\min\{\tau, \tau'\} - \tau\tau')I_d$ of the standard d-dimensional Brownian bridge, where I_d is the identity matrix of size d. Because of the within-cluster dependence, the structure of $\Sigma_n(\tau, \tau')$ is now significantly more involved. This does not change in the limit as $n \to \infty$, which is assumed to exist along with the limit of the Jacobian.

2.4 Assumption. *$J(\tau) = \lim_{n\to\infty} J_n(\tau)$, $\Sigma(\tau, \tau') = \lim_{n\to\infty} \Sigma_n(\tau, \tau')$ exist for $\tau, \tau' \in T$.*

Remark. I show in the appendix that under Assumptions 2.1-2.3 the pointwise convergence in Assumption 2.4 already implies uniform convergence of J_n and Σ_n.

The matrix limit $J(\tau)$ is positive definite by Assumption 2.3. Hence, the asymptotic distribution of $\sqrt{n}(\hat\beta_n(\tau) - \beta(\tau))$ can be determined via the continuous mapping theorem and an application of a central limit theorem to the right-hand side of (2.2). The following theorem confirms that this even remains valid when $\tau \mapsto \sqrt{n}(\hat\beta_n(\tau) - \beta(\tau))$ is viewed as a random process indexed by T. The distributional convergence then occurs relative to $\ell^\infty(T)^d$, the class of uniformly bounded functions on T with values in $\mathbb{R}^d$.

2.5 Theorem. *Suppose Assumptions 2.1-2.4 hold. Then $\{\sqrt{n}(\hat\beta_n(\tau) - \beta(\tau)) : \tau \in T\}$ converges in distribution to a mean-zero Gaussian process $\{\mathbb{Z}(\tau) : \tau \in T\}$ with covariance function $\mathrm{E}\, \mathbb{Z}(\tau)\mathbb{Z}(\tau')^\top = J^{-1}(\tau)\Sigma(\tau, \tau')J^{-1}(\tau')$, $\tau, \tau' \in T$.*

Remark. (i) The proof of the theorem proceeds via empirical process arguments similar to those used in Angrist, Chernozhukov, and Fernández-Val (2006). Their results do not carry over to the present case because heterogeneous cluster data is not covered by their iid assumptions. This prevents the use of standard Donsker theorems and leads to measurability issues typically not encountered in such proofs. Both problems are taken care of through results of Pollard (1990) and Kosorok (2003).

(ii) The theorem implies joint asymptotic normality of $\sqrt{n}(\hat\beta_n(\tau_j) - \beta(\tau_j))$ for every finite set of quantile indices $\tau_j \in T$, $j = 1, 2, \ldots$; see, e.g., Theorem 18.14 of van der Vaart (1998). The asymptotic covariance at τ_j and $\tau_{j'}$ is $J^{-1}(\tau_j)\Sigma(\tau_j, \tau_{j'})J^{-1}(\tau_{j'})$. If

[1] In view of Assumption 2.1, we can always take $(Y_{ik}, X_{ik}^\top)^\top = 0$ for $c_i < k \le c_{\max}$ whenever $c_i < c_{\max}$ to make this a well-defined class of functions from $\mathbb{R}^{c_{\max}} \times \mathbb{R}^{d \times c_{\max}}$ to $\mathbb{R}^d$.

only this finite dimensional convergence is needed, then Assumptions 2.3(iii) and (iv) can be relaxed considerably using the approach of Knight (1998).

The non-iid structure of the data causes the asymptotic covariance function of $\sqrt{n}(\hat{\beta}_n(\tau)-\beta(\tau))$ to take on the sandwich form $J^{-1}(\tau)\Sigma(\tau,\tau')J^{-1}(\tau')$. Estimates of these covariances are needed for Wald-type inference. However, in addition to the usual problem of having to control the nuisance quantities $f_{ik}(X_{ik}^\top\beta(\tau) \mid X_{ik})$ contained in the Jacobian $J(\tau)$, the matrix $\Sigma(\tau,\tau')$ now also contains products of quantile crossing indicators $\psi_\tau(Y_{ik}-X_{ik}^\top\beta(\tau))\psi_{\tau'}(Y_{il}-X_{il}^\top\beta(\tau'))$. For standard plug-in inference, the crossing indicators can be estimated by replacing $\tau \mapsto \beta(\tau)$ with $\tau \mapsto \hat{\beta}_n(\tau)$. The Jacobian is not directly affected by the within-cluster dependence and can be estimated using the bandwidth-driven methods of Hendricks and Koenker (1992) and Powell (1986).

Parente and Santos Silva (2015) propose such a plug-in estimator based on Powell's method and show that it leads to asymptotically valid covariance matrix estimates at individual quantiles in a setting with iid (necessarily equal-sized) clusters.[2] However, both the Hendricks-Koenker and Powell estimators are sensitive to the choice of bandwidth. Parente and Santos Silva (2015, pp. 5-6) give practical suggestions on how to select this bandwidth in the cluster context, but also note that some standard bandwidth rules derived for iid data seem to not perform well in some contexts. To my knowledge, bandwidth rules for QR that explicitly deal with cluster data are currently not available. Moreover, Parente and Santos Silva's method does not extend to uniform inference over ranges of quantiles because the limiting Gaussian process $\{\mathbb{Z}(\tau) : \tau \in T\}$ from Theorem 2.5 is nuisance parameter dependent and cannot be normalized to be free of these parameters. Critical values for inference based on $\mathbb{Z}$ therefore cannot be tabulated.

In the next section I present a bootstrap method that is able to approximate the distribution of the limiting process, consistently estimates the covariance function of that process, and avoids the issue of choosing a bandwidth (and kernel) altogether.

3. Bootstrap Algorithms for Cluster-Robust Inference

In this section I describe and establish the validity of procedures for cluster-robust bootstrap inference (Algorithm 3.4 below) and cluster-robust confidence bands (Algorithm 3.8) in QR models. Recall from the discussion above equation (2.1) that the population first-order condition of the QR objective function can be written as

$$\sqrt{n}M_n'(\beta(\tau),\tau) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{c_i}\mathrm{E}\,\psi_\tau\big(Y_{ik}-X_{ik}^\top\beta(\tau)\big)X_{ik} = 0, \tag{3.1}$$

[2]Their method is likely to generalize to allow for pointwise inference in the presence of clusters with unequal sizes.

where $\psi_\tau(z) = \tau - 1\{z < 0\}$ as before. The sample analogue of this condition,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \psi_\tau(Y_{ik} - X_{ik}^\top \beta) X_{ik} = 0, \tag{3.2}$$

can be thought of as being nearly solved by the QR estimate $\beta = \hat{\beta}_n(\tau)$.

The idea is now to bootstrap by repeatedly computing solutions to perturbations of (3.2). To account for the possible heterogeneity in the data, I use Chen et al.'s (2003) modification of the wild bootstrap (Wu, 1986; Liu, 1988; Mammen, 1992) for QR with correlated data. Chen et al. use their method to obtain confidence intervals for QR estimators at a single quantile. Here, I considerably extend the scope of their method to allow for inference on the entire QR process and uniformly consistent covariance matrix estimates of that process via the bootstrap; confidence intervals at individual quantiles $\tau_j \in T$, $j = 1, 2, \ldots$, then follow as a special case. Because Chen et al. do not give explicit regularity conditions for the validity of their method, this paper also serves as a theoretical justification for their pointwise confidence intervals.

To ensure that the bootstrap versions of the QR estimate accurately reflect the within-cluster dependence, the resampling scheme perturbs the gradient condition (3.2) at the cluster level. Let $W_1, \ldots, W_n$ be iid copies of a random variable W with $\mathrm{E}\, W = 0$, $\mathrm{Var}\, W = 1$, and $\mathrm{E}\,|W|^q < \infty$, where $q > 2$ as in Assumption 2.3(i). Here W is independent of the data. Define the bootstrap gradient process as

$$\mathbb{W}_n(\beta, \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \sum_{k=1}^{c_i} \psi_\tau(Y_{ik} - X_{ik}^\top \beta) X_{ik}.$$

An obvious strategy for bootstrap resampling would now be to repeatedly solve $\mathbb{W}_n(\beta, \tau) = 0$ for β with different draws of $W_1, \ldots, W_n$. However, this type of resampling is impractical because zeros of $\beta \mapsto \mathbb{W}_n(\beta, \tau)$ are difficult to compute due to the fact that $\mathbb{W}_n(\beta, \tau) = 0$ is not a first-order condition of a convex optimization problem.

Instead, I use the bootstrap gradient process $\mathbb{W}_n(\tau) := \mathbb{W}_n(\hat{\beta}_n(\tau), \tau)$ evaluated at the original QR estimate to construct the new objective function

$$\beta \mapsto \mathbb{M}_n^*(\beta, \tau) = \mathbb{M}_n(\beta, \tau) + \mathbb{W}_n(\tau)^\top \beta / \sqrt{n} \tag{3.3}$$

and define the process $\tau \mapsto \hat{\beta}_n^*(\tau)$ as any solution to $\min_{\beta \in B} \mathbb{M}_n^*(\beta, \tau)$. Then $\hat{\beta}_n^*(\tau)$ can be interpreted as the β that nearly solves the corresponding "first-order condition"

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \psi_\tau(Y_{ik} - X_{ik}^\top \beta) X_{ik} = \mathbb{W}_n(\tau).$$

This bootstrap, which I refer to as *wild gradient bootstrap*, essentially perturbs the right-hand side of (3.2) instead of the left. Because $\mathbb{W}_n(\tau)$ mimics the original gradient

process $n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{c_i}\psi_\tau(Y_{ik}-X_{ik}^\top\hat\beta_n(\tau))X_{ik}$ just like the original gradient process mimics the population first-order condition (3.1), choosing $\hat\beta_n^*(\tau)$ in such a way induces the left-hand side of the preceding display to match the behavior of $\mathbb{W}_n(\tau)$. The distributions of $\sqrt{n}(\hat\beta_n^*(\tau)-\hat\beta_n(\tau))$ and $\sqrt{n}(\hat\beta_n(\tau)-\beta(\tau))$ can then be expected to be similar. Theorem 3.1 ahead confirms that this is indeed the case, uniformly in $\tau\in T$.

The distributional convergence occurs both in the standard sense and with probability approaching one, conditional on the sample data $D_n := \{(Y_{ik}, X_{ik}^\top)^\top : 1\le k\le c_i, 1\le i\le n\}$. The latter concept is the standard measure of consistency for bootstrap distributions; see, e.g., van der Vaart (1998, p. 332). Let $\mathrm{BL}_1(\ell^\infty(T)^d)$ be the set of functions on $\ell^\infty(T)^d$ with values in $[-1,1]$ that are uniformly Lipschitz and define $\mathrm{E}^*(\cdot) := \mathrm{E}(\cdot \mid D_n)$.

3.1 Theorem. *If Assumptions 2.1-2.4 hold, then $\{\sqrt{n}(\hat\beta_n^*(\tau)-\hat\beta_n(\tau)) : \tau\in T\}$ converges in distribution to the Gaussian process $\{\mathbb{Z}(\tau) : \tau\in T\}$ described in Theorem 2.5. The convergence also holds conditional on the data in the sense that*

$$\sup_{h\in\mathrm{BL}_1(\ell^\infty(T)^d)}\left|\mathrm{E}^*\, h\big(\sqrt{n}\big(\hat\beta_n^*(\tau)-\hat\beta_n(\tau)\big)\big)-\mathrm{E}\, h(\mathbb{Z}(\tau))\right|\xrightarrow{\mathrm{P}}0.$$

Minimizing the bootstrap objective function (3.3) is a standard convex optimization problem. In fact, as the following algorithm shows, the problem can be implemented in statistical software as a linear QR with one additional observation. The idea is to pick a large enough Y^* to ensure $Y^* > X^{*\top}\hat\beta_n^*(\tau)$ for all $\tau\in T$, where $X^* = -\sqrt{n}\mathbb{W}_n(\tau)/\tau$. Then $\sqrt{n}\mathbb{W}_n(\tau)^\top\beta = \rho_\tau(Y^*-X^{*\top}\beta)-\tau Y^*$ and $-\tau Y^*$ can be ignored because $\hat\beta_n^*(\tau)$ not only minimizes (3.3), but also $\beta\mapsto n\mathbb{M}_n^*(\beta,\tau)-\tau Y^*$.

3.2 Algorithm (Wild gradient bootstrap)**.**

1. Run a QR of Y_{ik} on X_{ik} and save $\tau\mapsto\hat\beta_n(\tau)$. Compute $Y^* = n\max_{1\le i\le n} c_i \max_{1\le i\le n,1\le k\le c_i}|Y_{ik}|$.
2. Draw iid copies $W_1,\dots,W_n$ of W and compute $\mathbb{W}_n(\tau) := \mathbb{W}_n(\hat\beta_n(\tau),\tau)$ for that draw. Generate $X^* = -\sqrt{n}\mathbb{W}_n(\tau)/\tau$ and rerun the QR from Step 1 with the additional observation $(Y^*, X^{*\top})^\top$ to obtain $\tau\mapsto\hat\beta_n^*(\tau) = \arg\min_\beta\sum_{i=1}^n\sum_{k=1}^{c_i}\rho_\tau(Y_{ik}-X_{ik}^\top\beta)+\rho_\tau(Y^*-X^{*\top}\beta)$.
3. Repeat Step 2 m times, each with a new realization of $W_1,\dots,W_n$.
4. Approximate the distribution of $\{\sqrt{n}(\hat\beta_n(\tau)-\beta(\tau)) : \tau\in T\}$ by the empirical distribution of the m observations of $\{\sqrt{n}(\hat\beta_n^*(\tau)-\hat\beta_n(\tau)) : \tau\in T\}$.

Remarks. (i) The idea of representing a perturbed QR problem as linear QR with one additional observation is due to Parzen, Wei, and Ying (1994). The value of Y^* given in the first step of the algorithm is similar to the one suggested by Belloni, Chernozhukov, and Fernández-Val (2011, Algorithm A.4)

(ii) The Monte Carlo experiments in the next section suggest that in practice W should be drawn from the Mammen (1992) 2-point distribution that takes on the value $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and the value $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$. Other distributions such as the Rademacher or Webb (2014) distributions can be used, but there is no evidence that this would lead to better inference.

By choosing the number of bootstrap simulations m in Algorithm 3.2 large enough,[3] the distribution of $\sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))$ or functionals thereof can be approximated with arbitrary precision. I therefore let $m \to \infty$ in the following and define the bootstrap estimate of the asymptotic covariance function $V(\tau, \tau') := J^{-1}(\tau)\Sigma(\tau, \tau')J^{-1}(\tau')$ directly as

$$\hat{V}_n^*(\tau, \tau') = \mathrm{E}^* \, n\big(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau)\big)\big(\hat{\beta}_n^*(\tau') - \hat{\beta}_n(\tau')\big)^\top, \qquad \tau, \tau' \in T.$$

In practice one simply computes $\hat{V}_n^*(\tau, \tau')$ as the sample covariance of the m bootstrap observations of $\sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))$ and $\sqrt{n}(\hat{\beta}_n^*(\tau') - \hat{\beta}_n(\tau'))$. Cluster-robust standard errors of $\hat{\beta}_n(\tau)$ are the square-roots of the diagonal elements of $\hat{V}_n^*(\tau, \tau)/n$.

Availability of a consistent estimate of the covariance function of the limiting process is not strictly required for valid bootstrap inference on the QR process. Algorithm 3.4 ahead shows how this is done. However, especially in the presence of data clusters, applied researchers frequently emphasize the importance of bootstrap covariance matrix estimates for Wald-type inference in mean regression models; see, among others, Bertrand et al. (2004) and Cameron et al. (2008). As the Monte Carlo results in the next section show, reweighting by the bootstrap covariance matrix is equally important for cluster-robust inference in the QR context. Still, because convergence in distribution does not imply convergence in moments, consistency of $\hat{V}_n^*(\tau, \tau')$ does not immediately follow from Theorem 3.1.

Fortunately, the wild gradient bootstrap is able to consistently approximate the asymptotic variance of $\sqrt{n}(\hat{\beta}_n(\tau) - \beta_n(\tau))$. If the covariates have moments of high enough order, then the approximation of the asymptotic covariance function $V(\tau, \tau')$ through its bootstrap counterpart $\hat{V}_n^*(\tau, \tau')$ is in fact uniform in $\tau, \tau' \in T$.

3.3 Theorem. *Suppose Assumptions 2.1-2.4 hold. Then,*

(i) *for all* $\tau, \tau' \in T$, $\hat{V}_n^*(\tau, \tau') \to^{\mathrm{P}} V(\tau, \tau')$, *and*

(ii) *if* $q > 4$ *in Assumption 2.3, then* $\sup_{\tau, \tau' \in T} |\hat{V}_n^*(\tau, \tau') - V(\tau, \tau')| \to^{\mathrm{P}} 0$.

Remarks. (i) For the proof of this theorem I extend ideas developed by Kato (2011), who in turn relies to some extent on the strategy used in the proof of Theorem 3.2.5 of van der Vaart and Wellner (1996) and Alexander's (1985) "peeling device." Kato's results do not apply to the present case because he works with a single quantile, iid data, and a different bootstrap method. For the proof I develop new tail bounds on the QR gradient process and differences of such processes. They yield $\mathrm{E} \sup_{\tau \in T} |\sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))|^p < \infty$ and $\mathrm{E} \sup_{\tau \in T} |\sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau))|^p < \infty$ uniformly in n for $p < q$. The first part of the theorem then follows from Theorem 3.1 and a bootstrap version of a standard uniform integrability result. The proof of the second part is considerably more involved, but relies on the same tail bounds.

(ii) A byproduct of the proof of the theorem is the result that the wild gradient bootstrap correctly approximates other (possibly fractional) order moments of $\mathbb{Z}(\tau)$ if the covariates have moments of slightly higher order: As long as $p < q$, the results in the appendix immediately give $\mathrm{E}^* |\sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))|^p \to^{\mathrm{P}} \mathrm{E} |\mathbb{Z}(\tau)|^p$.

[3] Andrews and Buchinsky (2000) and Davidson and MacKinnon (2000) provide methods for determining an appropriate number of bootstrap simulations m in practice.

(iii) Ghosh, Parr, Singh, and Babu (1984) show that the bootstrap variance estimate of an unconditional quantile can be inconsistent if the bootstrap observations are too likely to take on extreme values. This problem is generic and does not depend on the specific type of bootstrap. The boundedness of the parameter space imposed in Assumption 2.2 prevents such behavior in the bootstrap estimates obtained from the perturbed QR problem (3.3). As Kato (2011) points out, a possible (although not particularly desirable) alternative would be to restrict the moments on the response variable.

(iv) Similar but somewhat simpler arguments can be used to prove analogues of Theorems 2.5, 3.1, and 3.3 for the bootstrap method presented in Parzen et al. (1994) for QR with independent data. For iid data, such analogues of Theorems 2.5 and 3.1 are essentially contained in the results of Belloni et al. (2011) as special cases.

I now turn to inference with the wild gradient bootstrap. Let $\tau \mapsto R(\tau)$ be a continuous, $(h \times d)$-matrix-valued function with $h \leq d$ and let $r\colon T \to \mathbb{R}^d$. Suppose $R(\tau)$ has full rank for every $\tau \in T$. I consider testing general pairs of hypotheses of the form

$$\mathrm{H}_0\colon R(\tau)\beta(\tau) = r(\tau) \text{ for all } \tau \in T, \qquad \mathrm{H}_1\colon R(\tau)\beta(\tau) \neq r(\tau) \text{ for some } \tau \in T.$$

Many empirically relevant hypotheses can be tested with this framework. For example, a standard hypothesis in practice is that a single QR coefficient is zero for all $\tau \in T$. If the coefficient of interest is the first entry of $\beta(\tau)$, then $R(\tau) \equiv (1, 0, \ldots, 0)$ and $r(\tau) \equiv 0$.

For inference I use generalized Kolmogorov-Smirnov statistics. Cramér-von-Mises versions of these statistics can be used as well, but are not discussed here to conserve space. For a positive definite weight matrix function $\tau \mapsto \Omega(\tau)$ with positive square root $\Omega^{1/2}(\tau)$, define the test statistic

$$K_n(\Omega, T) = \sup_{\tau \in T} \big| \Omega^{-1/2}(\tau) \sqrt{n} \big(R(\tau)\hat{\beta}_n(\tau) - r(\tau) \big) \big|. \tag{3.4}$$

I focus on two versions of the statistic: (i) an unweighted version with $\Omega(\tau) \equiv I_d$ and (ii) a Wald-type statistic with $\Omega(\tau)$ equal to

$$\hat{\Omega}_n^*(\tau) := R(\tau)\hat{V}_n^*(\tau, \tau)R(\tau)^\top.$$

Other choices are clearly possible. For example, $\hat{V}_n^*(\tau, \tau)$ can be replaced by any other uniformly consistent estimate of $V(\tau, \tau)$. However, the Monte Carlo study in the next section suggests that option (ii) leads to tests with better finite-sample size and power than tests based on (i) or analytical estimates of $V(\tau, \tau)$.

In the absence of within-cluster correlation, the process inside the Euclidean norm in (3.4) with $\Omega = \hat{\Omega}_n^*$ would converge weakly to a standard vector Brownian bridge. Consequently, $K_n(\hat{\Omega}_n^*, T)$ would converge in distribution to the supremum of a standardized, tied-down Bessel process whose critical values can be simulated or computed exactly; see Koenker and Machado (1999) for details. In the presence of data clusters, the limiting Gaussian process of the quantity inside the Euclidean norm is no longer a Brownian bridge for any choice of weight matrix. Both $K_n(\hat{\Omega}_n^*, T)$ and $K_n(I_d, T)$ are then, in general,

asymptotically non-pivotal statistics. Bootstrap tests based on $K_n(\hat{\Omega}_n^*, T)$ therefore do not necessarily outperform tests based on $K_n(I_d, T)$ because of asymptotic refinements; see, e.g., Hall (1992). However, as I will show below, $K_n(\hat{\Omega}_n^*, T)$ still has the advantage that its square converges to a chi-square distribution if T consists of only a single quantile.

The following algorithm describes how to conduct inference and how to test restrictions on the QR process uniformly over the entire set T. This includes, for example, individual quantiles, finite sets of quantiles, closed intervals, and disjoint unions of closed intervals.

3.4 Algorithm (Wild gradient bootstrap inference)**.** 1. Do Steps 1-3 of Algorithm 3.2.

2. If $\Omega(\tau) = \hat{\Omega}_n^*(\tau)$, compute $\hat{V}_n^*(\tau, \tau)$ as the sample variance of the m bootstrap observations of $\sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))$ from Step 1.
3. For each of the m bootstrap observations from Step 1, calculate

$$K_n^*(\Omega, T) := \sup_{\tau \in T} \big| \Omega^{-1/2}(\tau) \sqrt{n} R(\tau) \big(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau) \big) \big|. \tag{3.5}$$

4. Reject H_0 in favor of H_1 if $K_n(\Omega, T)$ is larger than $q_{n,1-\alpha}(\Omega, T)$, the $1-\alpha$ empirical quantile of the m bootstrap statistics $K_n^*(\Omega, T)$.

As before, I take the number of bootstrap simulations m as large and view the bootstrap quantile $q = q_{n,1-\alpha}(\Omega, T)$ directly as the minimizer of

$$\mathrm{E}^* \Big(\rho_{1-\alpha}\big(K_n^*(\Omega, T) - q \big) - \rho_{1-\alpha}\big(K_n^*(\Omega, T) \big) \Big).$$

Subtracting the second term here again ensures that this expression is necessarily finite without further conditions on the underlying variables.

To prove consistency of Algorithm 3.4 for the Wald-type weight $\hat{\Omega}_n^*$, we also need to guarantee that $\hat{\Omega}_n^*$ is non-singular with probability approaching one as $n \to \infty$. This requires the eigenvalues of $\Sigma(\tau, \tau)$ in $V(\tau, \tau) = J^{-1}(\tau)\Sigma(\tau, \tau)J^{-1}(\tau)$ to be bounded away from zero, uniformly in $\tau \in T$. In the absence of clusters, such a property would automatically follow from non-singularity of $n^{-1} \sum_{i=1}^n \mathrm{E}\, X_{i1} X_{i1}^\top$. (Recall the discussion above Assumption 2.4.) In the cluster context, it is a separate restriction that rules out some scenarios where several clusters have similar forms of extreme within-cluster dependence.

3.5 Assumption. *For all non-zero $a \in \mathbb{R}^d$, $\inf_{\tau \in T} a^\top \Sigma(\tau, \tau) a > 0$.*

The next result shows that Algorithm 3.4 is indeed a consistent test of the null hypothesis $R(\tau)\beta(\tau) = r(\tau)$ for all $\tau \in T$ against the alternative that $R(\tau)\beta(\tau) \neq r(\tau)$ for some $\tau \in T$.

3.6 Theorem. *Suppose Assumptions 2.1-2.4 and 3.5 hold. For $\alpha \in (0, 1)$, we have*

(i) *under* H_0, $\mathrm{P}(K_n(\hat{\Omega}_n^*, T) > q_{n,1-\alpha}(\hat{\Omega}_n^*, T)) \to \alpha$ *and*

(ii) *under* H_1, $\mathrm{P}(K_n(\hat{\Omega}_n^*, T) > q_{n,1-\alpha}(\hat{\Omega}_n^*, T)) \to 1$.

Both results also hold without Assumption 3.5 if I_d is used instead of $\hat{\Omega}_n^*$ in all instances.

Remark. The theorem in fact remains valid if the Euclidean norm in the definition of $K_n(\Omega, T)$ in (3.4) is replaced by any other norm on $\mathbb{R}^d$ as long as the same norm is also employed in the bootstrap statistic $K_n^*(\Omega, T)$ in (3.5). A natural choice other than the Euclidean norm is the maximum norm $|x|_{\max} = \max\{|x_1|, \ldots, |x_d|\}$, i.e., the maximum absolute entry of a vector $x = (x_1, \ldots, x_d)$. I will use this norm below to construct bootstrap confidence bands for the QR coefficient functions.

I now discuss three useful corollaries of Theorems 3.3 and 3.6 regarding (i) chi-square inference with the bootstrap covariance matrix, (ii) bootstrap confidence bands, and (iii) computation of the supremum in the Kolmogorov-Smirnov statistics. First, if T consists of only a single quantile τ_0, then the square of $K_n(\hat{\Omega}_n^*, \tau_0)$ is simply the ordinary Wald statistic

$$n\big(R(\tau_0)\hat{\beta}_n(\tau_0) - r(\tau_0)\big)^\top \hat{\Omega}_n^{*-1}(\tau_0)\big(R(\tau_0)\hat{\beta}_n(\tau_0) - r(\tau_0)\big).$$

Because $\sqrt{n}(R(\tau_0)\hat{\beta}_n(\tau_0) - r(\tau_0))$ is asymptotically multivariate normal under the null hypothesis and $\hat{\Omega}_n^*(\tau_0)$ is consistent for the variance of that multivariate normal distribution, the statistic in the preceding display has an asymptotic chi-square distribution. Hence, chi-square critical values can be used instead of bootstrap critical values for the test decision. The following corollary makes this precise.

3.7 Corollary. *Suppose we are in the situation of Theorem 3.6 with $T = \{\tau_0\}$ for some $\tau_0 \in (0, 1)$. Then*

(i) *under* H_0, $K_n(\hat{\Omega}_n^*, \tau_0)^2 \rightsquigarrow \chi^2_{\operatorname{rank} R(\tau_0)}$ *and*

(ii) *under* H_1, $\mathrm{P}(K_n(\hat{\Omega}_n^*, \tau_0)^2 > q) \to 1$ *for every* $q \in \mathbb{R}$.

Remarks. (i) From this result it also follows immediately that a single QR coefficient at a single quantile can be studentized with its bootstrap standard error and compared to a standard normal critical value.

(ii) The Monte Carlo study below suggests that asymptotic inference using the bootstrap covariance matrix generally performs well and is only slightly worse in terms of finite-sample size than bootstrapping both the covariance matrix and the critical values. Still, when there are very few clusters, asymptotic inference with bootstrap standard errors tends to over-reject while simultaneously having significantly lower power than the test with bootstrap critical values. The over-rejection could, in principle, be avoided by replacing standard normal and chi-square critical values with larger critical values from the Student t_{n-1} and similarly scaled F distributions (Donald and Lang, 2007; Bester, Conley, and Hansen, 2011). However, such small-sample adjustments would decrease the power of the test even further. It is therefore recommended to bootstrap the critical values when only few clusters are available.

Next, the results in Theorems 3.3 and 3.6 allow for the construction of bootstrap confidence bands (uniform in $\tau \in T$) for the QR coefficient function. These bands can be computed jointly for the entire d-dimensional function or only a subset $\Delta \subset \{1, \ldots, d\}$ of coefficients. As before, a positive definite weight matrix function, denoted here by

$\tau \mapsto \Lambda(\tau)$, can be specified to improve the finite-sample performance. An obvious choice is $\Lambda(\tau) = \hat{V}_n^*(\tau, \tau)$. In the following algorithm and in the corollary immediately below, I write a_j for the jth entry of a d-vector a and A_{jj} for the jth diagonal element of a $d \times d$ square matrix A.

3.8 Algorithm (Wild gradient bootstrap confidence bands)**.** 1. Do Steps 1-3 of Algorithm 3.2 and, if $\Lambda(\tau) = \hat{V}_n^*(\tau, \tau)$, compute $\hat{V}_n^*(\tau, \tau)$ as in Step 2 of Algorithm 3.4.

2. For each of the m bootstrap observations, calculate

$$K_n^*(\Lambda, T, \Delta) := \sup_{\tau \in T} \max_{j \in \Delta} \left| \frac{\hat{\beta}_n^*(\tau)_j - \hat{\beta}_n(\tau)_j}{\sqrt{\Lambda(\tau)_{jj}/n}} \right|$$

and $q_{n,1-\alpha}(\Lambda, T, \Delta)$, the $1-\alpha$ empirical quantile of $K_n^*(\Lambda, T, \Delta)$.

3. For each $\tau \in T$ and $j \in \Delta$, compute the interval

$$\left[\hat{\beta}_n(\tau)_j \pm q_{n,1-\alpha}(\Lambda, T, \Delta) \sqrt{\Lambda(\tau)_{jj}/n} \right].$$

The confidence band given in the last step of the algorithm has asymptotic coverage probability $1-\alpha$. The proof of this result is based on the fact that, as long as the maximum norm is used in (3.5) instead of the Euclidean norm, $K_n^*(\Lambda, T, \Delta)$ is nothing but the bootstrap statistic $K_n^*(\Omega, T)$ with a diagonal weight matrix and a matrix of restrictions $R(\tau) \equiv R$ that selects the coefficients given in Δ.

3.9 Corollary. *Suppose we are in the situation of Theorem 3.6. For every $\Delta \subset \{1, \dots, d\}$*

$$\mathrm{P}\left(\beta(\tau)_j \in \left[\hat{\beta}_n(\tau)_j \pm q_{n,1-\alpha}(\hat{V}_n^*, T, \Delta) \sqrt{\hat{V}_n^*(\tau, \tau)_{jj}/n} \right] \text{ for all } \tau \in T, \text{ all } j \in \Delta \right)$$

converges to $1-\alpha$ as $n \to \infty$. This continues to hold without Assumption 3.5 if all instances of $\hat{V}_n^*$ are replaced by I_d.

Finally, if T is not a finite set, computing $K_n(\Omega, T)$ and the confidence bands is generally infeasible in practice due to the presence of a supremum in their definitions. This can be circumvented by replacing the supremum with a maximum over a finite grid $T_n \subset T$ that becomes finer as the sample size increases. For example, if T is a closed interval, we can take $T_n = \{j/n : j = 0, 1, \dots, n\} \cap T$. For any τ in the interior of T and n large enough, we can then find $\tau_n, \tau_n' \in T_n$ that differ by $1/n$ and satisfy $\tau_n \leq \tau < \tau_n'$. This gives $0 \leq \tau - \tau_n < 1/n$. Furthermore, the endpoints of T_n are less than $1/n$ away from the respective endpoints of T. Hence, every $\tau \in T$ is the limit of a sequence $\tau_n \in T_n$. This turns out to be the property needed to ensure that the approximation of T by a finite set has no influence on the asymptotic behavior of the bootstrap test.

3.10 Corollary. *Suppose we are in the situation of Theorem 3.6 and there exist sets $T_n \subset T$ such that for every $\tau \in T$ there is a sequence $\tau_n \in T_n$ such that $\tau_n \to \tau$ as $n \to \infty$. Then Theorem 3.6 and Corollary 3.9 continue to hold when T_n is used instead of T.*

The next section illustrates the finite-sample behavior of the wild gradient bootstrap in a brief Monte Carlo exercise. Section 5 then provides an application of the wild gradient bootstrap to Project STAR data.

4. Monte Carlo Experiments

This section presents several Monte Carlo experiments to investigate the small-sample properties of the wild gradient bootstrap in comparison to other methods of inference. I discuss significance tests at a single quantile (Experiment 4.1), inference about the QR coefficient function (Experiment 4.2), and confidence bands (Experiment 4.3).

The data generating process (DGP) for the following experiments is

$$Y_{ik} = 0.1U_{ik} + X_{ik} + X_{ik}^2 U_{ik},$$

where $X_{ik} = \sqrt{\varrho} Z_i + \sqrt{1-\varrho}\varepsilon_{ik}$ with $\varrho \in [0,1)$; Z_i and ε_{ik} are standard normal, independent of each other, and independent across their indices. This guarantees that the X_{ik} are standard normal and, within each cluster, any two observations X_{ik} and X_{il} have a correlation coefficient of ϱ. The U_{ik} are distributed as $\mathrm{N}(0, 1/3)$ and drawn independently of X_{ik} to ensure that the $X_{ik}^2 U_{ik}$ have mean zero and variance one. The correlation structure of U_{ik} is chosen such that the within-cluster correlation coefficient of $X_{ik}^2 U_{ik}$ is also approximately ϱ.[4] Both X_{ik} and U_{ik} are independent across clusters.

The DGP in the preceding display corresponds to the quadratic QR model

$$Q_{ik}(\tau \mid X_{ik}) = \beta_0(\tau) + \beta_1(\tau)X_{ik} + \beta_2(\tau)X_{ik}^2 \tag{4.1}$$

with $\beta_0(\tau) = \Phi^{-1}(\tau)/\sqrt{300}$, $\beta_1(\tau) \equiv 1$, and $\beta_2(\tau) = \Phi^{-1}(\tau)/\sqrt{3}$, where Φ is the standard normal distribution function. I denote the QR estimates of the two slope parameters $\beta_1(\tau)$ and $\beta_2(\tau)$ by $\hat{\beta}_{1,n}(\tau)$ and $\hat{\beta}_{2,n}(\tau)$. Their bootstrap versions are $\hat{\beta}^*_{1,n}(\tau)$ and $\hat{\beta}^*_{2,n}(\tau)$. As before, I refer to the square roots of the diagonal elements of $\hat{V}^*_n(\tau,\tau)/n$ as bootstrap standard errors and, for simplicity, now denote the bootstrap standard error of $\hat{\beta}_{2,n}(\tau)$ by $\mathrm{se}^*(\hat{\beta}^*_{2,n}(\tau))$.

In the following experiments, I consider inference about $\tau \mapsto \beta_1(\tau)$ and $\tau \mapsto \beta_2(\tau)$ for different values of the number of clusters n, the within-cluster correlation ϱ, and the variance of the cluster size $\mathrm{Var}(c_i)$. In all experiments below, the smallest possible cluster size is 5 and c_i is distributed uniformly on $\{5, 6, \ldots, c_{\max}\}$. Unless otherwise noted, the bootstrap weights are drawn from the Mammen distribution as defined in the remarks below Algorithm 3.2.

[4] By construction, the correlation coefficient of $X_{ik}^2 U_{ik}$ and $X_{il}^2 U_{il}$ is $\mathrm{Corr}(U_{ik}, U_{il})(2\varrho^2 + 1)/3$. I generate data such that $\mathrm{Corr}(U_{ik}, U_{il}) = \min\{1, 3\varrho/(2\varrho^2 + 1)\}$. The within-cluster correlation coefficient of $X_{ik}^2 U_{ik}$ is then exactly ϱ for $\varrho \in [0, 0.5]$ and has a value slightly below ϱ for $\varrho \in (0.5, 1)$. This choice for $\mathrm{Corr}(U_{ik}, U_{il})$ ensures that the other restrictions on the DGP hold for all values of ϱ used in the experiments.

4.1 Experiment (Significance tests at a single quantile)**.** This Monte Carlo experiment illustrates the small-sample size and power of different methods for testing whether a single QR coefficient equals zero at a given quantile. To test the correct null hypothesis $\beta_2(.5) = 0$ in (4.1) against the alternative $\beta_2(.5) \neq 0$, I consider (i) wild gradient bootstrap inference as in Algorithm 3.4, (ii) standard inference with bootstrap standard errors as in Corollary 3.7, (iii) cluster-robust inference based on analytically estimating the standard errors, (iv) standard inference without cluster correction, (v) cluster-robust Rao score inference, and (vi) wild bootstrap inference without cluster correction.

For (i), note that $R \equiv (0, 0, 1)$ and $r \equiv 0$. Hence, Algorithm 3.4 is equivalent to testing whether $|\hat{\beta}_{2,n}(.5)|$ exceeds the empirical $1 - \alpha$ quantile of the m observations of $|\hat{\beta}^*_{2,n}(.5) - \hat{\beta}_{2,n}(.5)|$ conditional on $\hat{\beta}_{2,n}(.5)$. No weight matrix is needed because the test decision is independent of $\Omega(\tau)$ whenever $R(\tau)V(\tau,\tau)R(\tau)^\top$ is a scalar. Similarly, for (ii), the test decision in Corollary 3.7 is equivalent to simply comparing $|\hat{\beta}_{2,n}(.5)|/\operatorname{se}^*(\hat{\beta}^*_{2,n}(.5))$ to $\Phi^{-1}(1 - \alpha/2)$. For (iii), I obtain standard errors by estimating $V(\tau,\tau) = J^{-1}(\tau)\Sigma(\tau,\tau)J^{-1}(\tau)$ analytically as suggested by Parente and Santos Silva (2015). They propose the plug-in estimate

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c_i}\sum_{l=1}^{c_i}\psi_\tau\big(Y_{ik} - X_{ik}^\top\hat{\beta}_n(\tau)\big)\psi_\tau\big(Y_{il} - X_{il}^\top\hat{\beta}_n(\tau)\big)X_{ik}X_{il}^\top$$

for $\Sigma(\tau,\tau)$ and replace $J(\tau)$ by a Powell (1986) kernel estimate. The kernel estimate requires a bandwidth choice. The results here are based on the standard implementation in the `quantreg` package in `R` with the Hall-Sheather rule; see Koenker (2005, pp. 80-81) and Koenker (2013).[5] For (iv), I use the regular version of the Powell sandwich estimator described in Koenker (2005). It employs the same kernel estimate of $J(\tau)$ as in (iii), but replaces $\Sigma(\tau,\tau)$ by $n^{-1}\tau(1-\tau)\sum_{i=1}^{n}\sum_{k=1}^{c_i}X_{ik}X_{ik}^\top$ and is therefore not designed to account for within-cluster correlation. For (v), I apply the QRS_0 test of Wang and He (2007), a cluster-robust version of the QR rank score test (Gutenbrunner, Jurêcková, Koenker, and Portnoy, 1993). Wang and He derive their test statistic under homoscedasticity assumptions; the DGP considered here is highly heteroscedastic. For (vi), I compute critical values from the `quantreg` implementation of the Feng et al. (2011, FHH hereafter) wild bootstrap for QR models. Their method perturbs the QR residuals via a carefully chosen weight distribution but presumes independent observations. An alternative wild bootstrap procedure due to Davidson (2012) had size properties similar to those of the FHH method but had lower power in nearly all of my experiments; results for this bootstrap are therefore omitted.

Panels (a)-(c) in Figure 1 show empirical rejection frequencies of a correct hypothesis $\mathrm{H}_0\colon \beta_2(.5) = 0$ for methods (i)-(vi) at the 5% level (short-dashed line) as a function of (a)

[5]This bandwidth choice required a robust estimate of scale. Koenker (2013) uses the minimum of the standard deviation of the QR residuals and their normalized interquartile range. Parente and Santos Silva (2015) suggest the median absolute deviation of the QR residuals with a scaling constant of 1. I chose Koenker's implementation because it yielded better results in nearly all cases.

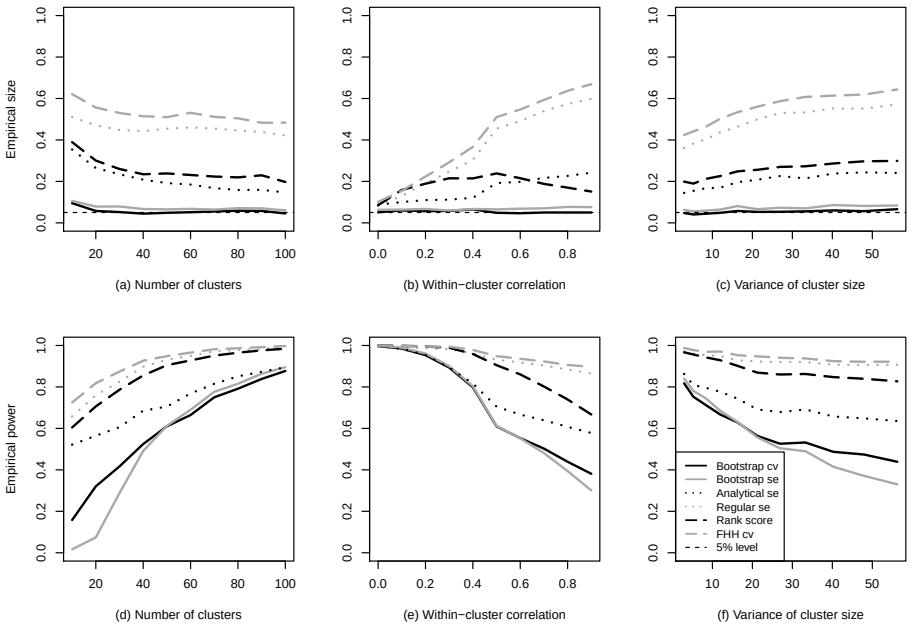

Figure 1: Empirical rejection frequencies of a correct hypothesis $\mathrm{H}_0\colon \beta_2(.5) = 0$ (panels (a)-(c)) and the incorrect hypothesis $\mathrm{H}_0\colon \beta_2(.75) = 0$ (panels (d)-(f)) using wild gradient bootstrap critical values (solid black lines), bootstrap standard errors (solid grey), analytical cluster-robust standard errors (dotted black), regular standard errors without cluster correction (dotted grey), cluster-robust rank score inference (long-dashed black), and FHH wild bootstrap without cluster correction (long-dashed grey) at the 5% level (short-dashed) as a function of the (a) number of clusters, (b) within-cluster correlation, and (c) maximal cluster size.

the number of clusters n, (b) the within-cluster correlation ϱ, and (c) the variance of the cluster size $\mathrm{Var}(c_i)$. Each horizontal coordinate was computed from 2,000 simulations and all six methods were faced with the same data. The three bootstrap tests used $m = 299$ bootstrap repetitions. The wild gradient bootstrap had Mammen weights. Results for other weight distributions are discussed below.

For panel (a), I set $\varrho = .5$, $\mathrm{Var}(c_i) = 10$ (i.e., $c_{\max} = 15$), and considered $n \in \{10, 20, \ldots, 100\}$. As can be seen, the wild gradient bootstrap critical values (solid black lines) and bootstrap standard errors (solid grey) provided tests that were essentially at the nominal level with as few as 20 clusters, with the bootstrap critical values performing slightly better. Tests based on analytical cluster-robust standard errors (dotted black) and cluster-robust rank scores (long-dashed black) over-rejected significantly, although this property became less pronounced for larger numbers of clusters. Regular standard errors

without cluster correction (dotted grey) and wild bootstrap without cluster correction (long-dashed grey) led to severe over-rejection in all cases. For (b), I chose $n = 50$, $\text{Var}(c_i) = 10$, and varied $\varrho \in \{0, .1, \ldots, .9\}$. At $\varrho = 0$, all tests except the rank score test apply and had almost correct size. For larger values of the within-cluster correlation, the three analytical tests and the FHH bootstrap test over-rejected considerably, although the rank score test improved for very high correlations. The test based on $\text{se}^*(\hat{\beta}^*_{2,n}(\tau))$ over-rejected mildly. The bootstrap test given in Algorithm 3.4 was nearly at the nominal level in most cases. For (c), I fixed $\varrho = .5$ and changed $c_{\max} \in \{9, 11, \ldots, 29\}$ so that $\text{Var}(c_i)$ increased from 2 to 52 over this range. I simultaneously decreased n in order to keep the average total number of observations constant at approximately 250; this resulted in numbers of clusters between 36 and 15. The test based on the bootstrap standard error again over-rejected slightly but far less than the ones based on the analytical cluster-robust standard error and the cluster-robust rank score. Wild gradient bootstrap critical values again provided a test with almost proper size, while regular standard errors and the wild bootstrap for independent observations were not useful at any value of $\text{Var}(c_i)$.

Panels (d)-(f) show empirical rejection frequencies of the incorrect hypothesis $\text{H}_0 \colon \beta_2(.75) = 0$ for the same data. Rejection frequencies of the three analytical methods and the FHH wild bootstrap are only reported for completeness and, because of their size distortion, should not be interpreted as estimates of their power. The wild gradient bootstrap critical values tended to lead to a more powerful test than inference with bootstrap standard errors. This was, in particular, the case in small samples, at high within-cluster correlations, and for large variances of the cluster size. The rejection frequencies of all tests were increasing in the number of clusters, decreasing in the within-cluster correlations, and decreasing in the variance of the cluster size.

Following MacKinnon and Webb (2015), I also experimented (not shown) with cases where I varied the within-cluster correlation of X and U in the Monte Carlo DGP independently. For the wild gradient bootstrap, I found that for any within-cluster correlation of X, the degree of correlation in U had little impact, whereas increases in the within-cluster correlation in X led to mild size distortions similar to the ones found in Figure 1. In contrast, increases in the within-cluster correlation of U led to severe over-rejection in tests based on analytical cluster-robust standard errors; higher correlation in X also induced over-rejection, but the impact was considerably less pronounced.

In light of the findings so far it should be noted that the small-sample results for the analytically estimated standard errors reported here do not contradict the ones reported by Parente and Santos Silva (2015), who find a much better performance of their method in terms of finite-sample size. In comparison to their experiments, I consider data with smaller numbers of clusters, different correlation structures, and much stronger cluster heterogeneity in terms of cluster sizes. Computing the standard errors analytically worked well when the number of clusters was large, the within-cluster correlation was low, and the clusters were small. Similarly, the rank score test of Wang and He (2007) is designed for homoscedastic models and performed much better in such settings. For heteroscedastic models, Wang (2009) shows that reweighting their test statistic can significantly improve

inference when more is known about the specific form of heteroscedasticity; her reweighting schemes do not apply to the DGP in the present example and are therefore not discussed.

Table 1: Empirical size and power as in Figure 1 for different bootstrap weights

	$H_0\colon \beta_2(.5) = 0$ (size)						$H_0\colon \beta_2(.75) = 0$ (power)					
	Mammen		Rademacher		Webb		Mammen		Rademacher		Webb	
n	cv	se	cv	se	cv	se	cv	se	cv	se	cv	se
10	.098	.114	.131	.166	.128	.146	.155	.019	.104	.016	.094	.011
20	.068	.088	.076	.102	.071	.098	.328	.086	.302	.036	.270	.026
100	.054	.068	.059	.070	.054	.070	.876	.896	.864	.886	.868	.890
ϱ												
.1	.061	.069	.063	.071	.062	.068	.998	1	.998	1	.999	1
.5	.055	.071	.059	.074	.061	.076	.602	.613	.590	.544	.589	.502
.9	.057	.078	.067	.088	.065	.091	.378	.308	.376	.183	.371	.156
$\mathrm{Var}(c_i)$												
2	.056	.070	.057	.072	.056	.072	.820	.840	.808	.830	.803	.831
24	.054	.076	.062	.078	.060	.074	.580	.578	.578	.479	.570	.446
52	.056	.082	.062	.085	.066	.086	.456	.364	.459	.235	.446	.183

Bootstrap weight distributions other than the Mammen distribution are often found to work well in regression settings. These include the standard normal distribution, the recentered Exponential(1) distribution, the Rademacher distribution, which takes on the values -1 and 1 with equal probability, and the Webb (2014) 6-point distribution, which takes on $-\sqrt{1.5}, -1, -\sqrt{0.5}, \sqrt{0.5}, 1$, and $\sqrt{1.5}$ with equal probability. In my experiments, the standard normal had size properties very similar to those of the Rademacher and Webb distributions, but lower power. I therefore do not present detailed results for this distribution. The same holds for the recentered Exponential(1), which behaved almost like the Mammen distribution in terms of size, but also had lower power. Comparisons of the other distributions are shown in Table 1. The experimental setup and data were the same as in Figure 1. The left-hand side of the table measures finite-sample size for different numbers of clusters, within-cluster correlations, and variances of the cluster size as in panels (a), (b), and (c) of Figure 1; the right-hand side corresponds to the power estimates in panels (d)-(f). As can be seen, the Mammen distribution had slightly better size and power, in particular when the number of clusters was small, the within-cluster correlation was high, and the variance of the cluster size was large.

To further investigate finite-sample power of the three bootstrap weight distributions, I plot in Figure 2 their rejection frequencies of $H_0\colon \beta_2(\tau) = 0$ at 17 separate quantile indices $\tau \in \{.1, .15, \ldots, .9\}$ for $n = 75$. I again chose $\varrho = .5$, $\mathrm{Var}(c_i) = 10$, $m = 299$, and 2,000 Monte Carlo simulations. For this experiment all 17 possible null hypotheses were tested with the same data. Only $H_0\colon \beta_2(.5) = 0$ is true. As the plot shows, wild gradient bootstrap critical values led to tests with good size and power at all quantiles and for all weight distribution. Size and power of the tests based on bootstrap standard errors were similar for $\tau \in [.2, .8]$. However, for quantile indices outside this interval the tests with bootstrap standard errors from the Rademacher (dotted grey) and Webb (long-dashed

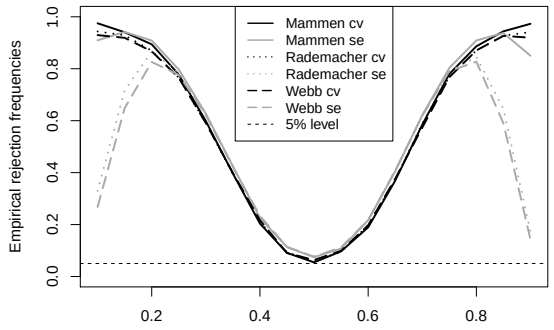

Figure 2: Rejection frequencies of $\mathrm{H}_0\colon \beta_2(\tau) = 0$ for different values of τ using the same methods as in Table 1. H_0 is only true at $\tau = .5$.

grey) distributions showed a sharp decline in power; the Mammen distribution (solid grey) did not have this issue. I experimented with the parameters of the DGP and found that the power of the Rademacher and Webb distributions for small and large τ increased quickly when I increased the number of clusters or decreased the within-cluster correlation. For example, for the Rademacher distribution the rejection frequency at $\tau = .1$ and .9 was about 80% when I either set ϱ to .3 or n to 100.

The reason for the large differences in finite-sample power between the distributions appears to be the extreme skewness of the distribution of the summands in the gradient process for small and large τ. The asymmetry in the Mammen distribution seems to mimic this property particularly well. The standard errors also improved when I used a recentered Exponential(1) or other asymmetric distributions, but the Mammen distribution provided the best results. □

4.2 Experiment (Uniform inference on the QR process)**.** This experiment illustrates the finite-sample performance of Algorithm 3.4 for inference on the entire QR process. I tested the true hypothesis $\mathrm{H}_0\colon \beta_1(\tau) = 1$ for all $\tau \in T$ and the false hypothesis $\mathrm{H}_0\colon \beta_1(\tau) = 0$ for all $\tau \in T$ at the 5% level. I chose $\mathrm{Var}(c_i) = 10$, $m = 199$ bootstrap simulations with the Mammen distribution, and, in view of Corollary 3.10, I approximated $T = [.1, .9]$ by $\{.1, .2, \ldots, .9\}$. The test statistics $K_n(\Omega, T)$ were either (i) weighted by the bootstrap estimate $\hat{\Omega}_n^*$, (ii) weighted by the analytical estimate of $\tau \mapsto R(\tau)V(\tau,\tau)R(\tau)^\top$ described in the preceding Monte Carlo exercise, or (iii) unweighted ($\Omega = I$). All three methods were faced with the same data.

Table 2 reports the empirical rejection frequencies of the true null hypothesis for methods (i)-(iii) from an experiment with 1,000 Monte Carlo simulations for each $\varrho \in \{0, .1, \ldots, .9\}$ and each $n \in \{20, 30, 50\}$. At $n = 20$, the test based on the bootstrapped Wald weight ("boot.") was quite conservative for all degrees of within-cluster correlation but not overly

Table 2: Empirical size of Algorithm 3.4 at the 5% level

	$n = 20$			$n = 30$			$n = 50$		
ϱ	boot.	ana.	unw.	boot.	ana.	unw.	boot.	ana.	unw.
0	.029	.026	.012	.034	.014	.011	.038	.026	.012
.1	.030	.022	.004	.038	.029	.013	.029	.029	.011
.2	.020	.016	.006	.025	.026	.010	.035	.034	.015
.3	.020	.017	.002	.032	.028	.009	.042	.039	.017
.4	.010	.006	.001	.030	.026	.009	.048	.039	.011
.5	.008	.001	.000	.016	.011	.001	.036	.025	.006
.6	.005	.000	.000	.015	.006	.000	.041	.029	.004
.7	.011	.001	.000	.015	.010	.000	.039	.030	.001
.8	.012	.000	.000	.008	.005	.000	.037	.027	.003
.9	.012	.000	.000	.006	.002	.000	.032	.020	.002

so for ϱ smaller than .4. The performance of the bootstrap test with analytical weights ("ana.") was slightly worse, especially for higher within-cluster correlations. However, both of these tests under-rejected considerably less for larger n so that at $n = 50$ the size of bootstrap-weighted test was above .3 for all but one ϱ. In contrast, the unweighted version was very conservative for all within-cluster correlations and all numbers of clusters.

Table 3: Empirical power of Algorithm 3.4 at the 5% level

	$n = 10$			$n = 15$			$n = 20$		
ϱ	boot.	ana.	unw.	boot.	ana.	unw.	boot.	ana.	unw.
0	1	.735	.692	1	.986	.983	1	1	1
.1	.994	.431	.373	1	.924	.904	1	.994	.991
.2	.944	.173	.133	.993	.704	.665	1	.962	.945
.3	.828	.043	.027	.964	.352	.320	.999	.787	.747
.4	.665	.013	.001	.849	.085	.062	.963	.417	.362
.5	.438	.007	.000	.594	.005	.001	.790	.063	.041
.6	.412	.003	.000	.531	.004	.000	.703	.034	.016
.7	.409	.006	.000	.474	.001	.000	.624	.024	.008
.8	.388	.006	.000	.423	.003	.001	.527	.014	.002
.9	.380	.006	.001	.338	.006	.001	.428	.014	.001

Table 3 shows empirical rejection frequencies of the false null hypothesis $\mathrm{H}_0\colon \beta_1(\tau) = 0$ for all $\tau \in T$ at $n \in \{10, 15, 20\}$ in the same experimental setup as above. For $n = 10$, the Wald test with bootstrap weights had substantial power even for high within-cluster correlations. In sharp contrast, the unweighted and analytically weighted tests rejected considerably fewer false null hypotheses and exhibited a total loss of power starting from about $\varrho = .5$. Increases in the number of clusters translated into significant gains in the power of all tests, but the test based on the bootstrap weight matrix far outperformed the other two tests at all sample sizes. □

4.3 Experiment (Confidence bands)**.** In this experiment I investigate the finite-sample properties of Algorithm 3.8. The setup is as in the preceding experiment. The empirical coverage of $\tau \mapsto \beta_1(\tau)$ with a 95% wild gradient bootstrap confidence band is, by construction, identical to 1 minus the empirical size of the bootstrap test in Table 2 and therefore not shown here. I instead consider a more complex scenario where I report the empirical coverage of a joint 95% confidence band for the two slope functions $\tau \mapsto \beta_1(\tau)$ and $\tau \mapsto \beta_2(\tau)$ for $n \in \{10, 15, 20\}$ and $\varrho \in \{0, .1, \ldots, .9\}$. Table 4 contains the results.

Table 4: Empirical coverage of $\tau \mapsto (\beta_1, \beta_2)(\tau)$ by 95% confidence band

	$n = 10$			$n = 15$			$n = 20$		
ϱ	boot.	ana.	unw.	boot.	ana.	unw.	boot.	ana.	unw.
0	.939	.953	.993	.938	.949	.993	.945	.958	.995
.1	.949	.977	.997	.923	.938	.992	.938	.942	.992
.2	.955	.981	.991	.920	.946	.988	.930	.934	.990
.3	.960	.994	.997	.945	.975	.993	.934	.947	.994
.4	.954	.998	1	.966	.993	1	.948	.972	.997
.5	.962	1	1	.981	.999	1	.974	.998	1
.6	.950	.999	1	.958	1	1	.972	.998	1
.7	.937	.999	1	.960	.998	1	.969	1	1
.8	.919	.999	1	.949	.997	.999	.965	.998	1
.9	.900	.998	1	.941	.997	.999	.960	.996	1

As before, the procedure based on the bootstrapped Wald weight showed the most balanced performance with confidence bands that were close to 95% in most cases. The only exceptions occurred at $n = 10$ for very high within-cluster correlations, where the confidence bands were too thin. The unweighted confidence bands were consistently too wide. For analytical weights, the empirical coverage was near 95% for small ϱ. However, at values of ϱ larger than .4 the coverage was essentially 100% even for $n = 20$. Further increases in n (not shown) yielded improvements for all versions of the confidence band but even the bootstrap-weighted confidence band needed a large number of clusters for the coverage to be fully balanced across ϱ. □

In summary, the wild gradient bootstrap performs well even in fairly extreme (but empirically relevant) situations where the number of clusters is small, the within-cluster correlation is high, and the clusters are very heterogeneous. Here, reweighting the test statistic by the bootstrap covariance matrix is crucial for tests to have good size and power in finite samples. Analytical weights or no weights can be used when the number of clusters is large; otherwise they tend to lead to tests that are less reliable than those based on the bootstrapped Wald weight. For inference at a single quantile, testing with bootstrap standard errors and normal/chi-square critical values provides a simpler alternative to testing with bootstrap critical values that is, with some exceptions, nearly as good. These findings are also confirmed by an additional experiment in the next section, where I implement placebo interventions in the Project STAR data.

5. Application: Project STAR

This section applies the wild gradient bootstrap to investigate the effects of a class size reduction experiment on the conditional quantiles of student performance on a standardized test. The data come from the first year of the Tennessee *Student/Teacher Achievement Ratio* experiment, known as Project STAR.

I start by briefly describing Project STAR; the discussion closely follows Word et al. (1990) and Graham (2008), where more details can be found. At the beginning of the 1985-1986 school year, incoming kindergarten students who enrolled in one of the 79 project schools in Tennessee were randomly assigned to one of three class types within their school: a small class (13-17 students), a regular-size class (22-25 students), or a regular-size class (22-25 students) with a full-time teacher's aide. Teachers were then randomly assigned to one of these class types. Each of the project schools was required to have at least one of each kindergarten class type. During the 1985-1986 school year, a total of 6,325 students in 325 different classrooms across Tennessee participated in the project. Classroom identifiers are not available, but Graham's (2008) matching algorithm is able to uniquely identify 317 of these classrooms in the data. 5,727 students in these classrooms have the complete set of characteristics available that I use in the QR model below. I restrict the analysis to only these kindergarten students.

The outcome of interest is student performance on the *Stanford Achievement Test* in mathematics and reading administered at the end of the 1985-1986 school year. I standardized the raw test scores as in Krueger (1999): First, I computed the empirical distribution functions of the math and reading scores for the pooled sample of regular (with and without teacher's aide) students. Next, I transformed the math and reading scores for students in all three class types into percentiles using the math and reading empirical distribution functions, respectively, obtained in the first step. Finally, to summarize overall performance, I computed the average of the two percentiles for each student. I use this percentile score as the dependent variable in the following analysis. The idea behind Krueger's normalization is that in the absence of a class size effect, the transformed subject scores for both small and regular class types would have an approximately uniform distribution.

The two main covariates of interest are the treatment dummy *small* indicating whether the student was assigned to a small class and the treatment dummy *regaide* indicating whether the student was assigned to a regular class with an aide. I consider the following model for the conditional quantiles of the transformed scores:

$$Q_{ik}(\tau \mid X_{ik}) = \beta_0(\tau) + \beta_1(\tau)small_{ik} + \beta_2(\tau)regaide_{ik} + \beta_3(\tau)^\top Z_{ik}. \tag{5.1}$$

This specification is similar to the mean regression given in Krueger's (1999) Table V.4. The covariate vector Z_{ik} contains a dummy indicating if the student identifies as *black*,[6]

[6]The sample also contains a large number of students who identify as white and a very small number of students who identify as Hispanic, Asian, American Indian, or other.

a student gender dummy, a dummy indicating whether the student is *poor* as measured by their access to free school lunch, a dummy indicating if the teacher identifies as black (*tblack*, the other teachers in the sample identify as white), the teacher's years of teaching experience (*texp*), a dummy indicating whether the teacher has at least a master's degree (*tmasters*), and additive school "fixed effects." Because of possible peer effects and unobserved teacher characteristics, I cluster at the classroom level.

The results are shown in Figure 3. The solid black lines in each panel plot a coefficient estimate corresponding to a coefficient in (5.1) as a function of τ. The vertical scale is the average percentile score. The grey bands are pointwise 95% wild gradient bootstrap confidence intervals based on bootstrap quantiles computed from $m = 999$ bootstrap simulations with Mammen weights. Students assigned to small classes mostly perform better than students assigned to regular classes (with or without aide), although the effect varies across the distribution. For scores above the .2 quantile of the score distribution, the difference is about five percentage points. This is in accordance with Krueger's (1999) findings. However, the benefits for students below the .2 quantile are much smaller and become insignificant at the .1 quantile. The impact of a smaller class on students at the very bottom of the score distribution is essentially zero. In addition, as in Krueger's mean regression analysis, the effect of being assigned a full-time aide is insignificant.

I now briefly discuss the other covariates. Black students perform worse than non-black students with otherwise identical characteristics; this is particularly pronounced between the first and third quartiles of the conditional score distribution, where black students' scores are about 10 percentage points lower. Girls generally score higher than boys, although the gap is quite small near the tails of the conditional score distribution. Poor students score up to 15 percentage points lower than otherwise identical students; however, this difference is much smaller near the top of the conditional distribution. As Krueger (1999) and earlier studies have found, teacher characteristics seem to matter little: their race and education (measured by whether they have a master's degree) have no significant impact. Another year of teaching experience has a small, positive effect for all but the very best students.

As a referee points out, an issue with Monte Carlo studies such as those in the preceding section is that the data sets used in simulations are likely to be quite different from real data sets. I therefore also evaluate the performance of the wild gradient bootstrap and the alternative methods introduced in Experiment 4.1 above through placebo interventions in the Project STAR data.

5.1 Experiment (Placebo interventions)**.** For this experiment, I removed all small classes from the sample so that only 194 regular-size classes (with and without teacher's aide) in the 79 project schools remained. Of these schools, 16 had two regular-size classes without aide and 2 had three such classes.

In each of these 18 schools, I then randomly assigned one of the regular-size classes without aide the treatment indicator $small = 1$. This mimics the random assignment of class sizes within schools in the original sample, even though in this case no student

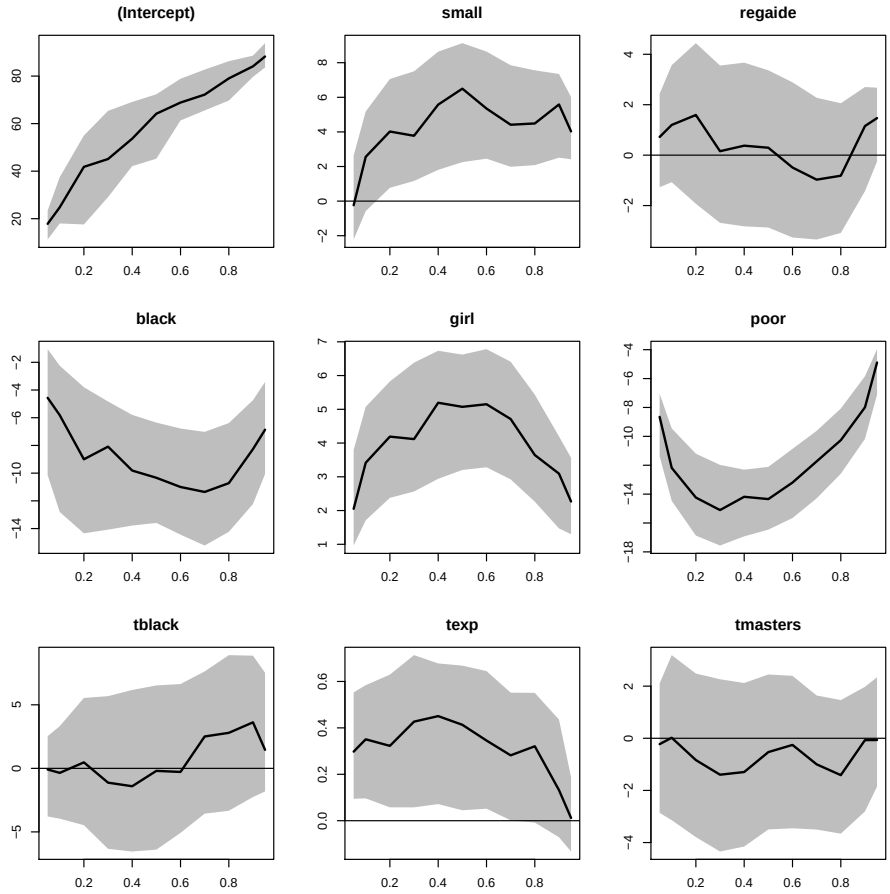

Figure 3: QR coefficient estimates $\tau \mapsto \hat{\beta}_n(\tau)$ (solid black lines) of model (5.1). The regression includes additive school "fixed effects" (not shown). The grey areas are pointwise 95% wild gradient bootstrap confidence intervals based on bootstrap quantiles clustered at the classroom level.

Table 5: Rejection frequencies of $H_0 : \beta_1(.5) = 0$ in placebo interventions for different values of $\beta_1(.5)$

	Mammen		Rademacher		Webb		Ana.	Reg.	Rank	FHH
	cv	se	cv	se	cv	se	se	se	score	cv
$\beta_1(.5) = 0$ (size)	.072	.084	.077	.093	.085	.095	.277	.284	.098	.311
$\beta_1(.5) = 5$ (power)	.413	.427	.449	.446	.450	.450	.701	.714	.516	.727

actually attended a small class. Next, I reran the QR in (5.1) and tested, at the 5% level, the correct null hypothesis that the coefficient on *small* is zero at the median, $\mathrm{H}_0\colon \beta_1(.5) = 0$, using the same methods as in Experiment 4.1. The rejection frequencies in the first line of results in Table 5 show the outcome of repeating this process 1,000 times. The bootstraps were again based on $m = 299$ simulations. As can be seen, the wild gradient bootstrap test from Algorithm 3.4 with the Mammen distribution outperformed all other methods of inference, some by a very large margin. Still, the test over-rejected slightly. This can be attributed to the fact that the treatment effect is now identified off of comparisons within only 18 instead of 79 schools, which makes the estimation problem much more challenging than in the actual data. The size of the tests in the placebo experiment can, in that sense, be viewed as an upper bound for the size of the tests in the original sample.

I also investigated power by increasing the percentile scores of all students in the randomly drawn small classes of the placebo experiment by 5. This increase is of the same order of magnitude as the estimated treatment effect at the median in the actual sample. Then I repeatedly tested the incorrect hypothesis $\mathrm{H}_0\colon \beta_1(.5) = 0$ (the correct value is $\beta_1(.5) = 5$) with the same experimental setup as before. The results are shown in the second line of Table 5. Despite the now much smaller sample, the wild gradient bootstrap was able to reject the null in a large number of cases. The other methods rejected more often, but this was likely driven by their size distortion. Notable here is the high power of the Wang and He (2007) rank score test despite its relatively mild over-rejection under the null. □

The large differences in the finite-sample size of the methods of inference considered in the preceding experiment can be attributed to the within-cluster dependence in the data. This is also supported by a back-of-the-envelope comparison of the results here to the Monte Carlo experiments in Section 4. For the Monte Carlo DGP (4.1), the within-cluster correlation coefficient of the outcome variable can be shown to be approximately ϱ. For the Project STAR data, the Karlin, Cameron, and Williams (1981) intraclass correlation coefficient

$$\hat{\varrho}_n := \frac{\sum_{i=1}^{n}\sum_{k=1}^{c_i}\sum_{l\neq k}(Y_{ik}-\bar{Y}_n)(Y_{il}-\bar{Y}_n)/(c_i-1)}{\sum_{i=1}^{n}\sum_{k=1}^{c_i}(Y_{ik}-\bar{Y}_n)^2}, \quad \text{where} \quad \bar{Y}_n = \frac{\sum_{i=1}^{n}\sum_{k=1}^{c_i} Y_{ik}}{\sum_{i=1}^{n} c_i},$$

of the percentile score is .319. This is a consistent estimate of the within-cluster correlation coefficient of the percentile score as long as both its mean and within-cluster covariance structure are identical across clusters. (Neither of these conditions is needed for any of the theoretical results in this paper.) At $\varrho = \hat{\varrho}_n$, the results of Experiments 4.1 and 5.1 are quite similar, with the exception that the rank score test performed much better in Experiment 5.1 than the test based on analytical cluster-robust standard errors.

Finally, before concluding this section, Figure 4 illustrates the difference between a 95% pointwise confidence interval based on a Powell sandwich estimator (as described in Experiment 4.1) that does not control for within-cluster correlation (dotted lines), the

wild gradient bootstrap confidence interval shown in Figure 3 (grey), and a 95% wild bootstrap confidence band for the entire coefficient function of *small* weighted by the bootstrap covariance matrix (dashed). As can be seen from the size of the grey area, not accounting for the possibility of peer effects and unobserved teacher characteristics via cluster-robust inference appears to give a false sense of precision at most quantiles. However, as the confidence band shows, we can conclude that the effect of the small class size is significantly positive over a large part of the support of the score distribution.

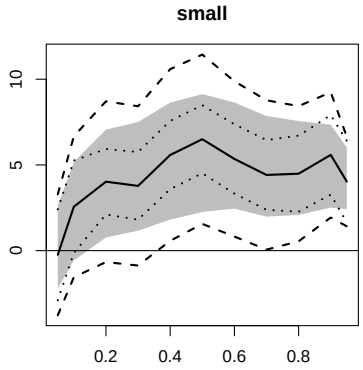

Figure 4: QR coefficient estimate and pointwise confidence interval for *small* from Figure 3, a 95% pointwise confidence interval not robust to within-cluster correlation (dotted lines), and a 95% wild bootstrap confidence band for the entire coefficient function (dashed).

6. Conclusion

In this paper I develop a wild bootstrap procedure for cluster-robust inference in linear QR models. I show that the bootstrap leads to asymptotically valid inference on the entire QR process in a setting with a large number of small, heterogeneous clusters and provides consistent estimates of the asymptotic covariance function of that process. The proposed bootstrap procedure is easy to implement and performs well even when the number of clusters is much smaller than the sample size. A brief application to Project STAR data is provided. It is still an open question how cluster-level fixed effects that correspond to the intuitive notion of identifying parameters from within-cluster variation can fit into the present framework; this is currently under investigation by the author. Another question is if the jackknife can improve on the bootstrap in the current context; recent results by Portnoy (2014) for censored regression quantiles suggest this possibility.

Appendix

A. Auxiliary Definitions and Results

I first introduce some notation and definitions that are used throughout the remainder of the paper. Then I state some auxiliary results. All proofs can be found in the next section.

Notation. For vectors a and b, I will occasionally write (a, b) instead of $(a^\top, b^\top)^\top$ if the dimensions are not essential. Take $(Y_{ik}, X_{ik}) = 0$ for $c_i < k \leq c_{\max}$ whenever $c_i < c_{\max}$ and let $Y_i = (Y_{i1}, \ldots, Y_{ic_{\max}})$ and $X_i = (X_{i1}^\top, \ldots, X_{ic_{\max}}^\top)^\top$. Let $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n (f(Y_i, X_i) - \mathrm{E}\, f(Y_i, X_i))$ be the empirical process evaluated at some function f and let $\mathbb{E}_n f = n^{-1} \sum_{i=1}^n f(Y_i, X_i)$ be the empirical average at f. I will frequently use the notation $|\,\mathbb{G}_n f|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\,\mathbb{G}_n f|$ for functional classes and $|f_\theta|_\Theta := \sup_{\theta \in \Theta} |f_\theta|$ for functions indexed by parameters. Define

$$m_{\beta,\tau}(Y_i, X_i) = \sum_{k=1}^{c_{\max}} \rho_\tau(Y_{ik} - X_{ik}^\top \beta), \quad z_{\beta,\tau}(Y_i, X_i) = \sum_{k=1}^{c_{\max}} \psi_\tau(Y_{ik} - X_{ik}^\top \beta) X_{ik},$$
$$g_{\beta_1,\tau_1,\beta_2,\tau_2}(Y_i, X_i) = \sum_{k=1}^{c_{\max}} \sum_{l=1}^{c_{\max}} \psi_{\tau_1}(Y_{ik} - X_{ik}^\top \beta_1) \psi_{\tau_2}(Y_{il} - X_{il}^\top \beta_2) X_{ik} X_{il}^\top$$

and the corresponding classes

$$\mathcal{M}_\delta = \{m_{\beta,\tau} - m_{\beta(\tau),\tau} : |\beta - \beta(\tau)| \leq \delta, \beta \in B, \tau \in T\}, \quad \mathcal{Z} = \{z_{\beta,\tau} : \beta \in B, \tau \in T\},$$
$$\mathcal{G} = \{g_{\beta_1,\tau_1,\beta_2,\tau_2} : (\beta_1, \tau_1, \beta_2, \tau_2) \in B \times T \times B \times T\}.$$

Write the jth coordinate projection as $x = (x_1, \ldots, x_j, \ldots, x_d) \mapsto \pi_j(x) = x_j$. Define a pseudometric ϱ on $\mathcal{Z}$ by

$$\varrho\big(z_{\beta,\tau}, z_{\beta',\tau'}\big) = \max_{1 \leq j \leq d} \sup_{n \geq 1} \big(\mathbb{E}_n\, \mathrm{E}(\pi_j \circ z_{\beta,\tau} - \pi_j \circ z_{\beta',\tau'})^2\big)^{1/2}.$$

Denote by π_{jh} the function that picks out the entry in the jth row and hth column of a matrix. For a matrix A, denote the Frobenius norm by $|A| = \mathrm{trace}^{1/2}(AA^\top)$; if A is a vector, this is the Euclidean norm. Let $\lambda_{\min}(A)$ be the smallest eigenvalue of a symmetric matrix A. Let $H_n(\beta) = n^{-1} \sum_{i=1}^n \sum_{k=1}^{c_i} \mathrm{E}\, f_{ik}(X_{ik}^\top \beta \mid X_{ik}) X_{ik} X_{ik}^\top$ and note that $J_n(\tau) = H_n(\beta(\tau))$. Define the mean value

$$I_n(\beta, \tau) = \int_0^1 H_n\Big(\beta(\tau) + t\big(\beta - \beta(\tau)\big)\Big)\, dt,$$

where the integral is taken componentwise. For scalars a and b, the notation $a \lesssim b$ means a is bounded by an absolute constant times b. □

Some expressions above and below may be non-measurable; probability and expectation of these expressions are understood in terms of outer probability and outer expectation (see, e.g. van der Vaart and Wellner, 1996, p. 6). Application of Fubini's theorem to such expectations requires additional care. A measurability condition that restores the Fubini theorem for independent non-identically distributed (inid) data is the "almost measurable Suslin" condition of Kosorok (2003). It is satisfied in all applications below.

A.1 Lemma. *If Assumptions 2.1-2.3 are satisfied, then* $\{z_{\beta(\tau),\tau} : \tau \in T\}$, $\{z_{\beta,\tau} : \beta \in B, \tau \in T\}$ *and* $\{m_{\beta,\tau} - m_{\beta(\tau),\tau} : \beta \in B, \tau \in T\}$ *are almost measurable Suslin.*

The following lemmas are used in the proofs of the results stated in the main text.

A.2 Lemma. *Suppose Assumptions 2.1-2.3 hold. Then,*
(i) $\sup_{n\geq 1} \mathrm{E}\,|\mathbb{G}_n m|^q_{\mathcal{M}_\delta} \lesssim \delta^q$,
(ii) $\sup_{n\geq 1} \mathrm{E}\,|\mathbb{W}_n(\beta,\tau)|^q_{B\times T} < \infty$, *and*
(iii) $\sup_{n\geq 1} \mathrm{E}\,|\mathbb{G}_n z|^q_{\mathcal{Z}} < \infty$.

A.3 Lemma. *If Assumptions 2.1-2.4 are satisfied, then* $|\beta - \beta(\tau)|^2 \lesssim M_n(\beta,\tau) - M_n(\beta(\tau),\tau)$ *for all* $\beta \in B$ *and all* $\tau \in T$.

A.4 Lemma. *Suppose Assumptions 2.1-2.4 are true. Then, for all* $0 < p < q$,
(i) $\sup_{n\geq 1} \mathrm{E}\,|\sqrt{n}(\hat\beta^*_n(\tau) - \hat\beta_n(\tau))|^p_T < \infty$,
(ii) $\sup_{n\geq 1} \mathrm{E}\,|\sqrt{n}(\hat\beta^*_n(\tau) - \beta(\tau))|^p_T < \infty$, *and*
(iii) $\sup_{n\geq 1} \mathrm{E}\,|\sqrt{n}(\hat\beta_n(\tau) - \beta(\tau))|^p_T < \infty$.

A.5 Lemma. *Suppose Assumptions 2.1-2.3 hold with* $q \geq 3$. *Then* $I_n(\beta,\tau)$ *is invertible and, uniformly in* n, $|I_n^{-1}(\beta,\tau) - J_n^{-1}(\tau)| \lesssim |\beta - \beta(\tau)|$ *for all* $\beta \in B$ *and* $\tau \in T$.

A.6 Lemma. *If Assumptions 2.1-2.4 hold with* $q \geq 3$, *then* $|J_n(\tau) - J(\tau)|_T$ *and* $|J_n^{-1}(\tau) - J^{-1}(\tau)|_T$ *converge to zero.*

A.7 Lemma. *If Assumptions 2.1-2.3 hold, then*
(i) $\varrho(z_{\beta,\tau}, z_{\beta',\tau'}) \lesssim |((\beta-\beta')^\top, \tau - \tau')|^{(q-2)/(2q)}$ *for all* $\beta, \beta' \in B$ *and* $\tau, \tau' \in T$, *and*
(ii) $z \mapsto \mathbb{G}_n z$ *is* ϱ*-stochastically equicontinuous on* $\mathcal{Z}$, *i.e., for all* $\varepsilon, \eta > 0$, *there is a* $\delta > 0$ *such that*

$$\limsup_{n\to\infty} \mathrm{P}\left(\sup_{\varrho(z_{\beta,\tau}, z_{\beta',\tau'})<\delta} |\,\mathbb{G}_n z_{\beta,\tau} - \mathbb{G}_n z_{\beta',\tau'}| > \eta \right) < \varepsilon.$$

A.8 Lemma. *If Assumptions 2.1-2.3 hold, then* $|\,\mathbb{E}_n(g - \mathrm{E}\,g)|_{\mathcal{G}} \to^{\mathrm{P}} 0$.

B. Proofs

Proof of Lemma A.1. To verify the almost measurable Suslin property for $m_{\beta(\tau),\tau}$, I start by establishing two useful facts: First, because $|1\{a < b\} - 1\{a < c\}| \leq 1\{|a-b| < |b-c|\}$ for $a, b, c \in \mathbb{R}$, each realization of Y_{ik} and X_{ik} satisfies

$$|1\{Y_{ik} < X_{ik}^\top \beta(\tau)\} - 1\{Y_{ik} < X_{ik}^\top \beta(\tau')\}| \leq 1\{|Y_{ik} - X_{ik}^\top \beta(\tau)| < |X_{ik}||\beta(\tau) - \beta(\tau')|\}.$$

Second, the eigenvalues of $J_n(\tau)$ are bounded away from zero uniformly in τ and n by Assumption 2.3. The same assumption and the inverse function theorem applied to $\sum_{i=1}^n \sum_{k=1}^{c_i} \mathrm{E}(\tau - 1\{Y_{ik} < X_{ik}^\top \beta(\tau)\}) X_{ik} = 0$ then give $d\beta(\tau)/d\tau = J_n^{-1}(\tau) n^{-1} \sum_{i=1}^n \sum_{k=1}^{c_i} \mathrm{E}\, X_{ik}$. Because the Frobenius norm is also the 2-Schatten norm, we have $|J_n^{-1}(\tau)| \le \sqrt{d \lambda_{\min}(J_n(\tau))^{-1}}$ and hence $d\beta(\tau)/d\tau$ is bounded uniformly in $\tau \in T$ by some $C > 0$. The mean-value theorem yields

$$|\beta(\tau) - \beta(\tau')| \le C|\tau - \tau'|, \qquad \tau, \tau' \in T. \tag{B.1}$$

For each $\tau \in T$, combine the preceding two displays with the Loève c_r inequality to obtain ($\mathbb{Q}$ here is the set of rationals)

$$\begin{aligned}
&\inf_{\tau' \in T \cap \mathbb{Q}} \mathbb{E}_n(\pi_j \circ z_{\beta(\tau),\tau} - \pi_j \circ z_{\beta(\tau'),\tau'})^2 \\
&\quad \lesssim \inf_{\tau' \in T \cap \mathbb{Q}} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{c_i} \big((\tau - \tau')^2 + 1\{|Y_{ik} - X_{ik}^\top \beta(\tau)| < |X_{ik}| C |\tau - \tau'|\} \big) \pi_j(X_{ik})^2.
\end{aligned}$$

The infimum on the right must be smaller than $n^{-1} \sum_{i=1}^n \sum_{k=1}^{c_i} (\delta^2 + 1\{|Y_{ik} - X_{ik}^\top \beta(\tau)| < |X_{ik}| C \delta\}) \pi_j(X_{ik})^2$ for every $\delta > 0$. Conclude that the infimum is $n^{-1} \sum_{i=1}^n \sum_{k=1}^{c_i} 1\{|Y_{ik} - X_{ik}^\top \beta(\tau)| < 0\} \pi_j(X_{ik})^2 = 0$. This does not change if we take suprema over $\tau \in T$ on both sides of the display. If follows that

$$\mathrm{P}\left(\sup_{\tau \in T} \inf_{\tau' \in T \cap \mathbb{Q}} \mathbb{E}_n(\pi_j \circ z_{\beta(\tau),\tau} - \pi_j \circ z_{\beta(\tau'),\tau'})^2 > 0 \right) = 0, \qquad 1 \le j \le d,$$

which makes $\{z_{\beta(\tau),\tau} : \tau \in T\}$ almost measurable Suslin by Lemma 2 of Kosorok (2003). Nearly identical calculations verify the same property for $\{z_{\beta,\tau} : \beta \in B, \tau \in T\}$. Because

$$\begin{aligned}
&\rho_\tau(y - x^\top \beta) - \rho_\tau\big(y - x^\top \beta(\tau)\big) - \big(\rho_{\tau'}(y - x^\top \beta') - \rho_{\tau'}\big(y - x^\top \beta(\tau')\big)\big) \\
&\qquad \le |x| \big(|\tau - \tau'| \big(\beta - \beta(\tau)\big) + |\beta - \beta'| + |\beta(\tau) - \beta(\tau')| \big),
\end{aligned}$$

the process $\{m_{\beta,\tau} - m_{\beta(\tau),\tau} : \beta \in B, \tau \in T\}$ is almost measurable Suslin as well. □

Proof of Lemma A.2. (i) The subgraph of a real-valued function f is defined as $\{(x,t) : f(x) > t\}$. Write the subgraph of $\rho(a-b) - \rho(a)$ as

$$\begin{aligned}
&(\{a \ge 0\} \cap \{a \ge b\} \cap \{b < -t/\tau\}) \cup (\{a \ge 0\} \cap \{a < b\} \cap \{(1-\tau)b - a > t\}) \\
&\quad \cup (\{a < 0\} \cap \{a < b\} \cap \{a - b < t/(\tau - 1)\}) \cup (\{a < 0\} \cap \{a \ge b\} \cap \{a - \tau b > t\}).
\end{aligned}$$

Now take $a = y - x^\top \beta(\tau)$ and $b = x^\top(\beta - \beta(\tau))$ so that, e.g., $\{a > 0\} = \{(y,x) : y - x^\top \beta(\tau) > 0\}$. Hence,

$$\begin{aligned}
&\{(1-\tau)b - a > t\} = \big\{(y,x) : y - x^\top \big(\tau \beta(\tau) + (1-\tau)\beta\big) < -t\big\} \quad \text{and} \\
&\{a - \tau b > t\} = \big\{(y,x) : y - x^\top \big(\tau \beta + (1-\tau)\beta(\tau)\big) > t\big\}.
\end{aligned}$$

By convexity of B, the two collections of sets in the display, indexed by $\beta \in B$, $\tau \in T$, and $t \in \mathbb{R}$, are contained in the collection of sets $\{(y,x) : y - x^\top\beta < -t\}$ and $\{(y,x) : y - x^\top\beta > t\}$, respectively, indexed by $\beta \in B$ and $t \in \mathbb{R}$.

The collection of sets $\{v \in \mathbb{R}^{d+2} : v^\top\lambda \leq 0\}$ indexed by $\lambda \in \mathbb{R}^{d+2}$ is a Vapnik–Červonenkis (VC) class of sets (see van der Vaart and Wellner, 1996, Problem 2.6.14); the same holds for the collection $\{v \in \mathbb{R}^{d+2} : v^\top\lambda > 0\}$ by Lemma 2.6.17(i) of van der Vaart and Wellner. Because $B \subset \mathbb{R}^d$, each individual set in the subgraph above indexed by $\beta \in B$, $\tau \in T$, and $t \in \mathbb{R}$ is contained in one of these two VC classes. Subclasses of VC classes are VC classes themselves. Conclude from van der Vaart and Wellner's Lemma 2.6.17(ii) and (iii) that the subgraph above is a VC class. Therefore the map

$$\rho_\tau(y - x^\top\beta) - \rho_\tau\big(y - x^\top\beta(\tau)\big)$$

indexed by β and τ is a VC subgraph class. Sums of functions from VC subgraph classes do not necessarily form VC subgraph classes, but their uniform entropy numbers behave like those of VC subgraph classes if the envelopes are increased accordingly (Kosorok, 2008, p. 157). Because the absolute value of the preceding display is bounded above by $|x^\top(\beta - \beta(\tau))|$, we can use $\sum_{k=1}^{c_{\max}} |X_{ik}|\delta$ as an envelope for $m(Y_i, X_i)$ with $m \in \mathcal{M}_\delta$. The class $\mathcal{M}_\delta$ then has a finite uniform entropy integral in the sense of van der Vaart and Wellner (1996, condition (2.5.1), p. 127).

Use van der Vaart and Wellner's (1996) Theorem 2.14.1 (which applies to inid observations if the reference to their Lemma 2.3.1 is replaced by a reference to their Lemma 2.3.6 and my Lemma A.1 is used to justify their symmetrization argument) to deduce that

$$\mathrm{E}\,|\,\mathbb{G}_n\, m|^q_{\mathcal{M}_\delta} \lesssim \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left|\sum_{k=1}^{c_{\max}} |X_{ik}|\delta\right|^q \lesssim \delta^q \sup_{i,k} \mathrm{E}\,|X_{ik}|^q.$$

The right-hand side is finite by assumption, which completes the proof.

(ii) By the Loève c_r inequality, it suffices to show that each of the d elements of $\mathbb{W}_n$ has the desired property. Arguments similar to the ones given in the first part of the proof establish that the collection of functions $(y,x,w) \mapsto w\pi_j(x)\psi_\tau(y - x^\top\beta)$ indexed by $(\beta,\tau) \in B \times T$ is a VC subgraph class. As such, it satisfies Pollard's (1982) uniform entropy condition. By Theorem 3 of Andrews (1994, p. 2273), the class of functions with finite uniform entropy is stable under addition as long as the envelope function is increased accordingly. An appropriate envelope for the set of functions $\sum_{k=1}^{c_{\max}} W_i\pi_j(X_{ik})\psi_\tau(Y_{ik} - X_{ik}^\top\beta)$ indexed by $(\beta,\tau) \in B \times T$ is $F_j(W_i, X_i) = 2|W_i| \sum_{k=1}^{c_{\max}} |\pi_j(X_{ik})|$. In addition, the components of $\mathbb{W}_n(\beta,\tau)$ are almost measurable Suslin by Lemma A.1 and the bound

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i^2\left(\sum_{k=1}^{c_{\max}}\big(\psi_\tau(Y_{ik} - X_{ik}^\top\beta) - \psi_{\tau'}(Y_{ik} - X_{ik}^\top\beta')\big)\pi_j(X_{ik})\right)^2 \\ \leq \sqrt{n}\max_{1\leq i\leq n} W_i^2\,\mathbb{E}_n(\pi_j \circ z_{\beta,\tau} - \pi_j \circ z_{\beta',\tau'})^2.\end{aligned}$$

Conclude from van der Vaart and Wellner's (1996) Theorem 2.14.1 (see part (i) of the proof above) and independence of the bootstrap weights from the data that

$$\mathrm{E}\,|\pi_j \circ \mathbb{W}_n(\beta,\tau)|^q_{B\times T} \lesssim \mathrm{E}\,\mathbb{E}_n\, F_j^q \lesssim \mathrm{E}\,|W|^q \sup_{i,k} \mathrm{E}\,|X_{ik}|^q,$$

which is finite by assumption.

(iii) This follows from (ii) with $W \equiv 1$. □

Proof of Lemma A.3. By a Taylor expansion about $\beta(\tau)$,

$$M_n(\beta,\tau) - M_n(\beta(\tau),\tau) = M_n'(\beta(\tau),\tau)\big(\beta-\beta(\tau)\big) + \big(\beta-\beta(\tau)\big)^\top I_n(\beta,\tau)\big(\beta-\beta(\tau)\big)/2.$$

The first term on the right is zero because $\beta = \beta(\tau)$ minimizes $M_n(\beta,\tau)$. By the properties of Rayleigh quotients, the second term on the right is at least as large as $\lambda_{\min}(I_n(\beta,\tau))|\beta-\beta(\tau)|^2/2$. Assumption 2.3 implies that $\lambda_{\min}(H_n(\beta))$ is bounded away from zero uniformly in β and n. For every non-zero $a \in \mathbb{R}^d$, we must have $a^\top I_n(\beta,\tau)a \geq \inf_{\beta,n} a^\top H_n(\beta)a$ and therefore $\lambda_{\min}(I_n(\beta,\tau))$ is bounded away from zero as well, uniformly in β, τ, and n. □

Proof of Lemma A.4. This proof uses a (non-trivial) modification of the strategy of proof used by Kato (2011). Without loss of generality, take $p \geq 2$. Because B is bounded, for every $\varepsilon > 0$ there exists some $\tau_n^* \in T$ such that $|\sqrt{n}(\hat\beta_n^*(\tau) - \beta(\tau))|_T$ and $|\sqrt{n}(\hat\beta_n^*(\tau_n^*) - \beta(\tau_n^*))|$ differ at most by ε. Choose $\varepsilon < 2$. For every nonnegative integer N, the inequality $1\{a+b>c\} \leq 1\{2a>c\} + 1\{2b>c\}$ with $a,b,c \in \mathbb{R}$ then yields

$$\begin{aligned}\mathrm{P}\Big(\big|\sqrt{n}\big(\hat\beta_n^*(\tau)-\beta(\tau)\big)\big|_T > 2^{N+1}\Big) &\leq \mathrm{P}\Big(\varepsilon + \big|\sqrt{n}\big(\hat\beta_n^*(\tau_n^*)-\beta(\tau_n^*)\big)\big| > 2^{N+1}\Big)\\ &\leq \mathrm{P}\Big(\big|\sqrt{n}\big(\hat\beta_n^*(\tau_n^*)-\beta(\tau_n^*)\big)\big| > 2^{N}\Big).\end{aligned}$$

Define shells $S_{jn} = \{(\beta,\tau) \in B\times T : 2^{j-1} < \sqrt{n}|\beta-\beta(\tau)| \leq 2^j\}$ for integer $j \geq 1$. If the event in the second line of the preceding display occurs, then there exists some $j \geq N$ such that $(\hat\beta_n^*(\tau_n^*),\tau_n^*) \in S_{jn}$. Because $\hat\beta_n^*(\tau)$ minimizes $\mathbb{M}_n^*(\beta,\tau)$ for every $\tau \in T$, including τ_n^*, this implies $\inf_{(\beta,\tau)\in S_{jn}} \mathbb{M}_n^*(\beta,\tau) - \mathbb{M}_n^*(\beta(\tau),\tau) \leq 0$. The union bound then gives

$$\begin{aligned}&\mathrm{P}\Big(\big|\sqrt{n}\big(\hat\beta_n^*(\tau_n^*)-\beta(\tau_n^*)\big)\big| > 2^N\Big)\\ &\quad\leq \sum_{j\geq N} \mathrm{P}\Big(\inf_{(\beta,\tau)\in S_{jn}} \mathbb{M}_n^*(\beta,\tau) - \mathbb{M}_n^*(\beta(\tau),\tau) \leq 0\Big).\end{aligned} \tag{B.2}$$

Add and subtract to decompose $\mathbb{M}_n^*(\beta,\tau) - \mathbb{M}_n^*(\beta(\tau),\tau)$ into

$$M_n(\beta,\tau) - M_n(\beta(\tau),\tau) + \mathbb{G}_n(m_{\beta,\tau} - m_{\beta(\tau),\tau})/\sqrt{n} + \mathbb{W}_n(\tau)^\top\big(\beta-\beta(\tau)\big)\sqrt{n}.$$

By Lemma A.3, $M_n(\beta,\tau) - M_n(\beta(\tau),\tau) \geq c|\beta - \beta(\tau)|^2 \geq c2^{2j-2}/n$ on S_{jn} for some $c > 0$. For each j, we therefore have the inclusion

$$\Big\{ \inf_{(\beta,\tau)\in S_{jn}} \mathbb{M}_n^*(\beta,\tau) - \mathbb{M}_n^*(\beta(\tau),\tau) \leq 0 \Big\}$$
$$\subset \Big\{ \big|\mathbb{G}_n(m_{\beta,\tau} - m_{\beta(\tau),\tau})\big|_{S_{jn}} + \big|\mathbb{W}_n(\tau)^\top \big(\beta - \beta(\tau)\big)\big|_{S_{jn}} \geq c2^{2j-2}/\sqrt{n} \Big\}.$$

Similarly, $|\mathbb{W}_n(\tau)^\top(\beta - \beta(\tau))| \leq |\mathbb{W}_n(\tau)|2^j/\sqrt{n}$ on S_{jn} by the Cauchy–Schwarz inequality. This can be bounded further by the supremum of $|\mathbb{W}_n(\beta,\tau)|2^j/\sqrt{n}$ over $B \times T$. After slightly decreasing c, conclude that the right-hand side of (B.2) is at most

$$\sum_{j\geq N} \mathrm{P}\Big(\big|\mathbb{G}_n(m_{\beta,\tau} - m_{\beta(\tau),\tau})\big|_{S_{jn}} \geq c2^{2j}/\sqrt{n} \Big) + \sum_{j\geq N} \mathrm{P}\Big(|\mathbb{W}_n(\beta,\tau)|_{B\times T} \geq c2^j \Big). \quad \text{(B.3)}$$

Consider the supremum inside the first term. For $\delta_{jn} = 2^j/\sqrt{n}$, the supremum over S_{jn} does not exceed the supremum over $\mathcal{M}_{\delta_{jn}}$. Together with Lemma A.2(i) this yields

$$\mathrm{E}\,|\mathbb{G}_n(m_{\beta,\tau} - m_{\beta(\tau),\tau})|_{S_{jn}}^q \lesssim (2^j/\sqrt{n})^q.$$

The supremum inside the second term in (B.3) satisfies $\sup_{n\geq 1} \mathrm{E}\,|\mathbb{W}_n(\beta,\tau)|_{B\times T}^q < \infty$ by Lemma A.2(ii). Combine these results with the Markov inequality to see that (B.3) is bounded by a constant multiple of $\sum_{j\geq N} 2^{-qj} \lesssim 2^{-qN}$, uniformly in n. Take $N = \lfloor \log_2 t \rfloor$ for $t \geq 2$ and conclude from the bounds developed so far that

$$\mathrm{P}\Big(\big|\sqrt{n}\big(\hat\beta_n^*(\tau) - \beta(\tau)\big)\big|_T > 2t \Big) \lesssim t^{-q}.$$

The case $0 < t < 2$ can be absorbed into a constant. The preceding display is then valid for all $t > 0$. Tonelli's theorem and Lemma 1.2.2 of van der Vaart and Wellner (1996) now give

$$\mathrm{E}\big|\sqrt{n}\big(\hat\beta_n^*(\tau) - \beta(\tau)\big)\big|_T^p = 2^p p \int_0^\infty t^{p-1}\, \mathrm{P}\Big(\big|\sqrt{n}\big(\hat\beta_n^*(\tau) - \beta(\tau)\big)\big|_T > 2t \Big)\, dt,$$

which is finite as long as $p < q$.

A simpler, nearly identical argument establishes $\mathrm{E}\,|\sqrt{n}(\hat\beta_n(\tau) - \beta(\tau))|_T^p < \infty$ uniformly in n. The Loève c_r inequality completes the proof. □

Proof of Lemma A.5. Use sub-additivity of the Frobenius norm, Assumptions 2.3(iii) and (iv), and the mean-value theorem to write

$$|I_n(\beta,\tau) - J_n(\tau)| \lesssim \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{c_i} \int_0^1 t\, dt\, \mathrm{E}\big|X_{ik}^\top\big(\beta - \beta(\tau)\big)\big||X_{ik}|^2 \lesssim |\beta - \beta(\tau)| \sup_{i,k} \mathrm{E}\,|X_{ik}|^3.$$

To transform this into a bound on the difference of inverses, note that the eigenvalues of $I_n(\beta,\tau)$ and $J_n(\tau)$ are bounded away from zero uniformly in β, τ, and n by Assumption 2.3. Since $A^{-1}-B^{-1}=B^{-1}(B-A)A^{-1}$ for any two nonsingular matrices A and B, conclude from the preceding display, sub-multiplicativity of the Frobenius norm, and Assumption 2.3(i) (with $q\geq 3$) that $|I_n^{-1}(\beta,\tau)-J_n^{-1}(\tau)|\lesssim |I_n^{-1}(\beta,\tau)||J_n^{-1}(\tau)||\beta-\beta(\tau)|$. The right-hand side is finite because $|I_n^{-1}(\beta,\tau)|\leq\sqrt{d\lambda_{\min}(I_n(\beta,\tau))^{-1}}$ and $|J_n^{-1}(\tau)|\leq\sqrt{d\lambda_{\min}(J_n(\tau))^{-1}}$ due to the fact that the Frobenius norm is also the 2-Schatten norm. □

Proof of Lemma A.6. The metric space $(T,|\cdot|)$ is totally bounded because T is bounded. As in the proof of Lemma A.5, $|J_n(\tau)-J_n(\tau')|\lesssim|\beta(\tau)-\beta(\tau')|$. Conclude from (B.1) that $J_n(\tau)$ is asymptotically uniformly equicontinuous. The pointwise convergence given in Assumption 2.4 is therefore uniform by a version of the Arzelà-Ascoli theorem (see, e.g., Davidson, 1994, Theorem 21.7, p. 335). The result for the difference of inverses is deduced as in Lemma A.5. □

Proof of Lemma A.7. (i) Use the Jensen inequality, the Loève c_r inequality and $\psi_{\tau'}=\tau'-\tau+\psi_\tau$, Assumption 2.3(i) and the Hölder inequality with exponents $q/2$ and $q/(q-2)$, the fact that $|1\{a\leq b\}-1\{a\leq c\}|\leq 1\{|a-b|\leq|b-c|\}$ for $a,b,c\in\mathbb{R}$, and finally Assumption 2.3(iii) and the mean-value theorem to see

$$
\begin{aligned}
&\big(\mathbb{E}_n\,\mathrm{E}(\pi_j\circ z_{\beta,\tau}-\pi_j\circ z_{\beta',\tau'})^2\big)^{1/2}\\
&\quad\lesssim \mathbb{E}_n\max_{1\leq k\leq c_i}\Big(\mathrm{E}\,\pi_j(X_{ik})^2\big(\psi_\tau(Y_{ik}-X_{ik}^\top\beta)-\psi_{\tau'}(Y_{ik}-X_{ik}^\top\beta')\big)^2\Big)^{1/2}\\
&\quad\lesssim |\tau-\tau'|+\mathbb{E}_n\max_{1\leq k\leq c_i}\Big(\mathrm{E}\big|\psi_\tau(Y_{ik}-X_{ik}^\top\beta)-\psi_\tau(Y_{ik}-X_{ik}^\top\beta')\big|^{2q/(q-2)}\Big)^{(q-2)/(2q)}\\
&\quad\lesssim |\tau-\tau'|+\mathbb{E}_n\max_{1\leq k\leq c_i}\big(\mathrm{E}\,1\{|Y_{ik}-X_{ik}^\top\beta|\leq|X_{ik}||\beta-\beta'|\}\big)^{(q-2)/(2q)}\\
&\quad\lesssim |\tau-\tau'|+\mathbb{E}_n\max_{1\leq k\leq c_i}\big(|\beta-\beta'|\,\mathrm{E}\,|X_{ik}|\big)^{(q-2)/(2q)}.
\end{aligned}
$$

Because $|\tau-\tau'|<1$, we have $|\tau-\tau'|\leq|\tau-\tau'|^{(q-2)/(2q)}$. Conclude from the Hölder inequality that the extreme right-hand side of the display does not exceed a constant multiple of $|((\beta-\beta')^\top,\tau-\tau')|^{(q-2)/(2q)}$. Now take suprema over n and then maxima over $1\leq j\leq d$.

(ii) In view of the proof of Lemma A.2(ii), the collection of functions $(y,x)\mapsto\pi_j(x)\psi_\tau(y-x^\top\beta)$ indexed by $(\beta,\tau)\in B\times T$ is a VC subgraph class. Sums of functions with finite uniform entropy still have finite uniform entropy if the envelope is increased accordingly. An appropriate envelope for $\pi_j\circ z_{\beta,\tau}$ is $2\sum_{k=1}^{c_{\max}}|\pi_j(X_{ik})|$. Because this envelope is L_q-integrable, stochastic equicontinuity follows from Andrews' (1994, Theorem 1, p. 2269) modification of Pollard's (1990, p. 53) functional central limit theorem; see also Kosorok (2003, Theorem 1). The process is suitably measurable because $\{z_{\beta,\tau}:\tau\in T\}$ is almost measurable Suslin by Lemma A.1. □

Proof of Lemma A.8. As noted in the proof of Lemma A.2(ii), the collection of functions $(y, x) \mapsto \pi_j(x)\psi_\tau(y - x^\top\beta)$ indexed by $\beta \in B$ and $\tau \in T$ is a VC subgraph class. Finite sums of products of such functions need not be VC subgraph, but their uniform entropy numbers behave like those of VC subgraph classes as long as the envelope is increased accordingly; see Kosorok (2008, pp. 157-158). An appropriate envelope for $\pi_{jh} \circ g$ is

$$4 \sum_{k=1}^{c_{\max}} \sum_{l=1}^{c_{\max}} |\pi_j(X_{ik})\pi_h(X_{il})|, \qquad j, h \in \{1, \dots, d\}.$$

I now verify that $\mathbb{E}_n(g - \mathrm{E}\, g)$ with $g \in \mathcal{G}$ satisfies the conditions of the uniform law of large numbers of Pollard (1990, Theorem 8.2, p. 39). By the Jensen inequality, the convergence occurs if it occurs for every entry of the matrix $\mathbb{E}_n(g - \mathrm{E}\, g)$. Hence, suppose for simplicity that $d = 1$; otherwise argue separately for each entry. The moment condition (i) on the envelope in Pollard's theorem holds immediately by Assumption 2.3. The other condition involves the packing numbers or, equivalently, the covering numbers of $\mathcal{G}$; see van der Vaart and Wellner (1996, p. 98) for definitions and the equivalence result. Because the uniform entropy numbers of $\mathcal{G}$ relative to L_r behave like those of VC subgraph classes for all $r \geq 1$, Pollard's condition (ii) is easily satisfied in view of the discussion in van der Vaart and Wellner (1996, p. 125). The uniform law of large numbers now follows if $\{g : g \in \mathcal{G}\}$ is suitably measurable to support the symmetrization argument used in Pollard's proof. A sufficient condition is that $\{g : g \in \mathcal{G}\}$ is almost measurable Suslin, which follows from

$$\begin{aligned}
&\mathbb{E}_n(\pi_{jh} \circ g_{\beta_1,\tau_1,\beta_2,\tau_2} - \pi_{jh} \circ g_{\beta_1',\tau_1',\beta_2',\tau_2'})^2 \\
&\quad \lesssim \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c_i} \sum_{l=1}^{c_i} \big((\tau_1 - \tau_1')^2 + 1\{|Y_{ik} - X_{ik}^\top \beta_1| < |X_{ik}||\beta_1 - \beta_1'|\} \\
&\qquad\qquad + (\tau_2 - \tau_2')^2 + 1\{|Y_{il} - X_{il}^\top \beta_2| < |X_{il}||\beta_2 - \beta_2'|\} \big) \pi_{jh}(X_{ik} X_{il}^\top)^2
\end{aligned}$$

and the argument used in the proof of Lemma A.1, mutatis mutandis. □

Proof of Theorem 3.3. (i) Recall that $\{\mathbb{Z}(\tau) : \tau \in T\}$ is the d-dimensional Gaussian process described in Theorems 2.5 and 3.1. Write $Z_n^*(\tau) := \sqrt{n}(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau))$. Because $\mathrm{E}^*\, Z_n^*(\tau) Z_n^*(\tau')^\top$ converges in probability to $\mathrm{E}\, \mathbb{Z}(\tau)\mathbb{Z}(\tau')^\top$ if and only if each coordinate converges, assume for simplicity that $d = 1$; otherwise, treat each coordinate individually. Then, by Theorem 3.1, $(Z_n^*(\tau), Z_n^*(\tau')) \rightsquigarrow (\mathbb{Z}(\tau), \mathbb{Z}(\tau'))$ in probability in $\mathbb{R}^2$ at every $\tau, \tau' \in T$. By arguing along subsequences, conclude from the portmanteau lemma (see, e.g., van der Vaart, 1998, Lemma 2.2, p. 6) that in the following we can use the fact that $\mathrm{E}^*\, f(Z_n^*(\tau), Z_n^*(\tau')) \to^{\mathrm{P}} \mathrm{E}\, f(\mathbb{Z}(\tau), \mathbb{Z}(\tau'))$ for every continuous and bounded function f.

Without loss of generality, assume $Z_n^*(\tau) Z_n^*(\tau')$ is nonnegative; if not, split into positive and negative parts and argue separately. Now, for a given $\Delta > 0$,

$$|\,\mathrm{E}^*\, Z_n^*(\tau) Z_n^*(\tau') - \mathrm{E}^*\, \mathbb{Z}(\tau)\mathbb{Z}(\tau')| \leq \mathrm{E}^*\, Z_n^*(\tau) Z_n^*(\tau') - \mathrm{E} \min\{Z_n^*(\tau) Z_n^*(\tau'), \Delta\}$$

$$+ |\,\mathrm{E}^* \min\{Z_n^*(\tau)Z_n^*(\tau'), \Delta\} - \mathrm{E}\min\{\mathbb{Z}(\tau)\mathbb{Z}(\tau'), \Delta\}|$$
$$+ |\,\mathrm{E}\min\{\mathbb{Z}(\tau)\mathbb{Z}(\tau'), \Delta\} - \mathrm{E}\,\mathbb{Z}(\tau)\mathbb{Z}(\tau')|.$$

Because $(z, z') \mapsto \min\{zz', \Delta\}$ is continuous and bounded for $zz' \geq 0$, the second term on the right converges to zero in probability as $n \to \infty$. The first term on the right does not exceed $\mathrm{E}^*\, Z_n^*(\tau)Z_n^*(\tau')1\{Z_n^*(\tau)Z_n^*(\tau') > \Delta\}$. For $\varepsilon > 0$, the law of iterated expectations and the Cauchy-Schwarz inequality give

$$\mathrm{E}\,\mathrm{E}^*\, Z_n^*(\tau)Z_n^*(\tau')1\{Z_n^*(\tau)Z_n^*(\tau') > \Delta\} \leq \sup_{n\geq 1} \sqrt{\mathrm{E}\, Z_n^*(\tau)^{2(1+\varepsilon)}\, \mathrm{E}\, Z_n^*(\tau')^{2(1+\varepsilon)}}\Delta^{-\varepsilon}.$$

Note that the expectation on the right operates on both the data and the bootstrap distribution. As long as $2(1 + \varepsilon) < q$, the right-hand side is finite by Lemma A.4 and converges to zero as $\Delta \to \infty$. A similar argument applies to the third term. The Markov inequality completes the proof.

(ii) Apply mean-value expansions to $\sqrt{n}M_n'(\beta(\tau), \tau) = 0$ to deduce

$$\begin{aligned}\sqrt{n}\big(\hat\beta_n^*(\tau) - \hat\beta_n(\tau)\big) &= \sqrt{n}\big(\hat\beta_n^*(\tau) - \beta(\tau)\big) - \sqrt{n}\big(\hat\beta_n(\tau) - \beta(\tau)\big) \\ &= I_n^{-1}(\hat\beta_n^*(\tau), \tau)\sqrt{n}M_n'(\hat\beta_n^*(\tau), \tau) - I_n^{-1}(\hat\beta_n(\tau), \tau)\sqrt{n}M_n'(\hat\beta_n(\tau), \tau).\end{aligned}$$

After adding and subtracting with $J_n^{-1}(\tau)$, this becomes

$$\begin{aligned}&J_n^{-1}(\tau)\sqrt{n}\big(M_n'(\hat\beta_n^*(\tau), \tau) - M_n'(\hat\beta_n(\tau), \tau)\big) \\ &\quad + \big(I_n^{-1}(\hat\beta_n^*(\tau), \tau) - J_n^{-1}(\tau)\big)\sqrt{n}M_n'(\hat\beta_n^*(\tau), \tau) \\ &\quad + \big(I_n^{-1}(\hat\beta_n(\tau), \tau) - J_n^{-1}(\tau)\big)\sqrt{n}M_n'(\hat\beta_n(\tau), \tau).\end{aligned}$$

Denote the first term by $G_n(\tau)$ and the remaining two terms by $R_n(\tau)$. Then the distance between the bootstrap and population covariance functions can be written as

$$|\hat V_n^*(\tau, \tau') - V(\tau, \tau')|_{T\times T} = \big|\mathrm{E}^*\big(G_n(\tau) + R_n(\tau)\big)\big(G_n(\tau') + R_n(\tau')\big)^\top - V(\tau, \tau')\big|_{T\times T}.$$

Monotonicity of the expectation operator and sub-multiplicativity of the Frobenius norm yield $|\,\mathrm{E}^*\, G_n(\tau)R_n(\tau')^\top|_{T\times T} \leq \mathrm{E}^*\,|G_n(\tau)R_n(\tau')^\top|_{T\times T} \leq \mathrm{E}^*\,|G_n(\tau)|_T|R_n(\tau)|_T$. This remains true when G_n is replaced by R_n. Hence, the preceding display is at most

$$\big|\mathrm{E}^*\, G_n(\tau)G_n(\tau')^\top - V(\tau, \tau')\big|_{T\times T} + 2\,\mathrm{E}^*\,|G_n(\tau)|_T|R_n(\tau)|_T + \mathrm{E}^*\,|R_n(\tau)|_T^2.$$

I will now argue that the second and third term converge to zero in probability. By the Loève c_r inequality and Lemma A.5, $\mathrm{E}^*\,|R_n(\tau)|_T^2$ is bounded above by 2 times

$$\mathrm{E}^*\big|\sqrt{n}\big(\hat\beta_n^*(\tau) - \beta(\tau)\big)\big|_T^2|M_n'(\hat\beta_n^*(\tau), \tau)|_T^2 + \big|\sqrt{n}\big(\hat\beta_n(\tau) - \beta(\tau)\big)\big|_T^2|M_n'(\hat\beta_n(\tau), \tau)|_T^2.$$

Theorem 3.3 of Koenker and Bassett (1978) yields

$$|\,\mathbb{E}_n\, z_{\hat\beta_n^*(\tau),\tau} - \mathbb{W}_n(\tau)| \lesssim \max_{i\le n, k\le c_i} |X_{ik}| + \sqrt{n}|\mathbb{W}_n(\tau)|$$

uniformly in $\tau \in T$ and therefore

$$\begin{aligned}|\sqrt{n}M_n'(\hat\beta_n^*(\tau),\tau)|_T^4 &= |\mathbb{G}_n z_{\hat\beta_n^*(\tau),\tau} - \sqrt{n}\,\mathbb{E}_n\, z_{\hat\beta_n^*(\tau),\tau} + \mathbb{W}_n(\tau) - \mathbb{W}_n(\tau)\big|_T^4 \\ &\lesssim |\mathbb{G}_n z_{\beta,\tau}|_{B\times T}^4 + \Big|n^{-1/2}\max_{i\le n,k\le c_i}|X_{ik}| + |\mathbb{W}_n(\tau)|\Big|_T^4 + |\mathbb{W}_n(\tau)|_T^4 \\ &\lesssim |\mathbb{G}_n z|_{\mathcal{Z}}^4 + n^{-2}\max_{i\le n,k\le c_i}|X_{ik}|^4 + |\mathbb{W}_n(\tau)|_T^4.\end{aligned}$$

The first term on the far right of the display satisfies $\mathrm{E}\,|\mathbb{G}_n z|_{\mathcal{Z}}^4 < \infty$ uniformly in n by Lemma A.2(iii), the second satisfies $n^{-2}\,\mathrm{E}\max_{i\le n,k\le c_i}|X_{ik}|^4 \le n^{-1}\sup_{i,k}\mathrm{E}\,|X_{ik}|^4$ by Pisier's inequality (see, e.g., van der Vaart and Wellner, 1996, Problem 2.2.8, p. 105), and the third satisfies $\mathrm{E}\,|\mathbb{W}_n(\tau)|_T^4 < \infty$ uniformly in n by Lemma A.2(ii). Conclude that $\mathrm{E}\,|M_n'(\hat\beta_n^*(\tau),\tau)|_T^4 = O(n^{-2})$. Repeat the argument above with $\mathbb{W}_n(\tau) \equiv 0$ and $\hat\beta_n(\tau)$ instead of $\hat\beta_n^*(\tau)$ to also establish $\mathrm{E}\,|M_n'(\hat\beta_n(\tau),\tau)|_T^4 = O(n^{-2})$. The Markov inequality, Lemma 1.2.6 of van der Vaart and Wellner (1996, p. 11), the Cauchy-Schwarz inequality, and Lemma A.4 now imply $\mathrm{E}^*\,|R_n(\tau)|_T^2 \to^{\mathrm{P}} 0$.

Further, decompose $J_n(\tau)G_n(\tau)$ into

$$\mathbb{G}_n z_{\hat\beta_n^*(\tau),\tau} - \mathbb{G}_n z_{\hat\beta_n(\tau),\tau} + \sqrt{n}\,\mathbb{E}_n\, z_{\hat\beta_n(\tau),\tau} - \big(\sqrt{n}\,\mathbb{E}_n\, z_{\hat\beta_n^*(\tau),\tau} - \mathbb{W}_n(\tau)\big) - \mathbb{W}_n(\tau). \quad \text{(B.4)}$$

The same arguments as before show $\mathrm{E}^*\,|J_n(\tau)G_n(\tau)|_T^2 = O_{\mathrm{P}}(1)$ and thus also $\mathrm{E}^*\,|G_n(\tau)|_T^2 \le |J_n^{-1}(\tau)|_T^2\,\mathrm{E}^*\,|J_n(\tau)G_n(\tau)|_T^2 = O_{\mathrm{P}}(1)$ by Lemma A.6. Finally, apply the Cauchy-Schwarz inequality to conclude

$$|\hat V_n^*(\tau,\tau') - V(\tau,\tau')|_{T\times T} = \big|\mathrm{E}^*\, G_n(\tau)G_n(\tau')^\top - V(\tau,\tau')\big|_{T\times T} + o_{\mathrm{P}}(1).$$

In view of (B.4) and the arguments above, to show that the right-hand side of the preceding display is within $o_{\mathrm{P}}(1)$ of $|J_n^{-1}(\tau)\,\mathrm{E}^*\,\mathbb{W}_n(\tau)\mathbb{W}_n(\tau')^\top J_n^{-1}(\tau) - V(\tau,\tau')|_{T\times T}$, it suffices to establish that the E^*-expectation of

$$\xi_n := \big|\mathbb{G}_n z_{\hat\beta_n^*(\tau),\tau} - \mathbb{G}_n z_{\hat\beta_n(\tau),\tau}\big|_T^2$$

converges to zero in probability. I will show below that this already follows if the display has a P-probability limit of zero. Indeed, for any $\eta > 0$, $\mathrm{P}(\xi_n > \eta)$ does not exceed

$$\mathrm{P}\bigg(\sup_{\varrho(z_{\beta,\tau},z_{\beta',\tau'})<\delta} |\,\mathbb{G}_n\, z_{\beta,\tau} - \mathbb{G}_n z_{\beta',\tau'}|^2 > \eta\bigg) + \mathrm{P}\Big(\big|\varrho\big(z_{\hat\beta_n^*(\tau),\tau}, z_{\hat\beta_n(\tau),\tau}\big)\big|_T \ge \delta\Big). \quad \text{(B.5)}$$

The limit superior of the first term on the right is arbitrarily small by Lemma A.7(ii). To see that the second term is eventually small as well, use Lemma A.7(i) to establish

$$\mathrm{E}\big|\varrho\big(z_{\hat\beta_n^*(\tau),\tau}, z_{\hat\beta_n(\tau),\tau}\big)\big|_T^{2q/(q-2)} \lesssim \mathrm{E}\big|\hat\beta_n^*(\tau)-\hat\beta_n(\tau)\big|_T \le n^{-1/2}\sup_{n\ge1}\mathrm{E}\big|\sqrt{n}\big(\hat\beta_n^*(\tau)-\hat\beta_n(\tau)\big)\big|_T.$$

The expression on the right converges to zero by Lemma A.4. The Markov inequality yields $|\hat\beta_n(\tau)-\beta(\tau)|_T \to^{\mathrm{P}} 0$ and the desired result.

It now follows that $\xi_n \to^{\mathrm{P}} 0$ and therefore also $\xi_n \rightsquigarrow 0$ by Lemma 1.10.2 of van der Vaart and Wellner (1996). Hence, $\mathrm{E}\min\{\xi_n,\Delta\}\to 0$ due to boundedness and continuity of $z\mapsto\min\{z,\Delta\}$ with nonnegative z and Δ. Conclude from the Markov inequality that the second term on the right of

$$\mathrm{E}^*\,\xi_n = (\mathrm{E}^*\,\xi_n - \mathrm{E}^*\min\{\xi_n,\Delta\}) + \mathrm{E}^*\min\{\xi_n,\Delta\}.$$

converges to zero in probability. The first term is bounded above by $\mathrm{E}^*\,\xi_n 1\{\xi_n>\Delta\}$. For a small enough $\varepsilon>0$, the right-hand side of $\mathrm{E}\,\xi_n 1\{\xi_n>\Delta\}\le \mathrm{E}\,\xi_n^{1+\varepsilon}\Delta^{-\varepsilon}$ is finite by Lemma A.2(iii) uniformly in n and converges to zero as $\Delta\to\infty$. Deduce from the Markov inequality that the preceding display has a probability limit of zero. Combine the results above and Lemma A.6 to obtain

$$|\hat V_n^*(\tau,\tau')-V(\tau,\tau')|_{T\times T} = \big|J^{-1}(\tau)\,\mathrm{E}^*\,\mathbb{W}_n(\tau)\mathbb{W}_n(\tau')^\top J^{-1}(\tau')-V(\tau,\tau')\big|_{T\times T}+o_{\mathrm{P}}(1)$$

Because $V(\tau,\tau')=J^{-1}(\tau)\Sigma(\tau,\tau')J^{-1}(\tau')$, it now suffices to show that $\mathrm{E}^*\,\mathbb{W}_n(\tau)\mathbb{W}_n(\tau')^\top$ and $\Sigma(\tau,\tau')$ are uniformly close in probablity as $n\to\infty$. From the definition of the bootstrap weights we have $\mathrm{E}^*\,\mathbb{W}_n(\tau)\mathbb{W}_n(\tau')^\top = \mathbb{E}_n\,g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'}$. Decompose the difference between $\mathbb{E}_n\,g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'}$ and $\Sigma_n(\tau,\tau')=\mathbb{E}_n\,\mathrm{E}\,g_{\beta(\tau),\tau,\beta(\tau'),\tau'}$ into

$$\begin{aligned}\mathbb{E}_n(g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'}-\mathrm{E}\,g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'})+\mathbb{E}_n(g_{\beta(\tau),\tau,\beta(\tau'),\tau'}-\mathrm{E}\,g_{\beta(\tau),\tau,\beta(\tau'),\tau'})\\ -\mathbb{E}_n\,\mathrm{E}(g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'}-g_{\beta(\tau),\tau,\beta(\tau'),\tau'}).\end{aligned}$$

The first two terms converge to zero by the uniform law of large numbers in Lemma A.8. An argument similar to the one given in Lemma A.7(i) yields

$$\begin{aligned}&\sup_{\tau,\tau'}\sup_{|\beta-\beta(\tau)|<\delta}\sup_{|\beta'-\beta(\tau')|<\delta}|\,\mathbb{E}_n\,\mathrm{E}(g_{\beta,\tau,\beta',\tau'}-g_{\beta(\tau),\tau,\beta(\tau'),\tau'})|\\ &\quad\lesssim\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^{c_i}\sum_{l=1}^{c_i}\sup_{\tau,\tau'}\mathrm{E}\big(1\{|Y_{ik}-X_{ik}^\top\beta(\tau)|<|X_{ik}|\delta\}\\ &\qquad\qquad+1\{|Y_{il}-X_{il}^\top\beta(\tau')|<|X_{il}|\delta\}\big)|X_{ik}^\top X_{il}|\lesssim\delta^{(q-2)/q}.\end{aligned}$$

Because $\mathrm{P}(|\hat\beta_n(\tau)-\beta(\tau)|_T\ge\delta)\to0$ for every $\delta>0$, let first $n\to\infty$ and then $\delta\to0$ to obtain $\mathbb{E}_n\,g_{\hat\beta_n(\tau),\tau,\hat\beta_n(\tau'),\tau'}=\Sigma_n(\tau,\tau')+o_{\mathrm{P}}(1)$ uniformly in $T\times T$.

The proof is complete if $|\Sigma_n(\tau,\tau') - \Sigma(\tau,\tau')|_{T\times T}$ converges to zero as $n \to \infty$. To this end, note that the metric space $(T \times T, |\cdot|)$ is totally bounded because T is bounded. A straightforward computation using Assumption 2.3 and (B.1) gives $|\Sigma_n(\tau_1,\tau_2) - \Sigma_n(\tau_1',\tau_2')| \lesssim |(\tau_1-\tau_1',\tau_2-\tau_2')|^{1/2}$. Thus $\Sigma_n(\tau)$ is asymptotically uniformly equicontinuous. The pointwise convergence given in Assumption 2.4 is therefore uniform by the Arzelà-Ascoli theorem (Davidson, 1994, Theorem 21.7, p. 335). □

Proof of Theorem 2.5. Because $\sqrt{n}M_n'(\beta(\tau),\tau) = 0$, a mean-value expansion gives

$$\sqrt{n}M_n'(\hat\beta_n(\tau),\tau) = I_n(\hat\beta_n(\tau),\tau)\sqrt{n}\big(\hat\beta_n(\tau) - \beta(\tau)\big).$$

In view of Lemma A.5, rearrange to write $\sqrt{n}(\hat\beta_n(\tau) - \beta(\tau))$ as

$$I_n^{-1}(\hat\beta_n(\tau),\tau)\Big(\mathbb{G}_n\big(z_{\hat\beta_n(\tau),\tau} - z_{\beta(\tau),\tau}\big) - \sqrt{n}\,\mathbb{E}_n\, z_{\hat\beta_n(\tau),\tau} + \mathbb{G}_n\, z_{\beta(\tau),\tau}\Big).$$

Computations similar to, but substantially simpler than the ones given in the proof of Theorem 3.3(ii) show that the preceding display is, uniformly in $\tau \in T$, $o_{\mathrm{P}}(1)$ away from $J^{-1}(\tau)\,\mathbb{G}_n\, z_{\beta(\tau),\tau}$. For every finite set of points $\tau_1,\tau_2,\ldots \in T$, the convergence of the marginal vectors

$$(\mathbb{G}_n\, z_{\beta(\tau_1),\tau_1}^\top, \mathbb{G}_n\, z_{\beta(\tau_2),\tau_2}^\top, \ldots)^\top$$

follows from the Lindeberg-Feller central limit theorem and Assumptions 2.1-2.4.

Take $\varrho_T(\tau,\tau') := \varrho(z_{\beta(\tau),\tau}, z_{\beta(\tau'),\tau'})$. To see that (T,ϱ_T) is a totally bounded pseudometric space, note that for every $\varepsilon > 0$ the number of intervals of length ε needed to cover T does not exceed $\mathrm{diam}(T)/\varepsilon$. Hence, in view of A.7(i) and (B.1), there exists some absolute constant Δ such that for every given radius δ, we can pick $\varepsilon = \delta^{2q/(q-2)}/\Delta$. The number of ϱ_T-balls of radius δ needed to cover T is then at most $\mathrm{diam}(T)\Delta/\delta^{2q/(q-2)}$.

The ϱ_T-stochastic equicontinuity of $z_{\beta(\tau),\tau} \mapsto \mathbb{G}_n\, z_{\beta(\tau),\tau}$ is implied by the ϱ-stochastic equicontinuity of $z_{\beta,\tau} \mapsto \mathbb{G}_n\, z_{\beta,\tau}$. The theorem now follows from Theorem 10.2 of Pollard (1990, p. 51) and the continuous mapping theorem in metric spaces (see, e.g. van der Vaart, 1998, Theorem 18.11, p. 259). □

Proof of Theorem 3.1. Essentially the same computations as in the proof of Theorem 3.3(ii) (although without the requirement that $q > 4$) yield

$$\sqrt{n}\big(\hat\beta_n^*(\tau) - \hat\beta_n(\tau)\big) = -J^{-1}(\tau)\mathbb{W}_n(\tau) + o_{\mathrm{P}}(1).$$

Here the minus sign on the right is a consequence of the definition of the perturbed QR problem in (3.3) and will of course have no impact on the asymptotic distribution.

Theorem 3 of Kosorok (2003) implies that the ϱ-stochastic equicontinuity of $\mathbb{G}_n\, z_{\beta,\tau}$ carries over to its multiplier process $\mathbb{W}_n(\beta,\tau)$ if the conditions of Theorem 10.6 of Pollard (1990, p. 53) hold. In fact, inspection of the proof shows that if only ϱ-stochastic equicontinuity is needed, then it suffices already to verify Pollard's conditions (i) and

(iii)-(v). (This observation was also made by Andrews (1994, p. 2284) for Pollard's Theorem 10.6.) Further, as pointed out by Andrews, these conditions can be verified for any pseudometric that satisfies Pollard's condition (v) because Pollard's total boundedness result is not needed. The pseudometric ϱ used here has property (v) by construction.

Following Andrews (1994, p. 2284), the "manageability" condition (i) is implied by the finite uniform entropy property that was established in the proof of Lemma A.7(ii). The remaining moment conditions (iii) and (iv) are implied by Assumption 2.3. By a standard stochastic equicontinuity argument as in (B.5) we can therefore restate, uniformly in $\tau \in T$, the preceding display as

$$\sqrt{n}\big(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau)\big) = -J^{-1}(\tau)\mathbb{W}_n(\beta(\tau), \tau) + o_{\mathrm{P}}(1).$$

Finally, apply the multiplier central limit theorem of Kosorok (2003, Theorem 3) to $\mathbb{W}_n(\beta(\tau), \tau)$. His conclusions again do not depend on the the specific choice of pseudometric because they are used for a total boundedness result that is not required here. (Recall from the proof of Theorem 2.5 that (T, ϱ_T) is totally bounded.) The only condition that still needs to be verified is that $\{z_{\beta(\tau),\tau} : \tau \in T\}$ is almost measurable Suslin, which holds by Lemma A.1. The continuous mapping theorem in metric spaces completes the proof. □

Proof of Theorem 3.6. Let $\Omega_\infty(\tau) = R(\tau)V(\tau, \tau)R(\tau)^\top$. Because $a^\top \Sigma(\tau, \tau)a$ is bounded below by zero and $J(\tau)$ and $R(\tau)$ have full rank, Assumption 3.5 yields $\inf_{\tau\in T} a^\top \Omega_\infty(\tau)a > 0$ for all non-zero a. Conclude from a singular value decomposition that $\inf_{\tau\in T} \lambda_{\min}(\Omega_\infty(\tau)) > 0$ and therefore also $\inf_{\tau\in T} \lambda_{\min}(\Omega_\infty^{1/2}(\tau)) > 0$. Since eigenvalues are Lipschitz continuous on the space of symmetric matrices, apply Theorem 3.3 to deduce

$$\Big|\inf_{\tau\in T} \lambda_{\min}(\hat{\Omega}_n^*(\tau)) - \inf_{\tau\in T} \lambda_{\min}(\Omega_\infty(\tau))\Big| \leq \sup_{\tau\in T}\big|\lambda_{\min}(\hat{\Omega}_n^*(\tau)) - \lambda_{\min}(\Omega_\infty(\tau))\big| \xrightarrow{\mathrm{P}} 0.$$

Hence, $\inf_{\tau\in T} \lambda_{\min}(\hat{\Omega}_n^{*1/2}(\tau)) > 0$ with probability approaching one as $n \to \infty$. On that event, we can write $|\hat{\Omega}_n^{*-1/2}(\tau)|_T^2 \leq d/\inf_{\tau\in T} \lambda_{\min}(\hat{\Omega}_n^{*1/2}(\tau)) \to^{\mathrm{P}} d/\inf_{\tau\in T} \lambda_{\min}(\Omega_\infty^{1/2}(\tau))$. But then $|\hat{\Omega}_n^{*-1/2}(\tau)|_T$ must be bounded in probability and the right-hand side of

$$|\hat{\Omega}_n^{*-1/2}(\tau) - \Omega_\infty^{-1/2}(\tau)|_T \leq |\hat{\Omega}_n^{*-1/2}(\tau)|_T |\Omega_\infty^{-1/2}(\tau)|_T |\hat{\Omega}_n^{*1/2}(\tau) - \Omega_\infty^{1/2}(\tau)|_T$$

converges to zero in probability if $|\hat{\Omega}_n^{*1/2}(\tau) - \Omega_\infty^{1/2}(\tau)|_T$ does. By Proposition 3.2 of van Hemmen and Ando (1980) (see also Higham, 2008, Theorem 6.2, p. 135) this difference satisfies

$$|\hat{\Omega}_n^{*1/2}(\tau) - \Omega_\infty^{1/2}(\tau)|_T \leq \frac{|\hat{\Omega}_n^*(\tau) - \Omega_\infty(\tau)|_T}{\inf_{\tau\in T} \lambda_{\min}(\hat{\Omega}_n^{*1/2}(\tau)) + \inf_{\tau\in T} \lambda_{\min}(\Omega_\infty^{1/2}(\tau))}$$

and therefore has a probability limit of zero by Theorem 3.3.

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and abbreviate $\sup_{\tau\in T}\|\cdot\|$ by $\|\cdot\|_T$. Apply first Theorem 3.1 and the continuous mapping theorem unconditionally, then Proposition 10.7 of Kosorok (2008, pp. 189-190), Theorem 3.1 conditional on the data, and Lipschitz continuity to obtain

$$K_n^*(\hat{\Omega}_n^*, T) = \big\|\Omega_\infty^{-1/2}(\tau)R(\tau)\sqrt{n}\big(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau)\big)\big\|_T + o_{\mathrm{P}}(1) \rightsquigarrow \|\Omega_\infty^{-1/2}(\tau)R(\tau)\mathbb{Z}(\tau)\|_T$$

in probability conditional on the data. Here the bootstrap convergence occurs with respect to the bounded Lipschitz metric as in Theorem 3.1, uniformly on $\mathrm{BL}_1(\mathbb{R})$. Similarly, under the null hypothesis, rewrite $K_n(\hat{\Omega}_n^*, T)$ and then apply Theorem 2.5 and the continuous mapping theorem to establish

$$K_n(\hat{\Omega}_n^*, T) = \big\|\hat{\Omega}_n^{*-1/2}(\tau)R(\tau)\sqrt{n}\big(\hat{\beta}_n(\tau) - \beta(\tau)\big)\big\|_T \rightsquigarrow \|\Omega_\infty^{-1/2}(\tau)R(\tau)\mathbb{Z}(\tau)\|_T =: K(\Omega_\infty, T).$$

For $x \in \ell^\infty(T)$, the map $x \mapsto \|x\|_T$ constitutes a continuous, convex functional. Theorem 11.1 of Davydov, Lifshits, and Smorodina (1998, p. 75) then implies that the distribution function of $K(\Omega_\infty, T)$ is continuous and strictly increasing on (q_0, ∞), where $q_0 = \inf\{q : \mathrm{P}(K(\Omega_\infty, T) \le q) > 0\}$. Because $\mathbb{Z}(\tau)$ is a Gaussian process with non-degenerate variance, we have $\mathrm{P}(K(\Omega_\infty, T) \le 0) \le \mathrm{P}(K(\Omega_\infty, \{\tau\}) \le 0) = 0$ for arbitrary $\tau \in T$. Furthermore, because all norms on $\mathbb{R}^d$ are equivalent, there exists a $c > 0$ such that

$$\mathrm{P}(K(\Omega_\infty, T) < q) \ge \mathrm{P}\big(|\pi_j \circ \Omega_\infty^{-1/2}(\tau)R(\tau)\mathbb{Z}(\tau)|_{\{1,\ldots,d\}\times T} < q/c\big).$$

The supremum on the right is the supremum of the absolute value of mean-zero Gaussian variables. By Lifshits (1982, Corollary) and Davydov et al. (1998, Theorem 11.1), this supremum has a continuous, strictly increasing distribution function on $(0, \infty)$. The right-hand side of the preceding display is therefore strictly positive for every $q > 0$. Hence, zero is in the support of $K(\Omega_\infty, T)$ and we must have $q_0 = 0$. Conclude that the distribution function of $K(\Omega_\infty, T)$ is in fact continuous and strictly increasing on $(0, \infty)$.

I will now argue that the quantiles of $K_n(\hat{\Omega}_n^*, T)$ and $K(\Omega_\infty, T)$ are eventually close in probability. Define the maps

$$\begin{aligned} h_n(q) &= \rho_{1-\alpha}\big(K_n^*(\hat{\Omega}_n^*, T) - q\big) - \rho_{1-\alpha}\big(K_n^*(\hat{\Omega}_n^*, T)\big) \quad \text{and} \\ h(q) &= \rho_{1-\alpha}\big(K(\Omega_\infty, T) - q\big) - \rho_{1-\alpha}\big(K(\Omega_\infty, T)\big). \end{aligned}$$

Despite the fact that h_n may not be a measurable function of the bootstrap weights, $q \mapsto \mathrm{E}^* h_n(q)$ and $q \mapsto \mathrm{E}\, h(q)$ are clearly convex. Furthermore, $\mathrm{E}\, h$ takes on its unique minimum at $q_{1-\alpha}(\Omega_\infty, T) = \arg\min_{q\in\mathbb{R}} \mathrm{E}\, h(q)$ by the properties of the distribution function established above. In addition, both h_n and h are Lipschitz continuous. Proposition 10.7 of Kosorok (2008) and the definition of conditional weak convergence in probability then yield $|\mathrm{E}^* h_n(q) - \mathrm{E}\, h(q)|_Q \to^{\mathrm{P}} 0$ for every compact set $Q \subset \mathbb{R}$. Because

$q_{n,1-\alpha}(\hat{\Omega}_n^*, T) = \arg\min_{q\in\mathbb{R}} \mathrm{E}^* h_n(q)$ and $\mathrm{E}\, h$ has a unique minimum, Lemma 2 of Hjort and Pollard (1993) gives $q_{n,1-\alpha}(\hat{\Omega}_n^*, T) \to^{\mathrm{P}} q_{1-\alpha}(\Omega_\infty, T)$. Thus,

$$K_n(\hat{\Omega}_n^*, T) - q_{n,1-\alpha}(\hat{\Omega}_n^*, T) \rightsquigarrow \|\Omega_\infty^{-1/2}(\tau)R(\tau)\mathbb{Z}(\tau)\|_T - q_{1-\alpha}(\Omega_\infty, T),$$

where the distribution of the right-hand side is again continuous. The first result now follows because by the definition of weak convergence

$$\mathrm{P}\big(K_n(\hat{\Omega}_n^*, T) > q_{n,1-\alpha}(\hat{\Omega}_n^*, T)\big) \to \mathrm{P}\big(\|\Omega_\infty^{-1/2}(\tau)R(\tau)\mathbb{Z}(\tau)\|_T > q_{1-\alpha}(\Omega_\infty, T)\big) = \alpha.$$

Under the alternative, use Theorem 2.5 and then the fact that Ω_∞ and R have full rank to find an $\varepsilon > 0$ such that

$$K_n(\hat{\Omega}_n^*, T)/n \xrightarrow{\mathrm{P}} \big\|\Omega_\infty^{-1/2}(\tau)R(\tau)\big(\beta(\tau) - r(\tau)\big)\big\|_T =: K_\infty > \varepsilon.$$

Hence, the first term on the right-hand side of

$$\begin{aligned}&\mathrm{P}\big(K_n(\hat{\Omega}_n^*, T) \le q_{n,1-\alpha}(\hat{\Omega}_n^*, T)\big)\\ &\quad\le \mathrm{P}\big(|K_n(\hat{\Omega}_n^*, T)/n - K_\infty| \ge \varepsilon\big) + \mathrm{P}\big(q_{n,1-\alpha}(\hat{\Omega}_n^*, T) > n(K_\infty - \varepsilon)\big)\end{aligned}$$

converges to zero. To see that the second term converges to zero as well, note that $q_{n,1-\alpha}(\hat{\Omega}_n^*, T)$ is bounded in probability by the arguments given in the previous paragraph. Hence, $q_{n,1-\alpha}(\hat{\Omega}_n^*, T)/n \to^{\mathrm{P}} 0$ and the desired conclusion follows because $K_\infty - \varepsilon > 0$. All of the above remains valid with I_d in place of $\hat{\Omega}_n^*$ and Ω_∞. □

Proof of Corollary 3.9. Denote the diagonal matrix of a square matrix A by $\operatorname{diag} A$. For every given Δ, we can find a matrix R such that

$$K_n^*(\hat{V}_n^*, T, \Delta) = \sup_{\tau\in T} \big|\operatorname{diag}(\hat{V}_n^*(\tau,\tau))^{-1/2} R\sqrt{n}\big(\hat{\beta}_n^*(\tau) - \hat{\beta}_n(\tau)\big)\big|_{\max}.$$

This statistic is of the same form as the statistic $K_n^*(\Omega, T)$ used in Theorem 3.3 with $\Omega = \operatorname{diag} \hat{V}_n^*$. If we also define

$$K_n(\hat{V}_n^*, T, \Delta) = \sup_{\tau\in T} \big|\operatorname{diag}(\hat{V}_n^*(\tau,\tau))^{-1/2} R\sqrt{n}\big(\hat{\beta}_n(\tau) - \beta(\tau)\big)\big|_{\max},$$

then we can view $K_n(\hat{V}_n^*, T, \Delta)$ as $K_n(\operatorname{diag} \hat{V}_n^*, T)$ under the null hypothesis and the event inside the displayed probability in the corollary is equivalent to $\{K_n(\hat{V}_n^*, T, \Delta) \le q_{n,1-\alpha}(\hat{V}_n^*, T, \Delta)\}$. Hence, $\mathrm{P}(K_n(\hat{V}_n^*, T, \Delta) \le q_{n,1-\alpha}(\hat{V}_n^*, T, \Delta)) \to 1 - \alpha$ follows from Theorem 3.6 if its proof also applies to the weight matrix $\operatorname{diag} \hat{V}_n^*$. Because $\inf_{\tau\in T} a^\top V(\tau,\tau)a > 0$ for all non-zero a implies $\inf_{\tau\in T} a^\top \operatorname{diag} V(\tau,\tau)a > 0$ for all non-zero a, I only have to show that $|\operatorname{diag} \hat{V}_n^*(\tau,\tau) - \operatorname{diag} V(\tau,\tau)|_T \to^{\mathrm{P}} 0$. But this is implied by $|\hat{V}_n^*(\tau,\tau) - V(\tau,\tau)|_T \to^{\mathrm{P}} 0$ because for any square matrices A, B of identical dimension, $|\operatorname{diag} A - \operatorname{diag} B| = |\operatorname{diag}(A - B)| \le |A - B|$. □

Proof of Corollary 3.10. Inspection of the proof of Theorem 3.6 reveals that the desired conclusion holds if (i) $K_n(\hat{\Omega}_n^*, T_n)$ converges in distribution to $K(\Omega_\infty, T)$ and (ii) $K_n^*(\hat{\Omega}_n^*, T_n)$ and $K_n^*(\hat{\Omega}_n^*, T)$ are $o_{\mathrm{P}}(1)$ away from one another. For (i), note that T_n is a subset of the pseudometric space (T, ϱ_T). Because $\tau_n \to \tau$, conclude from Lemma A.7(i) that $\varrho_T(\tau_n, \tau) \to 0$. It now follows from Exercise 7.5.5 of Kosorok (2008, p. 125, with his T_0 equal to T) and the extended continuous mapping theorem that $K_n(\hat{\Omega}_n^*, T_n) \rightsquigarrow K(\Omega_\infty, T)$.

For (ii), we obtain $K_n^*(\hat{\Omega}_n^*, T_n) \rightsquigarrow K(\Omega_\infty, T)$ unconditionally using the same argument. Therefore, again by the extended continuous mapping theorem, we in fact have the joint convergence

$$\big(K_n^*(\hat{\Omega}_n^*, T_n), K_n^*(\hat{\Omega}_n^*, T)\big) \rightsquigarrow \big(K(\Omega_\infty, T), K(\Omega_\infty, T)\big)$$

unconditionally, which immediately gives $K_n^*(\hat{\Omega}_n^*, T_n) - K_n^*(\hat{\Omega}_n^*, T) \to^{\mathrm{P}} 0$. □

References

Alexander, K. M. (1985). Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Volume 2, pp. 475–493. University of California Press, Berkeley.

Andrews, D. W. K. (1994). Empirical process methods in econometrics. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume IV, Chapter 37, pp. 2248–2294. Elsevier.

Andrews, D. W. K. and M. Buchinsky (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica 68*, 21–51.

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

Belloni, A., V. Chernozhukov, and I. Fernández-Val (2011). Conditional quantile processes based on series or many regressors. Preprint, `arXiv:1105.6154`.

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

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

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

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

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

Chetverikov, D., B. Larsen, and C. Palmer (2013). IV quantile regression for group-level treatments. Unpublished manuscript.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, Oxford.

Davidson, R. (2012). Statistical inference in the presence of heavy tails. *Econometrics Journal 15*, C31–C53.

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

Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998). *Local Properties of Distributions of Stochastic Functionals*, Volume 173 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI.

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

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. *Biometrika 94*, 995–999.

Ghosh, M., W. C. Parr, K. Singh, and G. J. Babu (1984). A note on bootstrapping the sample median. *Annals of Statistics 12*, 1130–1135.

Gonçalves, S. and H. White (2005). Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association 100*, 970–979.

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

Gutenbrunner, C., J. Jurêcková, R. Koenker, and S. Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics 2*, 307–333.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory 11*, 105–121.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hendricks, W. and R. Koenker (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association 87*, 58–68.

Higham, N. J. (2008). *Functions of Matrices: Theory and Computation*. Society for Industrial & Applied Mathematics, Philadelphia, PA.

Hjort, N. and D. Pollard (1993). Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Karlin, S., E. C. Cameron, and P. T. Williams (1981). Sibling and parent-offspring correlation estimation with variable family size. *Proceedings of the National Academy of Sciences 78*, 2664–2668.

Kato, K. (2011). A note on moment convergence of bootstrap M-estimators. *Statistics & Decisions 28*, 51–61.

Kloek, T. (1981). OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica 49*, 205–207.

Knight, K. (1998). Limiting distributions of l_1 regression estimators under general conditions. *Annals of Statistics 26*, 756–770.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*, 74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker, R. (2013). *quantreg: Quantile Regression*. R package version 5.05.

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

Koenker, R. and J. A. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*, 1296–1310.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis 84*, 299–318.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York.

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

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Lifshits, M. A. (1982). On the absolute continuity of distributions of functionals of random processes. *Theory of Probability and Its Applications 27*, 600–607.

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics 16*, 1696–1708.

MacKinnon, J. G. and M. D. Webb (2015). Wild bootstrap inference for wildly different cluster sizes. Department of Economics, Queen's University Working Paper 2-2015.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*, Volume 77 of *Lecture Notes in Statistics*. Springer, New York.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics & Statistics 72*, 334–338.

Parente, P. M. and J. M. Santos Silva (2015). Quantile regression with clustered data. *Journal of Econometric Methods*, forthcoming.

Parzen, M. I., L.-J. Wei, and Z. Ying (1994). A resampling method based on pivotal estimating functions. *Biometrika 81*, 341–350.

Pollard, D. (1982). A central limit theorem for empirical processes. *Journal of the Australian Mathematical Mathematical Society (Series A) 33*, 235–248.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute for Mathematical Statistics.

Portnoy, S. (2014). The jackknife's edge: Inference for censored regression quantiles. *Computational Statistics and Data Analysis 72*, 273–281.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics 32*, 143–155.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New

York.

van Hemmen, J. L. and T. Ando (1980). An inequality for trace ideals. *Communications in Mathematical Physics 76*, 143–148.

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

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

Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. Department of Economics, University of Calgary Working Paper.

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

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics 14*, 1261–1295.

Yoon, J. and A. Galvao (2013). Robust inference for panel quantile regression models with individual effects and serial correlation. Unpublished manuscript.