# Smoothed GMM for quantile models*

Luciano de Castro† Antonio F. Galvao‡ David M. Kaplan§ Xin Liu¶

February 18, 2018

**Abstract**

This paper develops theory for feasible estimation and testing of finite-dimensional parameters identified by general conditional quantile restrictions, under much weaker assumptions than previously seen in the literature. This includes instrumental variables nonlinear quantile regression as a special case. More specifically, we consider a set of unconditional moments implied by the conditional quantile restrictions, providing conditions for local identification. Since estimators based on the sample moments are generally impossible to compute numerically in practice, we study feasible estimators based on smoothed sample moments. We propose a method of moments estimator for exactly identified models, as well as a generalized method of moments estimator for over-identified models. We establish consistency and asymptotic normality of both estimators under general conditions that allow for weakly dependent data and nonlinear structural models. Simulations with iid and dependent data illustrate the finite-sample properties. Our in-depth empirical application concerns the consumption Euler equation derived from quantile utility maximization. Advantages of the quantile Euler equation include robustness to fat tails, decoupling of risk attitude from the elasticity of intertemporal substitution, and log-linearization without any approximation error. For the four countries we examine, the quantile estimates of discount factor and elasticity of intertemporal substitution are economically reasonable for a range of quantiles above the median, even when two-stage least squares estimates are not reasonable.

*Keywords*: instrumental variables, nonlinear quantile regression, quantile utility maximization.
*JEL classification codes:* C31, C32, C36

*The authors would like to express their appreciation to seminar and conference participants at the University of Illinois at Urbana–Champaign, Pennsylvania State University, University of Iowa, Brown University, University of Connecticut, University of Colorado (Boulder), 2017 Midwest Econometrics Group, and the 2017 North American Summer Meeting of the Econometric Society, for helpful comments, references, and discussions. Code is provided for all methods, simulations, and applications.

†Department of Economics, University of Iowa. E-mail: decastro.luciano@gmail.com
‡Department of Economics, University of Arizona. E-mail: agalvao@email.arizona.edu
§Department of Economics, University of Missouri. E-mail: kaplandm@missouri.edu
¶Department of Economics, University of Missouri. E-mail: xl6f6@mail.missouri.edu

# 1 Introduction

Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) has attracted considerable interest in statistics and econometrics. QR estimates conditional quantile functions that provide insight into heterogeneous effects of policy variables. This is especially valuable for program evaluation studies, where these methods help analyze how treatments or social programs affect the outcome's distribution. Nevertheless, endogeneity has been a pervasive concern in economics due to simultaneous causality, omitted variables, measurement error, self-selection, and estimation of equilibrium conditions, among other causes. Extending the standard QR, Chernozhukov and Hansen (2005, 2006, 2008) present results on identification, estimation, and inference for an instrumental variables QR (IVQR) model that allows for endogenous regressors.[1] However, computational difficulties have limited practical estimators to linear models with iid data (discussed below).

Under weaker conditions than prior IVQR papers, we develop theory around feasible smoothed estimators.[2] We consider the set of unconditional moments implied by a general parametric conditional quantile restriction and study exactly identified and over-identified models. Under misspecification of the conditional model, our results still hold for the pseudo-true parameter solving the unconditional moments, complementing the results in Angrist, Chernozhukov, and Fernández-Val (2006) for QR. For identification, we provide sufficient conditions for local identification based on these moments. For estimation, since using unsmoothed sample moments is generally intractable, we study smoothed estimators that compute quickly and may have improved precision (Kaplan and Sun, 2017). Specifically, we develop smoothed method of moments (MM) and smoothed generalized method of moments (GMM) quantile estimators for exactly identified and over-identified models, respectively. Unlike prior IVQR estimation papers, we allow for weakly dependent data and nonlinear structural models when establishing the large sample properties of the estimators, namely,

[1] We refer to Chernozhukov, Hansen, and Wüthrich (2017) for an overview of IVQR. They discuss alternative, complementary QR models with endogeneity in Section 1.2.5, specifically the triangular system and local quantile treatment effect (LQTE) model. Even in the LQTE model, the IVQR estimator has a meaningful interpretation; see Wüthrich (2016).

[2] The methods developed in this paper are also related to those for semiparametric and nonparametric models. Identification, estimation, and inference of general (non-smooth) conditional moment restriction models have received much attention in the econometrics literature, as in Newey and McFadden (1994, §7), Chen, Linton, and van Keilegom (2003), Chen and Pouzo (2009, 2012), and Chen and Liao (2015), for example. However, theoretical results are only for unsmoothed estimators that are often not computationally feasible in practice.

consistency and asymptotic normality.

Our in-depth empirical study estimates a quantile Euler equation using aggregate time series data. This equation is derived from a quantile utility maximization model. This model is an interesting alternative to the standard expected utility model because it is robust to fat tails and allows heterogeneity through the quantiles (partially decoupling the elasticity of intertemporal substitution from risk attitude), and the resulting Euler equation does not suffer from any approximation error when log-linearized.[3] Quantile preferences were first studied by Manski (1988) and were axiomatized by Chambers (2009) and Rostek (2010). De Castro and Galvao (2017) use quantile preferences in a dynamic economic setting and provide a comprehensive analysis of a dynamic rational quantile model. They derive the policy function (Euler equation) as a nonlinear conditional quantile restriction. Structural parameters from the quantile utility problem can be estimated by our smoothed GMM method. Using a standard economic model of intertemporal allocation of consumption, we estimate the elasticity of intertemporal substitution (EIS) coefficient. From this, we employ a variation of the standard economy model of Lucas (1978) where the economic agents decide on the intertemporal consumption and savings (assets to hold) over an infinity horizon economy, maximizing the discounted present value of the stream of quantile utilities, subject to a linear budget constraint. The decision generates a policy function, which is used to estimate the parameters of interest for a given utility function. Numerous papers have estimated the EIS, e.g., Hansen and Singleton (1983), Hall (1988), Campbell and Mankiw (1989), Ogaki and Reinhart (1998), and Yogo (2004). For the four countries we study, the smoothed quantile estimates of the discount factor and EIS are economically reasonable for a range of quantiles above the median, including cases where the 2SLS estimates are not reasonable.

For IVQR estimation of the model in Chernozhukov and Hansen (2005), the literature lacks results for feasible estimators allowing nonlinear structural models and dependent data.[4] The following are iid sampling assumptions: Condition (i) on p. 310 in Chernozhukov and Hong (2003), Assumption 2.R1 in Chernozhukov and Hansen (2006), and Assumption 1 in Kaplan and Sun (2017). A nonlinear structural model is allowed by the computationally demanding[5] Markov Chain Monte Carlo estimator in Chernozhukov and Hong (2003,

[3]Heavy tails in consumption data have been documented recently by Toda and Walsh (2015, 2017).

[4]For nonlinear QR (no IV), see Powell (1994, §2.2), Oberhofer and Haupt (2016), and references therein.

[5]Chernozhukov et al. (2017) comment, "This approach bypasses the need to optimize a non-convex and non-smooth criterion at the cost of needing to design a sampler that adequately explores the quasi-posterior

Ex. 3, p. 297ff.), but linear-in-parameters models are required in (3.4) in Chernozhukov and Hansen (2006) and Assumption 1 in Kaplan and Sun (2017). Additionally, Chernozhukov and Hansen (2006) note, "The computational advantages of our estimator rapidly diminish as the number of endogenous variables increases" (p. 501). Even if only one observed variable is endogenous, this restriction limits the use of interactions and transformations (like polynomial terms). The results from Chernozhukov and Hansen (2006) have been extended in unpublished work by Su and Yang (2011) to non-iid data for use with a correctly specified linear spatial autoregressive model, treating regressors and instruments as nonstochastic. Chen and Lee (2017) propose an estimator for linear-in-parameters IVQR models using mixed integer quadratic programming, but computation is very slow: with only four parameters and $n = 100$ observations, their Table 1 shows average computation times for IV median regression exceeding five minutes. Wüthrich (2017) proposes an estimator without assuming linearity but only for a binary treatment (and iid data). From a Bayesian perspective, Lancaster and Jun (2010) allow nonlinear models but only iid data (and require computation over a grid of coefficient values or else by Markov Chain Monte Carlo). We relax both iid sampling and linearity in our formal results, while maintaining the computational simplicity, speed, and scalability of the method in Kaplan and Sun (2017), and adding the efficiency of two-step GMM.

Historically, the idea of GMM with smoothed IVQR moment conditions was proposed first in unpublished notes by MaCurdy and Hong (1999), mentioned later in (also unpublished) MaCurdy and Timmins (2001, §2.4) and the handbook chapter by MaCurdy (2007, §5).[6] Whang (2006) and Otsu (2008) use moment smoothing for empirical likelihood QR. The closely related idea of smoothing non-differentiable *objective functions* goes back to Amemiya (1982, §3), if not earlier, and such smoothing has been employed for QR by Horowitz (1998), Galvao and Kato (2016), and Fernandes, Guerre, and Horta (2017), among others. Kaplan and Sun (2017, §2.2) argue that smoothing the moment conditions (instead of objective function) is better even for QR, in terms of simplicity and bias. In non-quantile models, indicator function smoothing is used by Horowitz (1992) for maximum score and by Bruins, Duffy, Keane, and Smith (2015) for discrete choice indirect inference.

Section 2 presents the model and discusses identification. Section 3 develops the smoothed MM and GMM estimators, whose asymptotic properties are provided in Section 4. Section 5

in a reasonable amount of computation time."

[6] Among others, Buchinsky (1998, §III.A) discusses QR (but not IVQR) as GMM.

contains simulation results. In Section 6 we illustrate the new approach empirically. Section 7 suggests directions for future research. The appendix collects all proofs.

We conclude this introduction with some remarks about the notation. Random variables and vectors are uppercase ($Y$, $X$, etc.), while non-random values are lowercase ($y$, $x$); for vector/matrix multiplication, all vectors are treated as column vectors. Also, $\mathbb{1}\{\cdot\}$ is the indicator function, $\mathrm{E}(\cdot)$ expectation, $\mathrm{Q}_\tau(\cdot)$ the $\tau$-quantile, $\mathrm{P}(\cdot)$ probability, $\doteq$ "is equal to, up to smaller-order terms," $\asymp$ "has exact (asymptotic) rate/order of," and $\mathrm{N}(\mu, \sigma^2)$ the normal distribution. For vectors, $\|\cdot\|$ is the Euclidean norm. Acronyms used include those for central limit theorem (CLT), continuous mapping theorem (CMT), elasticity of intertemporal substitution (EIS), generalized method of moments (GMM), mean value theorem (MVT), probability density function (PDF), uniform law of large numbers (ULLN), and weak law of large numbers (WLLN).

# 2 Model and identification

We consider the following nonlinear conditional quantile model

$$\mathrm{Q}_\tau[\Lambda\big(Y_i, X_i, \beta_{0\tau}\big) \mid Z_i] = 0, \tag{2.1}$$

where $Y_i \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ is the endogenous variable vector, $Z_i \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ is the full instrument vector that contains $X_i \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ as a subset, $\Lambda(\cdot)$ is the "residual function" that is known up to the finite-dimensional parameter of interest $\beta_{0\tau} \in \mathcal{B} \subseteq \mathbb{R}^{d_\beta}$, and $\tau \in (0,1)$ is the quantile index. The model in (2.1) can be represented by conditional moment restrictions as

$$0 = \mathrm{E}\big[\mathbb{1}\big\{\Lambda\big(Y_i, X_i, \beta_{0\tau}\big) \le 0\big\} - \tau \mid Z_i\big], \tag{2.2}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

To estimate $\beta_{0\tau}$, we use unconditional moments implied by (2.2):

$$0 = \mathrm{E}\big\{Z_i\big[\mathbb{1}\big\{\Lambda\big(Y_i, X_i, \beta_{0\tau}\big) \le 0\big\} - \tau\big]\big\}. \tag{2.3}$$

Kaplan and Sun (2017) consider a special case of (2.3) with $\Lambda\big(Y, X, \beta\big) = Y_1 - Y_{-1}^\top \beta_1 - X^\top \beta_2$, where $Y_1$ is the outcome and $Y_{-1} = (Y_2, \ldots, Y_{d_Y})$ are endogenous regressors. We take $Z_i$ as given; see Kaplan and Sun (2017, §2.1, pp. 108–110) for discussion of optimal instruments. Our asymptotic results assume only (2.3), so they are robust to misspecification of the structural model in (2.3), treating $\beta_{0\tau}$ as the pseudo-true parameter satisfying (2.3).

Given (2.3), $\beta_{0\tau}$ is "locally identified" if there exists a neighborhood of $\beta_{0\tau}$ within which only $\beta_{0\tau}$ satisfies (2.3). This holds if the partial derivative matrix of the right-hand side of (2.3) with respect to the $\beta$ argument is full rank; e.g., see Chen, Chernozhukov, Lee, and Newey (2014, p. 787). This full rank condition is formally stated below in Assumption A9(ii). The following proposition states the local identification result.

**Proposition 2.1.** *Given* (2.3) *and (the full rank) Assumption A9(ii),* $\beta_{0\tau}$ *is locally identified.*

Global identification is notoriously more difficult to establish, although Chernozhukov and Hansen (2005, Thm. 2 and App. C) provide some results for IVQR.

To fix ideas, we discuss two examples of structural models in the form of (2.1). The first example is a random coefficient model as in Chernozhukov and Hansen (2005, 2006). Let $D$ be an endogenous "treatment" (like education), $U \sim \text{Unif}(0,1)$ an unobserved variable (like ability), $X$ exogenous regressors, and $\bar{Y}_d = q\big(d, x, \beta_0(U)\big)$ potential outcomes (like wage), where $q(\cdot)$ is known and $\beta_0(U)$ is a random coefficient vector depending on $U$. Let $\beta_{0\tau} = \beta_0(\tau)$, $Y = (\bar{Y}_D, D)$, and $\Lambda(Y, X, \beta) = \bar{Y}_D - q(D, X, \beta)$. Under their Assumptions A1–A5, Theorem 1 in Chernozhukov and Hansen (2005) yields a conditional quantile restriction like in (2.1):

$$\tau = \mathrm{P}\big[\Lambda\big(Y, X, \beta_{0\tau}\big) \le 0 \mid Z\big].$$

Another example of a structural model applies to our empirical application in Section 6. Under certain assumptions, if individuals maximize the $\tau$-quantile of utility instead of expected utility, then the resulting consumption Euler equation can be written in the form

$$\mathrm{Q}_\tau[\beta_{0\tau}(1 + r_{t+1})U'(C_{t+1})/U'(C_t) \mid \Omega_t] = 1,$$

where $\beta$ is the discount factor, $r_t$ is real interest rate, $C_t$ is consumption, $U(\cdot)$ is the utility function, $\Omega_t$ is the information set, and $\mathrm{Q}_\tau[W_t \mid \Omega_t]$ denotes the conditional $\tau$-quantile of $W_t$ given $\Omega_t$. With isoelastic utility and instruments $Z_t$ chosen from $\Omega_t$, we obtain (2.1):

$$\mathrm{Q}_\tau[\beta_{0\tau}(1 + r_{t+1})(C_{t+1}/C_t)^{-\gamma_{0\tau}} - 1 \mid Z_t] = 0.$$

# 3 The smoothed MM and GMM estimators

This section presents smoothed estimators based on the moment conditions in (2.3). The smoothed MM and smoothed GMM estimators are designed for exactly identified and over-identified models, respectively. We now introduce notation, followed by the estimators.

Let the population map $M\colon \mathcal{B} \times \mathcal{T} \mapsto \mathbb{R}^{d_Z}$ be

$$M(\beta, \tau) \equiv \mathrm{E}[g_i^u(\beta, \tau)], \tag{3.1}$$

$$g_i^u(\beta, \tau) \equiv g^u(Y_i, X_i, Z_i, \beta, \tau) \equiv Z_i[\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\} - \tau], \tag{3.2}$$

where superscript "$u$" denotes "unsmoothed." The population moment condition (2.3) is

$$0 = M(\beta_{0\tau}, \tau). \tag{3.3}$$

Without smoothing, the corresponding sample moments simply replace population expectation $\mathrm{E}(\cdot)$ with sample expectation $\hat{\mathrm{E}}(\cdot)$, i.e., the sample average. Analogous to the population map $M(\cdot)$ in (3.1), the unsmoothed sample map is

$$\hat{M}_n^u(\beta, \tau) \equiv \hat{\mathrm{E}}[g^u(Y, X, Z, \beta, \tau)] \equiv \frac{1}{n}\sum_{i=1}^{n} g_i^u(\beta, \tau). \tag{3.4}$$

The well-known computational difficulty (e.g., Chernozhukov and Hong, 2003, Fig. 1(a) and Ex. 3, p. 297) of minimizing a GMM criterion based on $\hat{M}_n^u(\beta, \tau)$ comes from the discontinuous indicator function $\mathbb{1}\{\cdot\}$ inside $g_i^u(\beta, \tau)$. To address this difficulty, we smooth the indicator function.

With smoothing (no "$u$" superscript), the sample analogs of (3.1) and (3.2) are

$$\begin{aligned} g_{ni}(\beta, \tau) &\equiv g_n(Y_i, X_i, Z_i, \beta, \tau) \equiv Z_i\big[\tilde{I}\big(-\Lambda(Y_i, X_i, \beta)/h_n\big) - \tau\big], \\ \hat{M}_n(\beta, \tau) &\equiv \frac{1}{n}\sum_{i=1}^{n} g_{ni}(\beta, \tau), \end{aligned} \tag{3.5}$$

where $h_n$ is a bandwidth (sequence) and $\tilde{I}(\cdot)$ is a smoothed version of the indicator function $\mathbb{1}\{\cdot \geq 0\}$. The $\tilde{I}(\cdot)$ in Figure 1 has been used by Horowitz (1998), Whang (2006), and Kaplan and Sun (2017), who use the fact that its derivative is a fourth-order kernel to establish higher-order improvements in the linear iid setting. The double subscript on $g_{ni}$ is a reminder that we have a triangular array setup because $g_{ni}$ depends on the bandwidth sequence $h_n$ in addition to $(Y_i, X_i, Z_i)$.

## 3.1 Method of moments (exact identification)

With exact identification ($d_Z = d_\beta$), our estimator solves the smoothed sample moment conditions[7]

$$\hat{M}_n(\hat{\beta}_{\mathrm{MM}}, \tau) = 0. \tag{3.6}$$

[7]The right-hand side of (3.6) can be relaxed to $o_p(n^{-1/2})$.

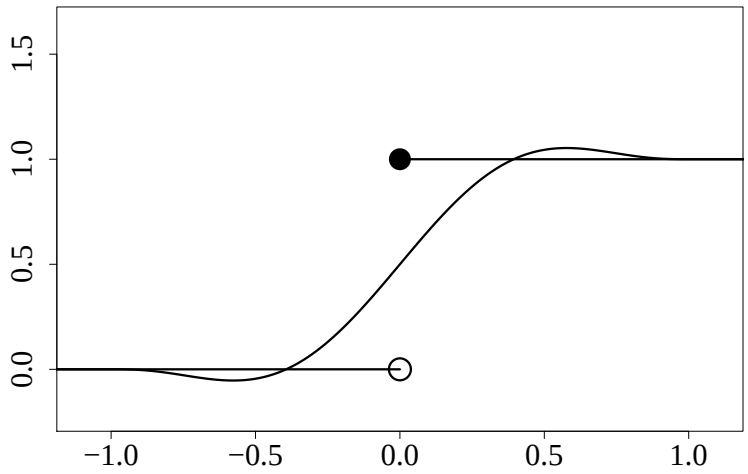


Figure 1: $\mathbb{1}\{u \geq 0\}$ and $\tilde{I}(u) = \mathbb{1}\{-1 \leq u \leq 1\}\big[0.5 + \frac{105}{64}\big(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\big)\big] + \mathbb{1}\{u > 1\}$.

Numerically, as long as the bandwidth is not too near zero and $\Lambda(\cdot)$ is differentiable in $\beta$, (3.6) is easy to solve since the Jacobian exists. Further, it is easy to check whether the estimate indeed satisfies (3.6). In contrast, with over-identification, it is impossible to know if the numerical solution is the global (not just local) minimum of the GMM criterion function.[8] Consequently, combining moments and using (3.6) provides a reliable initial value for the GMM minimization.

## 3.2 "One-step" GMM (over-identification)

With over-identification ($d_Z > d_\beta$), (3.6) has no solution. Thus, a natural GMM estimator is the "one-step" estimator proposed in Newey and McFadden (1994, p. 2151) that takes one Newton–Raphson-type step from an initial consistent (but not efficient) estimator. Newey and McFadden (1994, Thm. 3.5, p. 2151) show that this is sufficient for asymptotic efficiency, although they assume $g(\cdot)$ is smooth and fixed. We use the one-step estimator only for intermediate computation, focusing on the more common two-step estimator (in Section 3.3) for asymptotic theory.

Let

$$\bar{G} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\beta^\top} g_{ni}(\bar{\beta}, \tau), \tag{3.7}$$

where $\nabla_{\beta^\top}$ denotes the partial derivative with respect to $\beta^\top$, and $\bar{\beta}$ is an initial estimator

[8] See footnote 5 in Chernozhukov and Hong (2003).

consistent for $\beta_{0\tau}$. With iid data, let

$$\bar{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g_{ni}(\bar{\beta},\tau)g_{ni}(\bar{\beta},\tau)^\top \tag{3.8}$$

be an estimator of $\Omega = \mathrm{E}[g_{ni}(\beta_{0\tau},\tau)g_{ni}(\beta_{0\tau},\tau)^\top]$. As in (3.11) of Newey and McFadden (1994), the one-step estimator is

$$\hat{\beta}_{1\mathrm{s}} = \bar{\beta} - (\bar{G}^\top\bar{\Omega}^{-1}\bar{G})^{-1}\bar{G}^\top\bar{\Omega}^{-1}\sum_{i=1}^{n} g_{ni}(\bar{\beta},\tau)/n. \tag{3.9}$$

For $\bar{\beta}$, we use (3.6). For $\bar{\Omega}$ with dependent data, (3.8) is replaced by a long-run variance estimator as in Newey and West (1987) and Andrews (1991b), which we use in our code. However, this is ad hoc; see discussion in Section 4.3.

## 3.3 Two-step GMM estimator (over-identification)

We also consider the two-step GMM estimator to achieve asymptotic efficiency in over-identified models ($d_Z > d_\beta$).[9] Let $\hat{W}$ be a symmetric, positive definite weighting matrix. The smoothed GMM estimator minimizes a weighted quadratic norm of the smoothed sample moment vector:

$$\hat{\beta}_{\mathrm{GMM}} = \underset{\beta\in\mathcal{B}}{\arg\min}\left[\sum_{i=1}^{n} g_{ni}(\beta,\tau)\right]^\top \hat{W}\left[\sum_{i=1}^{n} g_{ni}(\beta,\tau)\right] = \underset{\beta\in\mathcal{B}}{\arg\min}\,\hat{M}_n(\beta,\tau)^\top\hat{W}\hat{M}_n(\beta,\tau). \tag{3.10}$$

The usual optimal weighting matrix is an estimator of the inverse long-run variance of the sample moments:[10] $\hat{W}^* = \bar{\Omega}^{-1} \xrightarrow{p} \Omega^{-1}$, where $\bar{\Omega}$ depends on an initial estimate $\bar{\beta}$ as in Section 3.2. The resulting efficient two-step GMM estimator is

$$\hat{\beta}_{2\mathrm{s}} = \underset{\beta\in\mathcal{B}}{\arg\min}\,\hat{M}_n(\beta,\tau)^\top\bar{\Omega}^{-1}\hat{M}_n(\beta,\tau). \tag{3.11}$$

Computing (3.11) is difficult because the function may be non-convex. To find the global minimum, we use the simulated annealing algorithm from the GenSA package in R (Xiang, Gubian, Suomela, and Hoeng, 2013), which is suited to such problems. Despite its strengths, simulated annealing cannot reliably solve (3.11) without an initial value reasonably close to the solution. Thankfully, such an initial value is provided by (3.6) or (3.9).

[9] This is not always true with time series under fixed-smoothing asymptotics (Hwang and Sun, 2015).

[10] There are other approaches to achieve efficiency without explicitly estimating the long-run variance, like the (Bayesian) exponentially tilted empirical likelihood of Schennach (2007, Thm. 3) and Schennach (2005, p. 36), on which Lancaster and Jun (2010) is based.

After computing (3.11), one could run simulated annealing again with the *unsmoothed* objective function. For linear iid IVQR, Kaplan and Sun (2017) suggest that smoothing improves (pointwise in $\tau$) mean squared error but may reduce estimated heterogeneity (across $\tau$), so the benefit of such a final step is ambiguous.

# 4 Large sample properties

We now establish consistency and asymptotic normality of both the smoothed MM and GMM estimators.

## 4.1 Assumptions

Different subsets of the following assumptions are used for different results.

**Assumption A1.** *For each observation $i$ among $n$ in the sample, endogenous vector $Y_i \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ and instrument vector $Z_i \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$; a subset of $Z_i$ is $X_i \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$, with $d_X \leq d_Z$. The sequence $\{Y_i, Z_i\}$ is strictly stationary and weakly dependent.*

**Assumption A2.** *The function $\Lambda\colon \mathcal{Y} \times \mathcal{X} \times \mathcal{B} \mapsto \mathbb{R}$ is known and has (at least) one continuous derivative in its $\mathcal{B}$ argument for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$.*

**Assumption A3.** *The parameter space $\mathcal{B} \in \mathbb{R}^{d_\beta}$ is compact; $d_\beta \leq d_Z$. Given $\tau \in (0,1)$, the population parameter $\beta_{0\tau}$ is in the interior of $\mathcal{B}$ and uniquely satisfies the moment condition*

$$0 = \mathrm{E}\big[Z_i\big(\mathbb{1}\big\{\Lambda\big(Y_i, X_i, \beta_{0\tau}\big) \leq 0\big\} - \tau\big)\big]. \tag{4.1}$$

**Assumption A4.** *The matrix $\mathrm{E}\big(Z_i Z_i^\top\big)$ is positive definite (and finite).*

**Assumption A5.** *The function $\tilde{I}(\cdot)$ satisfies $\tilde{I}(u) = 0$ for $u \leq -1$, $\tilde{I}(u) = 1$ for $u \geq 1$, and $-1 \leq \tilde{I}(u) \leq 2$ for $-1 < u < 1$. The derivative $\tilde{I}'(\cdot)$ is a symmetric, bounded kernel function of order $r \geq 2$, so $\int_{-1}^{1} \tilde{I}'(u)\,du = 1$, $\int_{-1}^{1} u^k \tilde{I}'(u)\,du = 0$ for $k = 1, \ldots, r-1$, and $\int_{-1}^{1} |u^r \tilde{I}'(u)|\,du < \infty$ but $\int_{-1}^{1} u^r \tilde{I}'(u)\,du \neq 0$, and $\int_{-1}^{1} |u^{r+1} \tilde{I}'(u)|\,du < \infty$.*

**Assumption A6.** *The bandwidth sequence $h_n$ satisfies $h_n = o(n^{-1/(2r)})$.*

**Assumption A7.** *Given any $\beta \in \mathcal{B}$ and almost all $Z_i = z$ (i.e., up to a set of zero probability), the conditional distribution of $\Lambda(Y_i, X_i, \beta)$ given $Z_i = z$ is continuous in a neighborhood of zero.*

**Assumption A8.** *For a fixed $\tau \in (0,1)$, using the definition in* (3.5)*,*

$$\sup_{\beta\in\mathcal{B}} \|\hat{M}_n(\beta,\tau) - \mathrm{E}[\hat{M}_n(\beta,\tau)]\| = o_p(1). \tag{4.2}$$

**Assumption A9.** *Let $\Lambda_i \equiv \Lambda\big(Y_i, X_i, \beta_{0\tau}\big)$ and $D_i \equiv \nabla_\beta \Lambda\big(Y_i, X_i, \beta_{0\tau}\big)$, using the notation*

$$\nabla_\beta \Lambda(y,x,\beta_0) \equiv \frac{\partial}{\partial\beta}\Lambda(y,x,\beta)\Big|_{\beta=\beta_0}, \tag{4.3}$$

*for the $d_\beta \times 1$ partial derivative vector. Let $f_{\Lambda|Z}(\cdot \mid z)$ denote the conditional PDF of $\Lambda_i$ given $Z_i = z$, and let $f_{\Lambda|Z,D}(\cdot \mid z,d)$ denote the conditional PDF of $\Lambda_i$ given $Z_i = z$ and $D_i = d$. (i) For almost all $z$ and $d$, $f_{\Lambda|Z}(\cdot \mid z)$ and $f_{\Lambda|Z,D}(\cdot \mid z,d)$ are at least $r$ times continuously differentiable in a neighborhood of zero, where the value of $r$ is from A5. For almost all $z \in \mathcal{Z}$ and $u$ in a neighborhood of zero, there exists a dominating function $C(\cdot)$ such that $|f^{(r)}_{\Lambda|Z}(u \mid z)| \le C(z)$ and $\mathrm{E}\big[C(Z)|Z|\big] < \infty$. (ii) The matrix*

$$G \equiv \frac{\partial}{\partial\beta^\top}\mathrm{E}\big[Z_i \mathbb{1}\{\Lambda(Y_i,X_i,\beta) \le 0\}\big]\Big|_{\beta=\beta_{0\tau}} = -\mathrm{E}\big\{Z_i D_i^\top f_{\Lambda|Z,D}(0 \mid Z_i, D_i)\big\} \tag{4.4}$$

*has rank $d_\beta$.*

**Assumption A10.** *A pointwise CLT applies:*

$$\sqrt{n}\big\{\hat{M}_n(\beta_{0\tau},\tau) - \mathrm{E}[\hat{M}_n(\beta_{0\tau},\tau)]\big\} \xrightarrow{d} \mathrm{N}(0,\Sigma_\tau). \tag{4.5}$$

**Assumption A11.** *Let $Z_i^{(k)}$ denote the $k$th element of $Z_i$, and similarly $\beta^{(k)}$. Let $G_{kj}$ denote the row $k$, column $j$ element of $G$ (from A9). Assume*

$$-\frac{1}{nh_n}\sum_{i=1}^n \tilde{I}'\big(-\Lambda(Y_i,X_i,\tilde{\beta}_k)/h_n\big) Z_i^{(k)} \frac{\partial}{\partial\beta^{(j)}}\Lambda(Y_i,X_i,\beta)\Big|_{\beta=\tilde{\beta}_{\tau,k}} \xrightarrow{p} G_{kj}. \tag{4.6}$$

*for each $k = 1,\dots,d_\beta$ and $j = 1,\dots,d_\beta$, where each $\tilde{\beta}_k$ lies between $\beta_{0\tau}$ and $\hat{\beta}_{\mathrm{MM}}$ (defined in A3 and* (3.6)*, respectively).*

**Assumption A12.** *For the weighting matrix, $\hat{W} \xrightarrow{p} W$, and both are symmetric, positive definite matrices.*

For transparency, A1 includes sampling assumptions that help establish the high-level assumptions A8, A10, and A11, which may require additional restrictions on dependence (mixing conditions); see Appendix B. Assumption A2 is stronger than a nonparametric model

but more general than a linear-in-parameters model. Assumption A3 assumes global identification, following the GMM tradition going back to Hansen (1982, Thm. 2.1(iii), p. 1035) due to "the difficulty of specifying primitive identification conditions for GMM" (Newey and McFadden, 1994, p. 2120), although Chernozhukov and Hansen (2005) have some such results for IVQR. Assumption A4 matches Assumption 2(ii) in Kaplan and Sun (2017); it is relatively weak, imposing only a finite second moment on $Z_i$ (and, unlike 2SLS, no moment restrictions on $Y_i$ or $X_i$). Assumption A5 is essentially Assumption 4(i,ii) of Kaplan and Sun (2017).

Assumption A6 ensures that the asymptotic effect of smoothing is negligible; it is relatively weak given $r = 4$ (as in Figure 1), and given the optimal $h_n \asymp n^{-1/(2r-1)}$ rate from Kaplan and Sun (2017) for linear IVQR with iid data. With weakly dependent data, other bandwidth restrictions are needed to establish A11; see Appendix B.3 for details. Appendix C.2 contains suggestions for practical bandwidth selection.

Assumption A7 can be checked easily in most cases. For example, if $\Lambda(Y_i, X_i, \beta) = \tilde{Y}_i - (D_i, X_i^\top)\beta$ and $Z^\top = (X^\top, \tilde{Z})$, then A7 is satisfied if the outcome $Y$ has a continuous distribution conditional on almost all $(X^\top, \tilde{Z}) = (x^\top, \tilde{z})$. Assumption A8 generally requires some restriction on dependence and moments, but it is much weaker than the iid sampling assumption of Chernozhukov and Hansen (2006, Assumption 2.R1) or Kaplan and Sun (2017, Assumption 1). Assumption A9 is used for the asymptotic normality result; it generalizes parts of Assumptions 3 and 7 in Kaplan and Sun (2017) to our nonlinear model. The nonsingularity of $G$ is also sufficient for local (but not global) identification (e.g., Chen et al., 2014, p. 787). The CLT in A10 is a high-level assumption, similar to condition (iv) in Theorem 7.2 of Newey and McFadden (1994), for example; like A8, it requires some restriction on dependence and moments but does not require iid sampling. Examples of more primitive sufficient conditions for A8 and A10 are given later in Appendix B. Assumption A11 is actually a generalization of the consistency of Powell's estimator for the asymptotic covariance matrix of the usual quantile regression estimator, as detailed in Section 4.3. Assumption A11 embodies the stochastic equicontinuity that is often separately assumed, as in Theorem 7.2(v) in Newey and McFadden (1994); it also involves interrelated restrictions of dependence, moments, and the bandwidth rate, as described in Appendix B.3.

Assumption A12 is standard for GMM. For two-step GMM, it is satisfied given a consistent estimator of the (inverse) asymptotic covariance matrix; see discussion in Section 4.3.

Finally, note the lack of a conditional quantile restriction. Only the unconditional moments in A3 are assumed to be satisfied by the (pseudo) true parameter. Thus, all our results hold even under misspecification (of a conditional model).

## 4.2 Consistency

To establish consistency, we use Theorem 5.9 in van der Vaart (1998), showing the two required conditions are satisfied here. One condition is an identification condition. The other requires uniform (in $\beta \in \mathcal{B}$) convergence in probability of the sample maps $\hat{M}_n(\beta, \tau)$ to the population map $M(\beta, \tau)$. No iid sampling assumption is required; the second assumption may be established under weak dependence.

A detailed example of primitive conditions for the high-level uniform weak law of large numbers assumed in Assumption A8 is given in Appendix B.1.

In addition to A8, we must show that the sequence of (non-random) maps $\mathrm{E}[\hat{M}_n(\beta, \tau)]$ converges to the desired population map $M(\beta, \tau)$, as in Lemma 4.1.

**Lemma 4.1.** *Under Assumptions A1–A7, for a fixed $\tau$, using definitions in* (3.1) *and* (3.5)*,*

$$\sup_{\beta \in \mathcal{B}} \Big| \mathrm{E}[\hat{M}_n(\beta, \tau)] - M(\beta, \tau) \Big| = o(1). \tag{4.7}$$

Lemma 4.1 is intuitive. Without smoothing, $M(\cdot) = \mathrm{E}[\hat{M}_n^u(\cdot)]$ for all $n$. With smoothing, we should expect this to hold asymptotically if the smoothing is asymptotically negligible. The next result establishes consistency.

**Theorem 4.2.** *Under Assumptions A1–A8 for smoothed MM, and additionally A12 for smoothed GMM, the estimators from* (3.6) *and* (3.10) *are consistent:*

$$\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau} = o_p(1), \quad \hat{\beta}_{\mathrm{GMM}} - \beta_{0\tau} = o_p(1). \tag{4.8}$$

## 4.3 Asymptotic normality

To establish asymptotic normality, smoothing facilitates the usual approach of expanding the sample moments around $\beta_{0\tau}$ because the smoothed sample moments are differentiable. That is, we may take a mean value expansion of the first-order condition, rearrange, and take limits.

The following lemma aids the proof of Theorem 4.4. It relies on A10 and a proof that the asymptotic "bias" is negligible, i.e., $\sqrt{n}\,\mathrm{E}[\hat{M}_n(\beta_{0\tau}, \tau)] \to 0$.

**Lemma 4.3.** *Under Assumptions A1–A6, A9, and A10,*

$$\sqrt{n}\hat{M}_n(\beta_{0\tau},\tau) \xrightarrow{d} \mathrm{N}(0,\Sigma_\tau), \quad \Sigma_\tau = \lim_{n\to\infty} \mathrm{Var}\left(n^{-1/2}\sum_{i=1}^n g_{ni}(\beta_{0\tau},\tau)\right). \tag{4.9}$$

*With iid data and the conditional quantile restriction* $\mathrm{P}\big(\Lambda(Y_i,X_i,\beta_{0\tau}) \le 0 \mid Z_i\big) = \tau$*, then* $\Sigma_\tau = \tau(1-\tau)\,\mathrm{E}(Z_iZ_i^\top)$*.*

The asymptotic normality of our estimators can now be stated. We also show their asymptotically linear (influence function) representations.

**Theorem 4.4.** *Under Assumptions A1–A11 for smoothed MM, and additionally A12 for smoothed GMM, for the estimators from* (3.6) *and* (3.10)*,*

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau}) &= -G^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n g_{ni}(\beta_{0\tau},\tau) + o_p(1),\\
\sqrt{n}(\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau}) &\xrightarrow{d} \mathrm{N}\Big(0,(G^\top\Sigma_\tau^{-1}G)^{-1}\Big),\\
\sqrt{n}(\hat{\beta}_{\mathrm{GMM}} - \beta_{0\tau}) &= -\{G^\top WG\}^{-1}G^\top W\frac{1}{\sqrt{n}}\sum_{i=1}^n g_{ni}(\beta_{0\tau},\tau) + o_p(1),\\
\sqrt{n}(\hat{\beta}_{\mathrm{GMM}} - \beta_{0\tau}) &\xrightarrow{d} \mathrm{N}\big(0,(G^\top WG)^{-1}G^\top W\Sigma_\tau WG(G^\top WG)^{-1}\big),
\end{aligned}$$

*where* $G$ *is from* (4.4)*,* $W$ *is from A12, and* $\Sigma_\tau$ *is from* (4.9)*.*

As usual, choosing a weighting matrix such that $\hat{W} \xrightarrow{p} W = \Sigma_\tau^{-1}$ is asymptotically efficient in the sense that the resulting asymptotic covariance matrix $(G^\top\Sigma_\tau^{-1}G)^{-1}$ minus the above GMM covariance matrix is negative semidefinite. This is the sense in which the two-step estimator in (3.11) is efficient.

Theorem 4.4 can also be used to construct Wald tests in the usual way. The result and its proof are also helpful for constructing "distance metric" hypothesis tests and overidentification tests. We detail these in a separate paper on inference (in progress).

A consistent long-run variance estimator for quantile models is currently lacking in the literature, as lamented in other quantile papers. For example, Kato (2012, p. 268) notes that results in Andrews (1991b) do not apply because they assume smoothness; specifically, Assumptions B(iii) and C(ii) are violated for unsmoothed quantile models. Similarly, Assumptions 4 and 5 in Newey and West (1987) are violated, as is Assumption 4 in de Jong and Davidson (2000). Our smoothing yields differentiability, but also a triangular array, violating (2.2) in Andrews (1991b), for example. However, de Jong and Davidson (2000)

allow triangular arrays and generally have very weak conditions. In future work, we hope to verify their Assumption 4 for our smoothed quantile GMM setting.

# 5 Monte Carlo simulations

This section reports Monte Carlo simulation results to illustrate the finite-sample performance of the proposed methods. Replication code is available on the third author's website.[11]

The following DGPs/models are used; details are in Appendix C.1 and the code. DGP 1 has a binary treatment, binary IV, and iid sampling, as with a randomized treatment offer but self-selection into treatment, and the treatment effect increases with the quantile index $\tau$. DGPs 2 and 3 are time series regressions with measurement error and either Gaussian (DGP 2) or Cauchy (DGP 3) errors in the outcome equation, but no slope heterogeneity. The first three models are exactly identified, so the smoothed MM and GMM estimators are identical; we compare these with the usual QR and IV estimators. DGP 4 is for estimating a log-linearized quantile Euler equation using time series data, as in our empirical application; the model is overidentified, so we can compare MM with different GMM estimators.

To quantify precision, instead of root mean squared error (RMSE), we report "robust RMSE." This replaces bias with median bias and replaces standard deviation with interquartile range (divided by 1.349); it equals RMSE if the sampling distribution is normal. The primary reason to use the "robust" version is that sometimes the usual IV estimator does not even possess a first moment in finite samples, let alone finite variance (e.g., Kinal, 1980).[12]

Table 1 shows the precision of our smoothed estimator from (3.6) ("S(G)MM") with a very small bandwidth $h = 0.0001$ (for simplicity), as well as the usual quantile regression ("QR") estimator (ignoring endogeneity) and the usual (mean) IV estimator.

Table 1 shows that for all DGPs, the smoothed estimator's robust RMSE declines toward zero as $n$ increases. In contrast, the QR estimator's robust RMSE never goes to zero due to endogeneity, and the IV estimator's robust RMSE fails to go to zero in DGP 1 where there is heterogeneity across quantiles. This reflects the theoretical result that only our estimator is consistent for $\gamma_\tau$ when there is endogeneity and heterogeneity.

[11] It is written in R (R Core Team, 2013) and uses packages from Borchers (2015) and Xiang et al. (2013).

[12] If one really cares about finite-sample RMSE per se, then OLS should be preferred to IV in the many cases where the IV RMSE is infinite but the OLS RMSE is finite.

Table 1: Simulated precision of estimators of $\gamma_\tau$.

| | | | Robust RMSE | | | Median Bias | | |
|---|---|---|---|---|---|---|---|---|
| DGP | $\tau$ | $n$ | S(G)MM | QR | IV | S(G)MM | QR | IV |
| 1 | 0.25 | 20 | 26.51 | 31.81 | 41.71 | 18.00 | 27.87 | 40.30 |
| | | 50 | 19.10 | 27.99 | 40.53 | 11.64 | 26.22 | 39.93 |
| | | 200 | 10.94 | 25.74 | 40.17 | 4.09 | 25.22 | 40.03 |
| | | 500 | 8.61 | 25.33 | 40.12 | 1.54 | 25.07 | 40.07 |
| | 0.50 | 20 | 19.20 | 22.14 | 18.69 | 9.04 | 17.48 | 15.30 |
| | | 50 | 13.87 | 21.16 | 16.47 | 4.92 | 18.86 | 14.93 |
| | | 200 | 8.13 | 20.50 | 15.38 | 1.19 | 19.99 | 15.03 |
| | | 500 | 5.11 | 20.16 | 15.21 | 0.52 | 19.96 | 15.07 |
| 2 | 0.25 | 20 | 1.34 | 0.59 | 1.34 | −0.37 | −0.52 | −0.24 |
| | | 50 | 0.86 | 0.55 | 0.73 | −0.09 | −0.53 | −0.03 |
| | | 200 | 0.42 | 0.52 | 0.31 | −0.01 | −0.51 | 0.02 |
| | | 500 | 0.26 | 0.50 | 0.20 | 0.00 | −0.50 | 0.02 |
| | 0.50 | 20 | 1.29 | 0.58 | 1.34 | −0.37 | −0.51 | −0.24 |
| | | 50 | 0.83 | 0.54 | 0.73 | −0.09 | −0.51 | −0.03 |
| | | 200 | 0.40 | 0.51 | 0.31 | 0.02 | −0.50 | 0.02 |
| | | 500 | 0.26 | 0.50 | 0.20 | 0.02 | −0.50 | 0.02 |
| 3 | 0.25 | 20 | 2.72 | 0.75 | 3.91 | −0.42 | −0.55 | −0.20 |
| | | 50 | 1.68 | 0.59 | 3.17 | −0.02 | −0.51 | 0.19 |
| | | 200 | 0.79 | 0.54 | 2.48 | −0.02 | −0.51 | −0.02 |
| | | 500 | 0.47 | 0.51 | 2.33 | 0.00 | −0.50 | 0.06 |
| | 0.50 | 20 | 2.10 | 0.67 | 3.91 | −0.49 | −0.53 | −0.20 |
| | | 50 | 1.36 | 0.56 | 3.17 | −0.15 | −0.51 | 0.19 |
| | | 200 | 0.63 | 0.52 | 2.48 | 0.01 | −0.50 | −0.02 |
| | | 500 | 0.35 | 0.51 | 2.33 | 0.00 | −0.50 | 0.06 |

1000 replications. "S(G)MM" is the estimator in (3.6) and (3.10) (equivalent here due to exact identification); "QR" is quantile regression (no IV); "IV" is the usual (mean) IV estimator.

Table 1 also shows important finite-sample differences not captured by first-order asymptotics. With $n = 20$, for DGP 2 or 3, the lowest robust RMSE is actually that of QR: despite its (median) bias being the largest due to ignoring the endogeneity, its dispersion is so much smaller than the other estimators' dispersions that its overall robust RMSE is the smallest. This advantage persists to $n = 50$, but eventually $n$ is large enough for the (median) bias to dominate. In DGPs 2 and 3 that lack slope heterogeneity, the IV estimator is the most efficient when errors are Gaussian (and $n$ is large enough), but not with Cauchy errors, reflecting the greater efficiency of the median (over the mean) when errors are heavy-tailed.

Table 2 compares simulated robust RMSE of three smoothed estimators (all with $h = 0.1$) of log-linearized quantile Euler equations, specifically the EIS parameter. Compared to the GMM estimator with identity weighting matrix, the two-step GMM estimator is always more efficient. The biggest such advantage is with the smallest sample size, $n = 50$; this seems surprising since the two-step GMM's estimated weighting matrix has the largest variance in that case. Two-step GMM is not always better (or always worse) than the MM estimator that takes the linear projection of the (lone) endogenous regressor onto the vector of (five) instruments to be the second excluded instrument (in addition to the constant). Two-step GMM has a smaller robust RMSE in some cases, even half that of MM with $\tau = 0.25$ and $n = 500$, but in other cases MM has smaller robust RMSE, especially when $n = 50$. Perhaps the additional variance of the two-step estimator due to its use of a long-run variance estimator (for the weighting matrix) makes it less efficient in these cases, a phenomenon explored in non-quantile GMM by Hwang and Sun (2015). Alternatively, perhaps future work can improve the long-run variance estimator's precision, in turn improving the two-step estimator's precision.

Table 2: Simulated precision of smoothed estimators of EIS.

| | | | Robust RMSE | | | Median Bias | | |
|---|---|---|---|---|---|---|---|---|
| | | | | GMM | | | GMM | |
| DGP | $\tau$ | $n$ | MM | (2s) | (ID) | MM | (2s) | (ID) |
| 4 | 0.25 | 50 | 0.180 | 0.285 | 0.315 | 0.057 | 0.055 | 0.054 |
| 4 | 0.25 | 200 | 0.129 | 0.097 | 0.107 | 0.024 | 0.013 | 0.017 |
| 4 | 0.25 | 500 | 0.097 | 0.047 | 0.047 | 0.008 | 0.003 | 0.003 |
| 4 | 0.50 | 50 | 0.153 | 0.258 | 0.300 | 0.055 | 0.023 | 0.025 |
| 4 | 0.50 | 200 | 0.099 | 0.124 | 0.144 | 0.025 | 0.004 | −0.008 |
| 4 | 0.50 | 500 | 0.066 | 0.079 | 0.089 | 0.014 | −0.003 | −0.007 |

1000 replications. "MM" is the estimator in (3.6); "GMM(2s)" is the estimator in (3.11); "GMM(ID)" is the estimator in (3.10) with identity weighting matrix.

# 6 Application: quantile Euler equation

This section illustrates the usefulness of the proposed methods through an empirical example: the estimation of a quantile Euler equation. We apply the proposed methodology to an economic model of intertemporal allocation of consumption and estimate the elasticity of

intertemporal substitution (EIS). The EIS is a parameter of central importance in macroeconomics and finance. We refer to Campbell (2003), Cochrane (2005), and Ljungqvist and Sargent (2012), and the references therein, for a comprehensive overview.

There is a large empirical literature that attempts to estimate the EIS; among others, Hansen and Singleton (1983), Hall (1988), Campbell and Mankiw (1989), Campbell and Viceira (1999), Campbell (2003), and Yogo (2004). The majority of the literature relies on the traditional expected utility framework. The purpose of this application is to estimate and make inference on the EIS for selected developed countries in Campbell's (2003) data set using the quantile utility maximization model. The quantile model has useful advantages, such as robustness, ability to capture heterogeneity, and separation the notion of risk attitude from the intertemporal substitution.

Section 6.1 describes in detail the model that leads to the quantile Euler equation, establishing parallels with the standard expected utility Euler equation. Section 6.2 describes the estimation procedure, and Section 6.3 discusses log-linearization for the quantile model. Section 6.4 discusses an interpretation of the parameters in question. In Section 6.5 we review the data, and finally Section 6.6 presents the empirical results.

## 6.1 Description of the economic model

De Castro and Galvao (2017) employ a variation of the standard economy model of Lucas (1978). The economic agents decide on the intertemporal consumption and savings (assets to hold) over an infinity horizon economy, subject to a linear budget constraint. The decision generates an intertemporal policy function, which is used to estimate the parameters of interest for a given utility function. Their work is related to that of Giovannetti (2013), who works with a similar model but restricts the analysis to two periods, whereas de Castro and Galvao (2017) consider an infinite horizon.

The specific model is as follows. Let $C_t$ denote the amount of consumption good that the individual consumes in period $t$. At the beginning of period $t$, the consumer has $x_t$ units of the risky asset, which pays dividend $d_t$. The price of the consumption good is normalized to one, while the price of the risky asset in period $t$ is $p(d_t)$. Then, the consumer decides its consumption $C_t$ and how many units of the risky asset $x_{t+1}$ to save for the next period, subject to the budget constraint

$$C_t + p(d_t)x_{t+1} \leq [d_t + p(d_t)]x_t \tag{6.1}$$

and positivity restriction

$$C_t, x_{t+1} \geq 0. \tag{6.2}$$

In equilibrium, we have that $x_t^* = 1, \forall t, k$.

So far, the model is exactly the same as the standard Lucas' model, but the objective function will differ. In the standard model, the consumer maximizes

$$\mathrm{E}\left[\sum_{t=0}^{\infty} \beta^t U(C_t) \;\middle|\; \Omega_0\right] \tag{6.3}$$

subject to (6.1) and (6.2), where $\beta \in (0,1)$ is the discount factor, $U\colon \mathbb{R}_+ \mapsto \mathbb{R}$ is the utility function, and $\Omega_0$ is the information set at time $t = 0$. For the expected utility choice problem, dynamic consistency and the principle of optimality for (6.3) imply that at time $s \geq 1$, the consumer chooses $\{C_t, x_t\}_{t \geq s}$ to maximize

$$\mathrm{E}\left[\sum_{t=s}^{\infty} \beta^{t-s} U(C_t) \;\middle|\; \Omega_s\right] \tag{6.4}$$

subject to (6.1) and (6.2). The connection between problems (6.3) and (6.4) is made explicit by the linearity of the expectation operator and the law of iterated expectations:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=0}^{\infty} \beta^t U(C_t) \;\middle|\; \Omega_0\right] \\
&= U(C_0) + \mathrm{E}\Big[\beta U(C_1) + \mathrm{E}\Big[\beta^2 U(C_2) + \mathrm{E}\big[\beta^3 U(C_3) + \cdots \mid \Omega_2\big] \;\Big|\; \Omega_1\Big] \;\Big|\; \Omega_0\Big].
\end{aligned} \tag{6.5}$$

De Castro and Galvao (2017) replace the operator E in (6.5) with $\mathrm{Q}_\tau$, such that the consumer maximizes the following quantile objective function

$$\begin{aligned}
&U(C_0) + \mathrm{Q}_\tau\Big[\beta_\tau U(C_1) + \mathrm{Q}_\tau\Big[\beta_\tau^2 U(C_2) + \mathrm{Q}_\tau\big[\beta_\tau^3 U(C_3) + \cdots \mid \Omega_2\big] \;\Big|\; \Omega_1\Big] \;\Big|\; \Omega_0\Big] \\
&= \mathrm{Q}_\tau\Big[\mathrm{Q}_\tau\Big[\mathrm{Q}_\tau\big[U(C_0) + \beta_\tau U(C_1) + \beta_\tau^2 U(C_2) + \beta_\tau^3 U(C_3) + \cdots \mid \Omega_2\big] \;\Big|\; \Omega_1\Big] \;\Big|\; \Omega_0\Big] \\
&\equiv \mathrm{Q}_\tau^\infty\left[\sum_{t=0}^{\infty} \beta_\tau^t U(C_t)\right],
\end{aligned} \tag{6.6}$$

again subject to (6.1) and (6.2), where $\beta_\tau \in (0,1)$ is the discount factor for the quantile $\tau$.

Unfortunately, linearity and the law of iterated expectations do not hold for the $\tau$-quantile operator, $\mathrm{Q}_\tau$. Thus, in order to preserve dynamic consistency and the principle of optimality,

we need to maintain the structure developed in (6.6). De Castro and Galvao (2017) show that the limit above exists and is well defined. Moreover, they show that the quantile preferences are dynamically consistent, the principle of optimality holds, and the corresponding dynamic problem yields a value function, via a fixed-point argument. They further provide conditions so that the value function is differentiable and concave.

The structure in the right-hand side of (6.5) reflects the following associated value function,

$$v(x_t, d_t) = \max_{x_{t+1} \geq 0} \big\{ U\big([d_t + p(d_t)]x_t - p(d_t)x_{t+1}\big) + \beta \, \mathrm{E}[v(x_{t+1}, d_{t+1}) \mid \Omega_t] \big\}. \tag{6.7}$$

The value function for the quantile problem is the same as (6.7) but with $\mathrm{Q}_\tau$ replacing E:

$$v(x_t, d_t) = \max_{x_{t+1} \geq 0} \big\{ U\big([d_t + p(d_t)]x_t - p(d_t)x_{t+1}\big) + \beta_\tau \, \mathrm{Q}_\tau[v(x_{t+1}, d_{t+1}) \mid \Omega_t] \big\}. \tag{6.8}$$

In addition, de Castro and Galvao (2017) derive the corresponding Euler equation, using the fact that in equilibrium, the holdings are $x_t = 1$ for all $t$:

$$- p(d_t)U'(C_t) + \beta_\tau \, \mathrm{Q}_\tau[U'(C_{t+1})(z_{t+1} + p(z_{t+1})) \mid \Omega_t] = 0. \tag{6.9}$$

Defining the asset's return by

$$1 + r_{t+1} \equiv \frac{z_{t+1} + p(z_{t+1})}{p(d_t)},$$

the Euler equation in (6.9) simplifies to

$$\mathrm{Q}_\tau \left[ \beta_\tau (1 + r_{t+1}) \frac{U'(C_{t+1})}{U'(C_t)} \;\middle|\; \Omega_t \right] = 1. \tag{6.10}$$

After parameterizing the utility function, (6.10) is a conditional quantile restriction in the form of (2.1), as in our econometric model.

The quantile Euler equation in (6.10) looks similar to the standard Euler equation from expected utility maximization,

$$\mathrm{E} \left[ \beta (1 + r_{t+1}) \frac{U'(C_{t+1})}{U'(C_t)} \;\middle|\; \Omega_t \right] = 1. \tag{6.11}$$

The expressions inside the conditional quantile and conditional expectation are identical.

For obtaining the mentioned results, de Castro and Galvao (2017) assume the following.

**Assumption A13.** *(i) The dividends assume values in $\mathcal{Z} \subseteq \mathbb{R}$, which is a bounded interval, and $\mathcal{X} = [0, \bar{x}]$ for some $\bar{x} > 1$;*

*(ii)* $\{d_t\}$ *is a Markov process with PDF* $f\colon \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}_+$*, which is continuous, symmetric* $(f(a,b) = f(b,a))$*,* $f(d_t, d_{t+1}) > 0$ *for all* $(d_t, d_{t+1}) \in \mathcal{Z} \times \mathcal{Z}$*, and satisfies the property that if* $h\colon \mathcal{Z} \mapsto \mathbb{R}$ *is weakly increasing and* $z \leq z'$*, then*

$$\int_{\mathcal{Z}} h(\alpha) f(\alpha \mid z)\, d\alpha \leq \int_{\mathcal{Z}} h(\alpha) f(\alpha \mid z')\, d\alpha; \tag{6.12}$$

*(iii)* $U\colon \mathbb{R}_+ \mapsto \mathbb{R}$ *is given by* $U(c) = \frac{1}{1-\gamma} c^{1-\gamma}$*, for* $\gamma > 0$*;*

*(iv)* $z \mapsto z + p(z)$ *is* $C^1$ *and non-decreasing, with* $\frac{d}{dz} z[\ln(z + p(z))] \geq \gamma$*.*

Assumptions A13(i)–(iii) are standard in economic applications. In Assumption A13(iv), it is natural to expect that the price $p(z)$ is non-decreasing with the dividend $z$, and $z + p(z)$ being non-decreasing is an even weaker requirement.

## 6.2 Estimation

We follow a large body of the literature (e.g., Campbell, 2003) and use isoelastic utility,

$$U(C_t) = \frac{1}{1-\gamma} C_t^{1-\gamma}, \quad \gamma > 0. \tag{6.13}$$

The ratio of marginal utilities is

$$\frac{U'(C_{t+1})}{U'(C_t)} = \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}. \tag{6.14}$$

From (6.10) and (6.14), the Euler equation is thus

$$\mathrm{Q}_\tau\big[\beta_\tau (1 + r_{t+1})(C_{t+1}/C_t)^{-\gamma_\tau} - 1 \mid \Omega_t\big] = 0. \tag{6.15}$$

The quantile Euler equation in (6.15) is a conditional quantile restriction with finite-dimensional parameter vector $(\beta_\tau, \gamma_\tau)$, as in (2.1), so we may use smoothed (G)MM estimation.

## 6.3 Log-linearization

One benefit of the quantile Euler equation is that it may be log-linearized with no approximation error, unlike the standard Euler equation. One may rewrite (6.15) as

$$\mathrm{Q}_\tau[\epsilon_{t+1} \mid \Omega_t] = 1, \quad \epsilon_{t+1} \equiv \beta(1 + r_{t+1})(C_{t+1}/C_t)^{-\gamma}. \tag{6.16}$$

For general random variable $W$, $\mathrm{Q}_\tau[\ln(W)] = \ln(\mathrm{Q}_\tau[W])$ ("equivariance") since $\ln(\cdot)$ is strictly increasing and continuous. In contrast, $\mathrm{E}[\ln(W)] \leq \ln(\mathrm{E}[W])$ by Jensen's inequality. Continuing from (6.16),

$$\begin{aligned} \ln(\epsilon_{t+1}) &= \ln(\beta) + \ln(1 + r_{t+1}) - \gamma \ln(C_{t+1}/C_t), \\ \ln(C_{t+1}/C_t) &= \gamma^{-1} \ln(\beta) + \gamma^{-1} \ln(1 + r_{t+1}) - \gamma^{-1} \ln(\epsilon_{t+1}). \end{aligned} \tag{6.17}$$

If $\gamma > 0$, then since $\mathrm{Q}_\tau(W) = -\mathrm{Q}_{1-\tau}(-W)$ (and $0 = -0$),

$$\begin{aligned} 0 = \ln(1) &= \mathrm{Q}_\tau[\ln(\epsilon_{t+1}) \mid \Omega_t] = \mathrm{Q}_{1-\tau}[-\gamma^{-1} \ln(\epsilon_{t+1}) \mid \Omega_t] \\ &= \mathrm{Q}_{1-\tau}[\ln(C_{t+1}/C_t) - \gamma^{-1} \ln(\beta) - \gamma^{-1} \ln(1 + r_{t+1}) \mid \Omega_t]. \end{aligned}$$

Thus, $\ln(\beta)/\gamma$ and $1/\gamma$ are the intercept and slope (respectively) of the $1 - \tau$ IV quantile regression of $\ln(C_{t+1}/C_t)$ on a constant and $\ln(1 + r_{t+1})$, with instruments from $\Omega_t$.

Similarly, in the sample, the $g_{ni}$ should be equivalent for nonlinear and log-linear estimation. The nonlinear estimator has "residuals" $\Lambda_N = \epsilon_{t+1} - 1$, so $\Lambda_N \leq 0 \iff \epsilon_{t+1} \leq 1$. The log-linear estimator has $\Lambda_L = \gamma^{-1} \ln(\epsilon_{t+1})$, so $\Lambda_L \leq 0 \iff \ln(\epsilon_{t+1}) \leq 0$ (if $\gamma > 0$), which is equivalent to the nonlinear condition.

Because the nonlinear and log-linear quantile Euler equations are equivalent, the corresponding estimators should be identical. In our application, this is generally true (matching 2+ significant figures), although sometimes there are differences due to the numerical methods, especially simulated annealing, since the log-linear minimization is done in a transformed parameter space. Additionally, the nonlinear and log-linear estimators do not match when the latter is negative (implying misspecification) since the above arguments assume $\gamma > 0$.

## 6.4 Interpretation

The parameters of interest in (6.15) are $\beta_\tau$ and $\gamma_\tau$. The former is the usual discount factor. The parameter $1/\gamma_\tau$ is the standard measure of EIS implicit in the CRRA utility function in (6.13). The EIS is a measure of responsiveness of the consumption growth rate to the real interest rate. As in Hall (1988), in a model with uncertainty, the interpretation is similar, and a high value of EIS means that when the real interest rate is expected to be high, the consumer will actively defer consumption to the later period.

The interpretation of $1/\gamma_\tau$ as the EIS remains valid for the quantile maximization model.[13]

[13] Hall's (1988) argument that $\gamma$ fundamentally represents the EIS rather than risk aversion applies here, too.

Most directly, this can be seen in equation (6.17), where $1/\gamma$ is the derivative of $\ln(C_{t+1}/C_t)$ with respect to $\ln(1 + r_{t+1})$, holding $\epsilon_{t+1}$ constant.

## 6.5 Data

We use data originally from Campbell (2003) and provided by Yogo (2004).[14] It consists of aggregate level quarterly data for the United States (US), United Kingdom (UK), Australia (AUS), and Sweden (SWE). The sample period for the US is 1947Q3–1998Q4, UK is 1970Q3–1999Q1, Australia is 1970Q3–1998Q4, and Sweden is 1970Q3–1999Q2. Consumption is measured at the beginning of the period, consisting of nondurables plus services for the US and total consumption for the other countries, in real, per capita terms. The real interest rate deflates a proxy for the nominal short-term rate by the consumer price index. Instruments are lags of log real consumption growth, nominal interest rate, inflation, and a log dividend-price ratio for equities. For a complete description of the data, see Campbell (2003).

## 6.6 Results

Other than using quantiles, our estimation follows Table 2 of Yogo (2004, p. 805). Yogo (2004) uses 2SLS to estimate the (structural) log-linearized model $\ln(C_{t+1}/C_t) = \delta_0 + \delta_1 \ln(1 + r_{t+1}) + u_{t+1}$, where $r_{t+1}$ is the real interest rate, instrumenting for $\ln(1+r_{t+1})$ with twice lagged measures of nominal interest rate, inflation, consumption growth, and log dividend-price ratio, where $\delta_1 = 1/\gamma$ is the EIS and $\delta_0 = \ln(\beta)/\gamma$. Yogo (2004) emphasizes that these are strong instruments that predict the real interest rate well, although formally characterizing "strong" for IVQR remains an open question.[15]

Tables 3 and 4 show the quantile Euler equation estimates for $\beta_\tau$ and $\gamma_\tau$ (respectively), using the smoothed MM estimator in (3.6) and the smoothed two-step GMM estimator in (3.11), for the deciles $\tau = 0.1, \ldots, 0.9$. For two-step GMM, the long-run variance estimator follows Andrews (1991b) with a quadratic spectral kernel. For both estimators, the plug-in bandwidth from Kaplan and Sun (2017) was used. For comparison, 2SLS estimates are in each table's bottom row.

Tables 3 and 4 generally show economically unrealistic estimates at lower $\tau$ but plausible

[14] `https://sites.google.com/site/motohiroyogo/research/EIS_Data.zip`

[15] In contrast, when trying to estimate the EIS by 2SLS of $\ln(1 + r_{t+1})$ on $\ln(C_{t+1}/C_t)$, or replacing the real interest rate with a real stock index return, the instruments are weak because it is difficult to predict consumption growth or stock returns.

Table 3: Smoothed MM and GMM estimates of $\beta_\tau$, log-linear model.

| | US | | UK | | AUS | | SWE | |
|---|---|---|---|---|---|---|---|---|
| $\tau$ | $\hat{\beta}_{\text{MM}}$ | $\hat{\beta}_{\text{GMM}}$ | $\hat{\beta}_{\text{MM}}$ | $\hat{\beta}_{\text{GMM}}$ | $\hat{\beta}_{\text{MM}}$ | $\hat{\beta}_{\text{GMM}}$ | $\hat{\beta}_{\text{MM}}$ | $\hat{\beta}_{\text{GMM}}$ |
| 0.1 | 0.92 | 0.92 | 0.00 | 0.12 | 0.88 | 0.91 | 0.95 | 0.94 |
| 0.2 | 0.90 | 0.92 | 1.86 | 0.75 | 0.88 | 0.87 | 0.95 | 0.95 |
| 0.3 | 0.85 | 1.13 | 1.16 | 0.82 | 0.82 | 0.75 | 0.95 | 0.95 |
| 0.4 | 0.11 | 1.04 | 1.07 | 0.87 | 2.38 | 1.35 | 0.95 | 0.95 |
| 0.5 | 1.14 | 1.02 | 1.04 | 1.03 | 1.13 | 1.09 | 0.90 | 0.80 |
| 0.6 | 1.04 | 1.01 | 1.02 | 1.01 | 1.02 | 1.03 | 0.94 | 1.07 |
| **0.7** | **1.02** | **1.01** | **1.00** | **1.00** | **1.00** | **1.00** | **0.97** | **0.97** |
| **0.8** | **1.00** | **1.00** | **0.98** | **0.99** | **0.98** | **0.99** | **0.97** | **0.97** |
| 0.9 | 0.99 | 1.00 | 0.96 | 0.97 | 0.96 | 0.98 | 0.97 | 0.96 |
| 2SLS | 1.08 | | 1.03 | | 1.11 | | 0.27 | |

Table 4: Smoothed MM and GMM estimates of $\gamma_\tau$, log-linear model.

| | US | | UK | | AUS | | SWE | |
|---|---|---|---|---|---|---|---|---|
| $\tau$ | $\hat{\gamma}_{\text{MM}}$ | $\hat{\gamma}_{\text{GMM}}$ | $\hat{\gamma}_{\text{MM}}$ | $\hat{\gamma}_{\text{GMM}}$ | $\hat{\gamma}_{\text{MM}}$ | $\hat{\gamma}_{\text{GMM}}$ | $\hat{\gamma}_{\text{MM}}$ | $\hat{\gamma}_{\text{GMM}}$ |
| 0.1 | −7.2 | −7.1 | −744.0 | −114.1 | −6.9 | −5.5 | −3.4 | −3.7 |
| 0.2 | −11.5 | −8.0 | 43.1 | −19.0 | −9.0 | −9.1 | −4.0 | −4.4 |
| 0.3 | −20.0 | 16.5 | 13.5 | −15.7 | −17.0 | −24.6 | −5.5 | −6.1 |
| 0.4 | −342.5 | 6.9 | 8.3 | −14.1 | 98.9 | 33.9 | −8.7 | −10.2 |
| 0.5 | 26.5 | 4.5 | 7.3 | 5.3 | 20.7 | 14.8 | −38.7 | −89.5 |
| 0.6 | 10.9 | 3.7 | 6.6 | 4.1 | 7.3 | 8.0 | −151.0 | 301.5 |
| **0.7** | **6.6** | **3.4** | **5.9** | **3.7** | **5.9** | **5.1** | **13.9** | **13.6** |
| **0.8** | **4.9** | **3.2** | **5.3** | **3.7** | **5.0** | **3.7** | **5.3** | **6.4** |
| 0.9 | 4.9 | 3.1 | 5.0 | 4.0 | 4.4 | 0.0 | 3.3 | 4.8 |
| 2SLS | 16.7 | | 6.0 | | 22.1 | | -544.4 | |

estimates at larger $\tau$. For $\tau \leq 0.4$, some of the $\beta_\tau$ estimates are unrealistically far from one, and many of the $\gamma_\tau$ estimates are negative. For $\tau \geq 0.5$, in contrast, most of the $\beta_\tau$ estimates are close to one, and most $\gamma_\tau$ estimates seem plausible. For $\tau \in \{0.7, 0.8\}$ in particular, looking across all four countries and both MM and GMM estimates, the estimates are all contained within the ranges $\hat{\beta}_\tau \in [0.97, 1.02]$ and $\hat{\gamma}_\tau \in [3.2, 13.9]$.

The differences between MM and two-step GMM estimates are often relatively small, especially when the estimates are reasonable. However, the table shows some economically significant differences, such as for the US (even with $\tau \geq 0.6$).

The differences between the quantile and 2SLS estimates can be economically significant. This includes the case of Sweden, where the 2SLS estimates are entirely unrealistic: $\hat{\beta}_{2\text{SLS}} = 0.27$ and $\hat{\gamma}_{2\text{SLS}} = -544.4$. Although smaller $\tau$ produce unrealistic estimates, the Sweden quantile estimates for $\tau = 0.7$ and $\tau = 0.8$ have $\hat{\beta}_\tau = 0.98$ and $\hat{\gamma}_\tau$ in the range $[5.3, 13.9]$, all perfectly reasonable. The larger $\tau$ may better capture the response of consumption to interest rate changes; expectation-based Euler equations can struggle with the fact that consumption tends to evolve over time more smoothly than the real interest rate.[16] In fact, among the other seven countries whose data Yogo (2004) examined (Netherlands, Canada, France, Germany, Italy, Japan, Switzerland), all seven had negative 2SLS estimates $\hat{\gamma}_{2\text{SLS}} < 0$, but five of the seven had positive $\hat{\gamma}_\tau > 0$ with $\tau = 0.9$ (and plausible $\hat{\beta}_\tau$).

In all, this empirical application illustrates that the quantile utility maximization model and new smoothed estimators serve as important tools to study economic behavior.

# 7 Conclusion

For finite-dimensional parameters defined by general quantile-type restrictions, we have developed smoothed MM and GMM estimation and asymptotic theory, for exactly and over-identified models, respectively, allowing for weakly dependent data and nonlinear models. This includes nonlinear IV quantile regression and quantile Euler equations as special cases, and our theory is robust to misspecification of the structural models.

The empirical results suggest that quantile utility maximization combined with our smoothed estimation can provide a useful, economically meaningful alternative to estimation based on expected utility. A bonus feature is the ability to log-linearize the quantile Euler equation without any approximation error, unlike the standard Euler equation. Future work may apply our methods to household panel data or carefully consider how to determine $\tau$.

There is more to explore econometrically, too: quantile GMM inference (in progress), IVQR averaging estimators (in progress), optimal bandwidth choice, non/semiparametric models, fixed-smoothing asymptotic approximations, higher-order bootstrap refinements, formally establishing A11 when $D_i$ depends on $\beta_{0\tau}$, and results uniform in $\tau$, among other topics.

[16] We thank Duke Kao for this idea.

# A Proofs

*Proof of Proposition 2.1.* See Appendix A.1 of Chen et al. (2014). □

*Proof of Lemma 4.1.* Noting that $\left|Z[\tilde{I}(\cdot) - \mathbb{1}\{\cdot\}]\right| \leq |Z|$ (i.e., $|Z|$ is a dominating function) and applying the dominated convergence theorem (since $Z$ has finite expectation by A4), since $h_n \to 0$ by A6,

$$\begin{aligned}
&\lim_{h_n \to 0} \sup_{\beta \in \mathcal{B}} \left\| \mathrm{E}[\hat{M}_n(\beta, \tau)] - \mathrm{E}[Z(\mathbb{1}\{\Lambda(Y, X, \beta) \leq 0\} - \tau)] \right\| \\
&= \lim_{h_n \to 0} \max_{\beta \in \mathcal{B}} \left\| \mathrm{E}\left\{ Z \left[ \tilde{I}\left( \frac{-\Lambda(Y, X, \beta)}{h_n} \right) - \mathbb{1}\{\Lambda(Y, X, \beta) \leq 0\} \right] \right\} \right\| \\
&= \lim_{h_n \to 0} \left\| \mathrm{E}\left\{ Z \left[ \tilde{I}\left( \frac{-\Lambda(Y, X, \beta_n^*)}{h_n} \right) - \mathbb{1}\{\Lambda(Y, X, \beta_n^*) \leq 0\} \right] \right\} \right\| \\
&= \left\| \mathrm{E}\left\{ \lim_{h_n \to 0} Z \left[ \tilde{I}\left( \frac{-\Lambda(Y, X, \beta_n^*)}{h_n} \right) - \mathbb{1}\{\Lambda(Y, X, \beta_n^*) \leq 0\} \right] \right\} \right\| \\
&= 0
\end{aligned} \tag{A.1}$$

as long as there is no probability mass at $\Lambda(Y, X, \beta) = 0$ for any $\beta \in \mathcal{B}$ and almost all $Z$, which is indeed true by Assumption A7. The notation $\beta_n^*$ denotes the value attaining the maximum, which exists since $\mathcal{B}$ is compact by A3. □

*Proof of Theorem 4.2.* We first prove consistency of the smoothed method of moments estimator. We then prove consistency of the smoothed GMM estimator.

**MM Consistency.** To prove consistency of $\hat{\beta}_{\mathrm{MM}}$, we show that the conditions of Theorem 5.9 in van der Vaart (1998) are satisfied. Alternatively, one could apply Theorem 2.1 in Newey and McFadden (1994, p. 2121), where $\hat{\beta}$ maximizes $\hat{Q}_n(\beta) \equiv -\|\hat{M}_n(\beta, \tau)\|$ with $\hat{Q}_n(\hat{\beta}) = 0$.

Combining results from A8 and Lemma 4.1 and the triangle inequality,

$$\begin{aligned}
&\sup_{\beta \in \mathcal{B}} \left| \hat{M}_n(\beta, \tau) - M(\beta, \tau) \right| \\
&= \sup_{\beta \in \mathcal{B}} \left| \hat{M}_n(\beta, \tau) - \mathrm{E}\big[\hat{M}_n(\beta, \tau)\big] + \mathrm{E}\big[\hat{M}_n(\beta, \tau)\big] - M(\beta, \tau) \right| \\
&\leq \overbrace{\sup_{\beta \in \mathcal{B}} \left| \hat{M}_n(\beta, \tau) - \mathrm{E}\big[\hat{M}_n(\beta, \tau)\big] \right|}^{=o_p(1) \text{ by A8}} + \overbrace{\sup_{\beta \in \mathcal{B}} \left| \mathrm{E}\big[\hat{M}_n(\beta, \tau)\big] - M(\beta, \tau) \right|}^{=o_p(1) \text{ by Lemma 4.1}} \\
&= o_p(1) + o_p(1) = o_p(1).
\end{aligned} \tag{A.2}$$

This satisfies the first condition of Theorem 5.9 in van der Vaart (1998, p. 46), or (combined with the continuity of $\|\cdot\|$) condition (iv) in Theorem 2.1 of Newey and McFadden (1994).

For the second condition of Theorem 5.9 in van der Vaart (1998), since $\mathcal{B}$ is a compact subset of Euclidean space, so is the set

$$\{\beta : \|\beta - \beta_{0\tau}\| \geq \epsilon, \beta \in \mathcal{B}\}$$

for any $\epsilon > 0$. Writing out

$$\begin{aligned} M(\beta, \tau) &= \mathrm{E}\{Z_i[\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\} - \tau]\} \\ &= \mathrm{E}(\mathrm{E}\{Z_i[\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\} - \tau] \mid Z_i\}) \\ &= \mathrm{E}\big\{Z_i\big[\mathrm{P}\big(\Lambda(Y_i, X_i, \beta) \leq 0 \mid Z_i\big) - \tau\big]\big\}, \end{aligned}$$

we see that the function $M(\beta, \tau)$ is continuous in $\beta$ given A2 and A7. Note that A2 alone is not sufficient: it implies $\lim_{\delta \to 0} \Lambda(Y_i, X_i, \beta + \delta) \to \Lambda(Y_i, X_i, \beta)$ (for any realization $\omega \in \Omega$ in the implicit underlying probability space), but $\mathbb{1}\{\cdot \leq 0\}$ is not a continuous function. Specifically, it is discontinuous at zero, so the continuous mapping theorem only guarantees convergence (for $\omega \in \Omega$) where $\Lambda(Y_i, X_i, \beta) \neq 0$. Assumption A7 assumes this is a zero probability event (conditional on almost all $Z_i$), so $\mathbb{1}\{\Lambda(Y_i, X_i, \beta + \delta) \leq 0\}$ still converges almost surely to $\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\}$ as $\delta \to 0$ (i.e., the set of $\omega \in \Omega$ for which it does not converge has measure zero). Altogether, by A2 and A7, the bounded convergence theorem, and the continuous mapping theorem, writing $\Lambda_i \equiv \Lambda(Y_i, X_i, \beta)$,

$$\begin{aligned} &\lim_{\delta \to 0} \mathrm{P}\big(\Lambda(Y_i, X_i, \beta + \delta) \leq 0 \mid Z_i\big) \\ &= \lim_{\delta \to 0} \mathrm{E}\big(\mathbb{1}\{\Lambda(Y_i, X_i, \beta + \delta) \leq 0\} \mid Z_i\big) \\ &= \mathrm{E}\big(\lim_{\delta \to 0} \mathbb{1}\{\Lambda(Y_i, X_i, \beta + \delta) \leq 0\} \mid Z_i\big) \\ &= \mathrm{E}\big(\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\} \mid Z_i, \Lambda_i \neq 0\big) \overbrace{\mathrm{P}(\Lambda_i \neq 0 \mid Z_i)}^{=1 \text{ a.s., by A7}} \\ &\quad + \mathrm{E}\big(\lim_{\delta \to 0} \mathbb{1}\{\Lambda(Y_i, X_i, \beta + \delta) \leq 0\} \mid Z_i, \Lambda_i = 0\big) \overbrace{\mathrm{P}(\Lambda_i = 0 \mid Z_i)}^{=0 \text{ a.s., by A7}} \\ &= \mathrm{E}\big(\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \leq 0\} \mid Z_i\big) \end{aligned}$$

almost surely.

Since a continuous function on a compact set attains a minimum, letting $\beta^*$ denote the minimizer,

$$\inf_{\beta : \|\beta - \beta_{0\tau}\| \geq \epsilon} \|M(\beta, \tau)\| = \min_{\beta : \|\beta - \beta_{0\tau}\| \geq \epsilon} \|M(\beta, \tau)\| = \|M(\beta^*, \tau)\| > 0 \tag{A.3}$$

by A3, which says that for any $\beta \neq \beta_{0\tau}$, $M(\beta,\tau) \neq 0$, so $\|M(\beta^*,\tau)\| > 0$ (since $\|\cdot\|$ is a norm). Alternatively, for the conditions in Theorem 2.1 of Newey and McFadden (1994), (i) and (ii) are directly assumed in our A3, and (iii) is satisfied by the continuity of $M(\cdot,\tau)$ (as shown above).

Consistency of $\hat{\beta}_{\mathrm{MM}}$ follows by Theorem 5.9 in van der Vaart (1998) or Theorem 2.1 in Newey and McFadden (1994).

**GMM Consistency.** To prove the consistency of $\hat{\beta}_{\mathrm{GMM}}$, we show that the two conditions of Theorem 5.7 in van der Vaart (1998) are satisfied. The first condition of Theorem 5.7 in van der Vaart (1998) requires

$$\sup_{\beta\in\mathcal{B}} |\hat{M}_n(\beta,\tau)^\top \hat{W} \hat{M}_n(\beta,\tau) - M(\beta,\tau)^\top W M(\beta,\tau)| \xrightarrow{p} 0. \tag{A.4}$$

From (A.2), $\sup_{\beta\in\mathcal{B}} \|\hat{M}_n(\beta,\tau) - M(\beta,\tau)\| = o_p(1)$. From A12, $\hat{W} = W + o_p(1)$, which does not depend on $\beta$.

Let $\|\cdot\|$ denote the Frobenius matrix norm $\|A\| = \|A^\top\| = \sqrt{\mathrm{tr}(AA^\top)}$, which is the Euclidean norm if $A$ is a vector. Given this norm, the Cauchy–Schwarz inequality states that for any matrices $A$ and $B$, $\|AB\| \leq \|A\|\|B\|$.

We now use the triangle inequality, Cauchy–Schwarz inequality, uniform convergence in probability of $\hat{M}_n(\beta,\tau)$, and convergence in probability of $\hat{W}$, to show the required condition in (A.4):

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}} \big| \hat{M}_n(\beta,\tau)^\top \hat{W} \hat{M}_n(\beta,\tau) - M(\beta,\tau)^\top W M(\beta,\tau) \big| \\
&= \sup_{\beta\in\mathcal{B}} \big| (M(\beta,\tau) + (\hat{M}(\beta,\tau) - M(\beta,\tau)))^\top (W + (\hat{W} - W)) \\
&\qquad \times (M(\beta,\tau) + (\hat{M}(\beta,\tau) - M(\beta,\tau))) \\
&\qquad - M(\beta,\tau)^\top W M(\beta,\tau) \big| \\
&= \sup_{\beta\in\mathcal{B}} \big| M(\beta,\tau)^\top (\hat{W} - W) M(\beta,\tau) + M(\beta,\tau)^\top W (\hat{M}(\beta,\tau) - M(\beta,\tau)) \\
&\quad + (\hat{M}(\beta,\tau) - M(\beta,\tau))^\top W M(\beta,\tau) + M(\beta,\tau)^\top (\hat{W} - W)(\hat{M}(\beta,\tau) - M(\beta,\tau)) \\
&\quad + (\hat{M}(\beta,\tau) - M(\beta,\tau))^\top W (\hat{M}(\beta,\tau) - M(\beta,\tau)) \\
&\quad + (\hat{M}(\beta,\tau) - M(\beta,\tau))^\top (\hat{W} - W) M(\beta,\tau) \\
&\quad + (\hat{M}(\beta,\tau) - M(\beta,\tau))^\top (\hat{W} - W)(\hat{M}(\beta,\tau) - M(\beta,\tau)) \big| \\
&\leq \sup_{\beta\in\mathcal{B}} |M(\beta,\tau)^\top (\hat{W} - W) M(\beta,\tau)| + \sup_{\beta\in\mathcal{B}} |M(\beta,\tau)^\top W (\hat{M}(\beta,\tau) - M(\beta,\tau))|
\end{aligned}$$

$$
\begin{aligned}
&+\sup_{\beta\in\mathcal{B}}|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top W M(\beta,\tau)|+\sup_{\beta\in\mathcal{B}}|M(\beta,\tau)^\top(\hat{W}-W)(\hat{M}(\beta,\tau)-M(\beta,\tau))|\\
&+\sup_{\beta\in\mathcal{B}}|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top W(\hat{M}(\beta,\tau)-M(\beta,\tau))|\\
&+\sup_{\beta\in\mathcal{B}}|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top(\hat{W}-W)M(\beta,\tau)|\\
&+\sup_{\beta\in\mathcal{B}}|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top(\hat{W}-W)(\hat{M}(\beta,\tau)-M(\beta,\tau))|\\
=&\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top(\hat{W}-W)M(\beta,\tau)\|+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top W(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top W M(\beta,\tau)\|\\
&+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top(\hat{W}-W)(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top W(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top(\hat{W}-W)M(\beta,\tau)\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top(\hat{W}-W)(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
\le&\sup_{\beta\in\mathcal{B}}\overbrace{\|M(\beta,\tau)^\top\|\|(\hat{W}-W)\|\|M(\beta,\tau)\|}^{\text{by Cauchy–Schwarz inequality}}+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top\|\|W\|\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|W\|\|M(\beta,\tau)\|\\
&+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top\|\|(\hat{W}-W)\|\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|W\|\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|(\hat{W}-W)\|\|M(\beta,\tau)\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|(\hat{W}-W)\|\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
\le&\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top\|\|(\hat{W}-W)\|\overbrace{\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)\|}^{=O(1)}\\
&+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top\|\|W\|\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\overbrace{\|W\|}^{=O(1)}\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)\|\\
&+\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)^\top\|\|(\hat{W}-W)\|\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|\\
&+\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|W\|\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\|
\end{aligned}
$$

$$
\begin{aligned}
&+ \sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|(\hat{W}-W)\|\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)\| \\
&+ \sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))^\top\|\|(\hat{W}-W)\|\sup_{\beta\in\mathcal{B}}\|(\hat{M}(\beta,\tau)-M(\beta,\tau))\| \\
&= o_p(1)+o_p(1)+o_p(1)+o_p(1)+o_p(1)+o_p(1)+o_p(1)=o_p(1).
\end{aligned}
$$

Above, we know $\|W\| = O(1)$ since $W$ is fixed, and we know $\sup_{\beta\in\mathcal{B}}\|M(\beta,\tau)\| = O(1)$ since $M(\beta,\tau)$ is continuous in $\beta$ and $\mathcal{B}$ is a compact set.

The second condition of Theorem 5.7 in van der Vaart (1998) is that $\beta_{0\tau}$ satisfies the well-separated minimum property. Since $M(\beta,\tau)^\top W M(\beta,\tau)$ is continuous in $\beta$ and $\{\beta : \|\beta-\beta_{0\tau}\| \geq \epsilon, \beta\in\mathcal{B}\}$ is a compact set, let $\beta^*$ denote the minimizer: for any $\epsilon > 0$,

$$
\inf_{\beta:\|\beta-\beta_{0\tau}\|\geq\epsilon} M(\beta,\tau)^\top W M(\beta,\tau) = M(\beta^*,\tau)^\top W M(\beta^*,\tau). \tag{A.5}
$$

By A3, $M(\beta,\tau)\neq 0$ for any $\beta\neq\beta_{0\tau}$, so $M(\beta^*,\tau)\neq 0$. Since $W$ is positive definite (A12),

$$
M(\beta^*,\tau)^\top W M(\beta^*,\tau) > 0.
$$

Thus, for any $\epsilon > 0$,

$$
\inf_{\beta:\|\beta-\beta_{0\tau}\|\geq\epsilon} M(\beta,\tau)^\top W M(\beta,\tau) > 0 = M(\beta_{0\tau},\tau)^\top W M(\beta_{0\tau},\tau). \tag{A.6}
$$

Consistency of $\hat{\beta}_{\text{GMM}}$ follows by Theorem 5.7 in van der Vaart (1998). □

*Proof of Lemma 4.3.* Decomposing into a mean-zero term and a "bias" term,

$$
\sqrt{n}\hat{M}_n(\beta_{0\tau},\tau) = \overbrace{\sqrt{n}\big\{\hat{M}_n(\beta_{0\tau},\tau)-\mathrm{E}[\hat{M}_n(\beta_{0\tau},\tau)]\big\}}^{\xrightarrow{d}\mathrm{N}(0,\Sigma_\tau)\text{ by A10}} + \overbrace{\sqrt{n}\,\mathrm{E}[\hat{M}_n(\beta_{0\tau},\tau)]}^{\text{want to show } o_p(1)}.
$$

With iid data, Kaplan and Sun (2017, Thm. 1) show $\Sigma_\tau = \tau(1-\tau)\,\mathrm{E}(Z_i Z_i^\top)$. The remainder of the proof shows that the second term is indeed $o_p(1)$, actually $o(1)$.

Let $\Lambda_i \equiv \Lambda(Y_i, X_i, \beta_{0\tau})$, with marginal PDF $f_\Lambda(\cdot)$ and conditional PDF $f_{\Lambda|Z}(\cdot \mid z)$ given $Z_i = z$. Given strict stationarity of the data, using the definitions in (3.5), assuming the

support of $\Lambda_i$ given $Z_i = z$ is the interval $[\Lambda_L(z), \Lambda_H(z)]$ with $\Lambda_L(z) \le -h_n \le h_n \le \Lambda_H(z)$,

$$
\begin{aligned}
&\mathrm{E}[\hat{M}_n(\beta_{0\tau}, \tau)] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} g_n(Y_i, X_i, Z_i, \beta_{0\tau}, \tau)\right] = \mathrm{E}[g_n(Y_i, X_i, Z_i, \beta_{0\tau}, \tau)] \\
&= \mathrm{E}\Big\{ Z_i[\tilde{I}(-\Lambda_i/h_n) - \tau] \Big\} \\
&= \mathrm{E}\Big\{ Z_i \,\mathrm{E}[\tilde{I}(-\Lambda_i/h_n) - \tau \mid Z_i] \Big\} \\
&= \mathrm{E}\Bigg\{ Z_i \overbrace{\int_{\Lambda_L(Z_i)}^{\Lambda_H(Z_i)} [\tilde{I}(-L/h_n) - \tau]\, dF_{\Lambda|Z}(L \mid Z_i)}^{\text{integrate by parts}} \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ \overbrace{\Big(\tilde{I}(-L/h_n) - \tau\Big) F_{\Lambda|Z}(L \mid Z_i)\Big|_{\Lambda_L(Z_i)}^{\Lambda_H(Z_i)}}^{=-\tau\text{: use A5 and } \Lambda_H(Z_i) \ge h_n} \\
&\qquad\qquad - \int_{\Lambda_L(Z_i)}^{\Lambda_H(Z_i)} F_{\Lambda|Z}(L \mid Z_i) \overbrace{\tilde{I}'(-L/h_n)}^{=0 \text{ for } L \notin [-h_n, h_n]} (-h_n^{-1})\, dL \Bigg] \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ -\tau + \overbrace{h_n^{-1} \int_{-h_n}^{h_n} F_{\Lambda|Z}(L \mid Z_i) \tilde{I}'(-L/h_n)\, dL}^{\text{change of variables to } v = -L/h_n} \Bigg] \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ -\tau + \int_{-1}^{1} F_{\Lambda|Z}(-h_n v \mid Z_i) \tilde{I}'(v)\, dv \Bigg] \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ -\tau + \int_{-1}^{1} \left( \sum_{k=0}^{r} F_{\Lambda|Z}^{(k)}(0 \mid Z_i) \frac{(-h_n)^k v^k}{k!} \right) \tilde{I}'(v)\, dv \Bigg] \Bigg\} \\
&\quad + \mathrm{E}\Bigg\{ Z_i \int_{-1}^{1} \overbrace{f_{\Lambda|Z}^{(r)}(-\tilde{h} v \mid Z_i)}^{\tilde{h} \in [0, h_n] \text{ (from MVT)}} \frac{(-h_n)^{r+1} v^{r+1}}{(r+1)!} \tilde{I}'(v)\, dv \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ -\tau + \sum_{k=0}^{r} F_{\Lambda|Z}^{(k)}(0 \mid Z_i) \frac{(-h_n)^k}{k!} \overbrace{\int_{-1}^{1} v^k \tilde{I}'(v)\, dv}^{=0 \text{ for } 1 \le k \le r-1 \text{ by A5}} \Bigg] \Bigg\} \\
&\quad + O(h_n^{r+1}) \overbrace{\mathrm{E}\Bigg\{ Z_i \int_{-1}^{1} \overbrace{f_{\Lambda|Z}^{(r)}(-\tilde{h} v \mid Z_i)}^{\text{bounded by A9}} v^{r+1} \tilde{I}'(v)\, dv \Bigg\}}^{O(1) \text{ by A5 and A9}} \\
&= \mathrm{E}\Bigg\{ Z_i \Bigg[ -\tau + F_{\Lambda|Z}(0 \mid Z_i) + f_{\Lambda|Z}^{(r-1)}(0 \mid Z_i) \frac{(-h_n)^r}{r!} \int_{-1}^{1} v^r \tilde{I}'(v)\, dv \Bigg] \Bigg\} + O(h_n^{r+1}) \\
&= \mathrm{E}\{ Z_i[-\tau + \mathrm{E}(\mathbb{1}\{\Lambda_i \le 0\} \mid Z_i)] \} + \overbrace{\frac{(-h_n)^r}{r!}}^{r \text{ is even}} \left[ \int_{-1}^{1} v^r \tilde{I}'(v)\, dv \right] \mathrm{E}\Big[ Z_i f_{\Lambda|Z}^{(r-1)}(0 \mid Z_i) \Big] + O(h_n^{r+1})
\end{aligned}
$$

$$= \overbrace{\mathrm{E}\{\mathrm{E}[Z_i(\mathbb{1}\{\Lambda_i \le 0\} - \tau) \mid Z_i]\}}^{=\mathrm{E}[Z_i(\mathbb{1}\{\Lambda_i \le 0\} - \tau)] = 0 \text{ by A3}} + \frac{h_n^r}{r!}\left[\int_{-1}^{1} v^r \tilde{I}'(v)\, dv\right] \mathrm{E}\Big[Z_i f_{\Lambda|Z}^{(r-1)}(0 \mid Z_i)\Big] + O(h^{r+1})$$

$$= \frac{h_n^r}{r!}\left[\int_{-1}^{1} v^r \tilde{I}'(v)\, dv\right] \mathrm{E}\Big[Z_i f_{\Lambda|Z}^{(r-1)}(0 \mid Z_i)\Big] + O(h_n^{r+1}) = O(h_n^r).$$

Thus, the result follows if $\sqrt{n}h_n^r = o(1)$, i.e., $h_n = o(n^{-1/(2r)})$ as in A6. □

*Proof of Theorem 4.4.* We first establish the asymptotic normality of the smoothed MM estimator. We then prove the asymptotic normality of the smoothed GMM estimator.

**MM asymptotic normality.** Recall from (3.6) that $0 = \hat{M}_n(\hat{\beta}_{\mathrm{MM}}, \tau)$. Define

$$\nabla_{\beta^\top} \hat{M}_n(\beta_{0\tau}, \tau) \equiv \frac{\partial}{\partial \beta^\top} \hat{M}_n(\beta, \tau)\Big|_{\beta=\beta_{0\tau}}. \tag{A.7}$$

Let $\hat{M}_n^{(k)}(\beta, \tau)$ refer to the $k$th element in the vector $\hat{M}_n(\beta, \tau)$, so $\nabla_{\beta^\top} \hat{M}_n^{(k)}(\beta_{0\tau}, \tau)$ is a row vector and $\nabla_\beta \hat{M}_n^{(k)}(\beta_{0\tau}, \tau)$ is a column vector. Define

$$\dot{M}_n(\tau) \equiv \Big(\nabla_\beta \hat{M}_n^{(1)}(\tilde{\beta}_{(1)}, \tau), \ldots, \nabla_\beta \hat{M}_n^{(d_\beta)}(\tilde{\beta}_{(d_\beta)}, \tau)\Big)^\top, \tag{A.8}$$

a $d_\beta \times d_\beta$ matrix with its first row equal to that of $\nabla_{\beta^\top} \hat{M}_n(\tilde{\beta}_{(1)}, \tau)$, its second row equal to that of $\nabla_{\beta^\top} \hat{M}_n(\tilde{\beta}_{(2)}, \tau)$, etc., where each vector $\tilde{\beta}_{(k)}$ lies on the line segment between $\beta_{0\tau}$ and $\hat{\beta}_{\mathrm{MM}}$. Due to smoothing, we can take a derivative (for any $n$) to obtain a mean value expansion, and then rearrange:

$$0 = \hat{M}_n(\beta_{0\tau}, \tau) + \dot{M}_n(\tau)(\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau}), \tag{A.9}$$

$$\sqrt{n}(\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau}) = -[\dot{M}_n(\tau)]^{-1} \sqrt{n} \hat{M}_n(\beta_{0\tau}, \tau). \tag{A.10}$$

From A11, after plugging in definitions, $\dot{M}_n(\tau) \xrightarrow{p} G$; applying the continuous mapping theorem, $-[\dot{M}_n(\tau)]^{-1} \xrightarrow{p} -G^{-1}$. Using A10, the rest of the right-hand side of (A.10) has an asymptotic normal distribution. Equation (A.10) also implies the asymptotically linear (influence function) representation

$$\sqrt{n}(\hat{\beta}_{\mathrm{MM}} - \beta_{0\tau}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [-\dot{M}_n(\tau)]^{-1} g_{ni}(\beta_{0\tau}, \tau) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} G^{-1} g_{ni}(\beta_{0\tau}, \tau) + o_p(1). \tag{A.11}$$

Next, apply the continuous mapping theorem (CMT), using the nonsingularity of $G$ assumed in A9 and the result $\dot{M}_n(\tau) \xrightarrow{p} G$ in A11 to obtain $[\dot{M}_n(\tau)]^{-1} \xrightarrow{p} G^{-1}$. Using the CMT again, combine this with the results in (A.10) and Lemma 4.3:

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{\text{MM}} - \beta_{0\tau}) &= -\overbrace{[\dot{M}_n(\tau)]^{-1}}^{\text{use A11 and CMT}} \overbrace{\sqrt{n}\hat{M}_n(\beta_{0\tau}, \tau)}^{\text{use Lemma 4.3}} \\ &\xrightarrow{d} -G^{-1}\text{N}(0, \Sigma_\tau) \overset{d}{=} \text{N}\big(0, G^{-1}\Sigma_\tau[G^\top]^{-1}\big).\end{aligned}$$

**GMM asymptotic normality.** For GMM, the approach is similar, but starting from the first-order condition for the mean value expansion. From the definition of $\hat{\beta}_{\text{GMM}}$ in (3.10), we have the first-order condition

$$\Big[\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)\Big]^\top \hat{W}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau) = 0. \tag{A.12}$$

We reuse the notation from (A.7) and (A.8), but now $\tilde{\beta}_{(k)}$ lies between $\beta_{0\tau}$ and $\hat{\beta}_{\text{GMM}}$. By the mean value theorem,

$$\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau) = \hat{M}_n(\beta_{0\tau}, \tau) + \dot{M}_n(\tau)(\hat{\beta}_{\text{GMM}} - \beta_{0\tau}). \tag{A.13}$$

Pre-multiplying (A.13) by $[\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)]^\top\hat{W}$ and using (A.12) for the first equality,

$$\begin{aligned}0 &= [\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)]^\top\hat{W}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau) \\ &= [\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)]^\top\hat{W}\hat{M}_n(\beta_{0\tau}, \tau) + [\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)]^\top\hat{W}\dot{M}_n(\tau)(\hat{\beta}_{\text{GMM}} - \beta_{0\tau}).\end{aligned}$$

Multiplying by $\sqrt{n}$ and rearranging,

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_{\text{GMM}} - \beta_{0\tau}) \\ &= -\big\{[\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)]^\top\hat{W}\dot{M}_n(\tau)\big\}^{-1}\Big[\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau)\Big]^\top\hat{W}\overbrace{\sqrt{n}\hat{M}_n(\beta_{0\tau}, \tau)}^{=O_p(1)} \qquad \text{(A.14)} \\ &= -\{G^\top WG\}^{-1}G^\top W\sqrt{n}\hat{M}_n(\beta_{0\tau}, \tau) + o_p(1), \qquad \text{(A.15)}\end{aligned}$$

where $\hat{W} = W + o_p(1)$ by A12, and $\dot{M}_n(\tau) = G + o_p(1)$ and $\nabla_{\beta^\top}\hat{M}_n(\hat{\beta}_{\text{GMM}}, \tau) = G + o_p(1)$ by A11. From Lemma 4.3, $\sqrt{n}\hat{M}_n(\beta_{0\tau}, \tau) \xrightarrow{d} \text{N}(0, \Sigma_\tau)$. Applying the continuous mapping theorem yields the stated result. □

# B Primitive conditions for high-level assumptions

The following subsections discuss primitive conditions for the high-level Assumptions A8, A10, and A11.

## B.1 Assumption A8

Assumption A8 is a high-level ULLN-type assumption. Intuitively, it holds under weak enough dependence and a moment restriction on $Z_i$. Howver, it is not trivial since most ULLNs assume a constant function $g(\cdot)$ instead of a function indexed by $n$. We provide an example of sufficient lower-level assumptions in Lemma B.1.

**Lemma B.1.** *Let Assumptions A1–A5 and A7 hold. Additionally, assume the following. (i)* $\big(\mathcal{B}, d(\cdot)\big)$ *is a metric space. (ii) Defining open balls* $B(\beta,\rho) \equiv \{\tilde{\beta} \in \mathcal{B} : d(\beta,\tilde{\beta}) < \rho\}$,

$$\begin{aligned} g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho) &\equiv \sup\Big\{ g_n(Y_i, X_i, Z_i, \tilde{\beta}, \tau) : \tilde{\beta} \in B(\beta,\rho) \Big\}, \\ g_{*n}(Y_i, X_i, Z_i, \beta, \tau, \rho) &\equiv \inf\Big\{ g_n(Y_i, X_i, Z_i, \tilde{\beta}, \tau) : \tilde{\beta} \in B(\beta,\rho) \Big\} \end{aligned} \tag{B.1}$$

*are random variables for all* $i$, $\beta \in \mathcal{B}$, *and sufficiently small* $\rho$ *(which may depend on* $\beta$*), where the* sup *and* inf *are taken separately for each element of the vector. (iii) A pointwise WLLN holds for the random vectors in* (B.1)*, for each* $\beta \in \mathcal{B}$*. (iv) The data are strictly stationary. Then, for a fixed* $\tau \in (0,1)$*, using the definition in* (3.5)*,*

$$\sup_{\beta\in\mathcal{B}} \Big| \hat{M}_n(\beta,\tau) - \mathrm{E}[\hat{M}_n(\beta,\tau)] \Big| = o_p(1).$$

*Proof.* We show that the theorem in Andrews (1987) applies. The theorem concerns a uniform law of large numbers (ULLN) for a sample average of functions of the data. By Comment 6 in Andrews (1987), both the data and the functions may be indexed by both $i$ and $n$. In our case, the function $g_n(\cdot)$ is not indexed by $i$ but must be indexed by $n$ since it depends on the sequence of bandwidths, $h_n$. We continue to index the observations only by $i$ but note that triangular arrays are permitted by Andrews (1987). Since Andrews (1987) presumes a scalar-valued function, we write $g_n(\cdot)$; since the dimension of $g_n(\cdot)$ is fixed and finite, uniformity extends immediately to the vector.

Assumption A1 in Andrews (1987) is simply that $\mathcal{B}$ is compact, which is in our A3. (More recent work shows that "compact" can be replaced by "totally bounded" under a metric; see Andrews (1992) and Pötscher and Prucha (1994).)

Assumption A2(a) in Andrews (1987) is a technical measurability assumption; this is assumption (ii) in the statement of Lemma B.1.

Assumption A2(b) in Andrews (1987) is assumption (iii) in the statement of the lemma. There are many WLLNs for weakly dependent triangular arrays, where dependence is quantified and restricted in various ways; for example, see Theorem 2 in Andrews (1988) and the

theorems in de Jong (1998). With iid sampling, sufficient primitive conditions for a WLLN are already in our A4 and A5, respectively: a) $\mathrm{E}\big(\|Z_i\|^2\big) < \infty$, and b) $\tilde{I}(\cdot)$ is bounded. From A5, $-2 \leq \tilde{I}(\cdot) - \tau \leq 2$, so we have the dominating function $|g_n(Y_i, X_i, Z_i, \beta, \tau)| \leq 2|Z_i|$. Consequently,

$$|g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho)| \leq 2|Z_i|, \quad |g_{*n}(Y_i, X_i, Z_i, \beta, \tau, \rho)| \leq 2|Z_i|.$$

If the data are iid, then $g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho)$ is a row-wise iid triangular array. Thus, a sufficient condition for a WLLN is $\sup_n \mathrm{E}\big[\|g_n^*\|^2\big] < \infty$ (as can be shown with Markov's inequality). This condition holds since $\sup_n \mathrm{E}\big[\|g_n^*\|^2\big] \leq \mathrm{E}\big[\|2Z_i\|^2\big] < \infty$ by A4. An extension to independent but not identical sampling follows from a Lindeberg condition for $Z_i$. A pointwise WLLN continues to hold with dependence, too, as long as the dependence is not too strong.

Assumption A3 in Andrews (1987) in our notation is

$$\lim_{\rho \to 0} \sup_{n \geq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \{\mathrm{E}[g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho)] - \mathrm{E}[g_n(Y_i, X_i, Z_i, \beta, \tau)]\} \right| = 0, \tag{B.2}$$

and similarly when replacing $g_n^*$ with $g_{*n}$. Since $g_n$ varies with $n$ but not $i$, the strict stationarity in assumption (iv) in Lemma B.1 implies the summands do not vary with $i$, which simplifies (B.2) to be

$$\lim_{\rho \to 0} \sup_{n \geq 1} |\Delta_n| = 0, \ \Delta_n \equiv \mathrm{E}[g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho)] - \mathrm{E}[g_n(Y_i, X_i, Z_i, \beta, \tau)], \ \Delta_\infty \equiv \lim_{n \to \infty} \Delta_n. \tag{B.3}$$

Strict stationarity is not necessary, though, as long as (B.2) still holds.

A necessary condition for (B.3) is pointwise convergence $\lim_{\rho \to 0} \Delta_n = 0$ for each fixed $n$. By A3 and A5, $g_{ni}(\beta, \tau)$ is continuous and even differentiable in $\beta$. Additionally, as noted above, $2|Z_i|$ is a dominating function with finite expectation (by A4), so the dominated convergence theorem gives

$$\begin{aligned}
\lim_{\rho \to 0} \Delta_n &= \lim_{\rho \to 0} \mathrm{E}[g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho) - g_n(Y_i, X_i, Z_i, \beta, \tau)] \\
&= \mathrm{E}\left\{ \lim_{\rho \to 0} [g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho) - g_n(Y_i, X_i, Z_i, \beta, \tau)] \right\} \\
&= \mathrm{E}\{0\} = 0,
\end{aligned}$$

and similarly for $g_{*n}$, for any $n$.

For $\Delta_\infty$, as $n \to \infty$, we can again move the limit inside expectations by the dominated convergence theorem, so

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E}[g_n(Y_i, X_i, Z_i, \beta, \tau)] &= \mathrm{E}\Big\{ \lim_{n\to\infty} g_n(Y_i, X_i, Z_i, \beta, \tau) \Big\} \\
&= \mathrm{E}\{Z_i[\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \le 0\} - \tau]\} \\
&= \mathrm{E}\{\mathrm{E}[Z_i(\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \le 0\} - \tau) \mid Z_i]\} \\
&= \mathrm{E}\{Z_i[\mathrm{E}(\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \le 0\} \mid Z_i) - \tau]\} \\
&= \mathrm{E}\{Z_i[\mathrm{P}(\Lambda(Y_i, X_i, \beta) \le 0 \mid Z_i) - \tau]\}.
\end{aligned}$$

Technically, since $\tilde{I}(0/h_n) = 0.5$ for any $h_n > 0$, the function $\tilde{I}(\cdot/h_n) \to \mathbb{1}\{\cdot \ge 0\} - 0.5\,\mathbb{1}\{\cdot = 0\}$ as $n \to \infty$, so we have

$$\begin{aligned}
&\mathrm{E}\{Z_i[\mathrm{E}(\mathbb{1}\{\Lambda(Y_i, X_i, \beta) \le 0\} - 0.5\,\mathbb{1}\{\Lambda(Y_i, X_i, \beta) = 0\} \mid Z_i) - \tau]\} \\
&= \mathrm{E}\left\{ Z_i \left[ \mathrm{P}(\Lambda(Y_i, X_i, \beta) \le 0 \mid Z_i) - 0.5 \overbrace{\mathrm{P}(\Lambda(Y_i, X_i, \beta) = 0 \mid Z_i)}^{=0 \text{ a.s. by A7}} - \tau \right] \right\}.
\end{aligned}$$

That is, by A7, the 0.5 adjustment corresponds to a zero probability event that does not affect the overall expectation. For $g_n^*$, similarly,

$$\lim_{n\to\infty} \mathrm{E}[g_n^*(Y_i, X_i, Z_i, \beta, \tau, \rho)] = \mathrm{E}\left\{ \sup_{\tilde{\beta} \in B(\beta,\rho)} Z_i \Big[ \mathrm{P}\Big( \Lambda(Y_i, X_i, \tilde{\beta}) \le 0 \mid Z_i \Big) - \tau \Big] \right\}.$$

Consequently,

$$\Delta_\infty = \mathrm{E}\left\{ Z_i[\mathrm{P}(\Lambda(Y_i, X_i, \beta) \le 0 \mid Z_i) - \tau] - \sup_{\tilde{\beta} \in B(\beta,\rho)} Z_i \Big[ \mathrm{P}\Big( \Lambda(Y_i, X_i, \tilde{\beta}) \le 0 \mid Z_i \Big) - \tau \Big] \right\}.$$

For this to have a limit of zero as $\rho \to 0$ again requires continuity in $\rho$, but the necessary and sufficient conditions are different than for fixed $n$. Sufficient conditions here are found in A2 and A7: $\Lambda(\cdot)$ is continuous in $\beta$, and for any $\beta \in \mathcal{B}$ and almost all $z \in \mathcal{Z}$, the conditional distribution of $\Lambda(Y_i, X_i, \beta)$ given $Z_i = z$ is continuous in a neighborhood of zero. For example, if $\Lambda(Y_i, X_i, \beta) = Y_i - X_i^\top \beta$, then it is sufficient that $Y_i$ has a continuous distribution given almost all $Z_i = z$.

Given $\lim_{\rho\to 0} \Delta_n = 0$ for any $n < \infty$ and $n = \infty$, the conclusion $\lim_{\rho\to 0} \sup_{n\ge 1} |\Delta_n| = 0$ follows because the supremum is attained: $\sup_{n\ge 1} |\Delta_n| = |\Delta_k|$ for some $k \ge 1$ or $k = \infty$. If instead $\lim_{\rho\to 0} \Delta_n = 0$ only for $n \ge 1$, and not with $\lim_{n\to\infty}$, then it would be possible for all

$\lim_{\rho\to 0}\Delta_n = 0$ pointwise but $\lim_{\rho\to 0}\sup_{n\geq 1}\Delta_n \neq 0$; for example if $\Delta_n = (1-1/n)^{1/\rho}$ then all $\lim_{\rho\to 0}\Delta_n = 0$ but $\sup_{n\geq 1}\Delta_n = 1$ for any $\rho$, so $\lim_{\rho\to 0}\sup_{n\geq 1}\Delta_n = 1$. This is why the calculations for $\Delta_\infty$ are necessary.

Having verified A1, A2, and A3 in Andrews (1987), his theorem applies, yielding the desired ULLN. □

## B.2 Assumption A10

To establish Assumption A10, with iid data, the Lindeberg–Feller CLT can be applied as in the proof of Theorem 1 in Kaplan and Sun (2017). More generally, A10 can hold under weak dependence. For example, Theorem 3.13 in Wooldridge (1986, Ch. 2), as reproduced in Proposition 1 of Andrews (1991a), is a CLT for near epoch dependent triangular arrays that holds under some moment and dependence restrictions. The moment restriction, condition (ii), in our notation is $\mathrm{E}\{\|g_{ni}(\beta_{0\tau},\tau)\|^{2+\epsilon}\} < \infty$ for some $\epsilon > 0$. Since $|g_{ni}(\beta_{0\tau},\tau)| < 2|Z_i|$, if the underlying $Z_i$ are strictly stationary then $\mathrm{E}(\|Z_i\|^{2+\epsilon}) < \infty$ is sufficient; triangular array data are allowed if $\sup_{i\leq n, n\geq 1}\mathrm{E}(\|Z_{ni}\|^{2+\epsilon}) < \infty$ for some $\epsilon > 0$.

## B.3 Assumption A11

Unfortunately, for multiple reasons, Assumption A11 cannot be deduced simply by applying a result like Lemma 4.3 in Newey and McFadden (1994, p. 2156). Fortunately, it is closely related to the well-studied result of consistency of the kernel estimator for the quantile regression asymptotic covariance matrix. Since the argument is the same for each row in the matrix, we consider row $k$. Plugging in definitions,

$$\begin{aligned}\dot{M}_n^{(k,\cdot)}(\tau) = \nabla_{\beta^\top}\hat{M}_n^{(k)}(\tilde{\beta},\tau) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta^\top}g_{ni}^{(k)}(\beta,\tau)\bigg|_{\beta=\tilde{\beta}} \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i^{(k)}\tilde{I}'\big(-\Lambda(Y_i,X_i,\tilde{\beta})/h_n\big)(-h_n^{-1})D_i(\tilde{\beta})^\top, \qquad \text{(B.4)}\\ D_i(b) &\equiv \frac{\partial}{\partial\beta}\Lambda(Y_i,X_i,\beta)\bigg|_{\beta=b}. \qquad \text{(B.5)}\end{aligned}$$

By A5, $\tilde{I}'(\cdot)$ is a kernel function. The RHS of (B.4) is closely related to the kernel estimator of the usual quantile regression estimator's asymptotic covariance matrix initially proposed under censoring by Powell (1984, eqn. (5.6)) and without censoring in Powell (1991), but with two differences: 1) we have $\tilde{\beta}$ instead of $\hat{\beta}$, 2) we have a more general model. As a

special case of our model, the "usual" quantile regression model would have $Z_i = X_i$ and $\Lambda(Y_i, X_i, \beta) = Y_i - X_i^\top \beta$, so $\nabla_\beta \Lambda(Y_i, X_i, \tilde{\beta}) = -X_i$, and (B.4) simplifies to

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{(k)} \tilde{I}'\big((X_i^\top \tilde{\beta} - Y_i)/h_n\big)(-h_n^{-1})(-X_i^\top) = \frac{1}{nh_n}\sum_{i=1}^{n} \tilde{I}'\left(\frac{Y_i - X_i^\top \tilde{\beta}}{h_n}\right) X_i^{(k)} X_i^\top, \qquad \text{(B.6)}$$

using the symmetry $\tilde{I}'(-u) = \tilde{I}'(u)$ from A5. Since $\tilde{\beta}$ lies between $\beta_{0\tau}$ and $\hat{\beta}_{\text{MM}}$, proofs using $\hat{\beta}_{\text{MM}}$ still hold since $\sqrt{n}$-consistency of $\hat{\beta}_{\text{MM}}$ implies $\sqrt{n}$-consistency of $\tilde{\beta}$.

For (B.6), Kato (2012) shows consistency (i.e., our A11) with both iid and weakly dependent data. In fact, he shows the stronger result of asymptotic normality, so some of his assumptions may be weakened if only consistency is required; for example, his $h_n\sqrt{n}/\log(n) \to \infty$ in Assumption 13 can be (slightly) weakened to $h_n\sqrt{n} \to \infty$, and not as many moments of $X_i$ are required. Specifically, Kato (2012) considers strictly stationary $\beta$-mixing data, and the mixing coefficients $\beta(j)$, moments of $X_i$, and bandwidth rate are jointly restricted in his Assumptions 9, 10, and 13 (p. 268): $\sum_{j=1}^{\infty} j^\lambda [\beta(j)]^{1-2/\delta} < \infty$ for some $\delta > 2$ and $\lambda > 1 - 2/\delta$; $\mathrm{E}[\|X_i\|^{\max\{6,2\delta\}}] < \infty$; and for some integer sequence $s_n \to \infty$ with $s_n = o(\sqrt{nh_n})$, $(n/h_n)^{1/2}\beta(s_n) \to 0$.

For the more general (B.4), similar conditions are sufficient if $D_i(\beta) = D_i$, a random variable depending on $Y_i$ and $X_i$ but not the argument $\beta$. This occurs if (and only if) the residual function $\Lambda(\cdot)$ is linear-in-parameters. Then, $D_i$ replaces one of the $X_i$ in Kato (2012), while $Z_i$ replaces the other. The most notable restriction is on moments of $D_i$ (which implies a certain number of finite moments for $Y_i$ and $X_i$) in addition to $Z_i$ (which is already in A4). Many economic variables are bounded or reasonably have infinite moments (e.g., a normal distribution), in which case such moment assumptions are not binding. If $D_i(b)$ does depend on its argument, then an extension of the argument itself in Kato (2012) is necessary.

In (B.6), $X_i$ plays the roles of both the derivative of $\Lambda(Y_i, X_i, \beta)$ and the instrument vector, so the PDF in $G$ is just conditional on $X_i$; more generally, both the instrument vector and derivative must be conditioned on. This can be seen by computing the expectation of (B.4) in a similar manner to the proof of Lemma 4.3. After replacing $\tilde{\beta} = \beta_{0\tau} + O_p(n^{-1/2})$ and dropping the remainder, letting $D_i \equiv \nabla_\beta \Lambda(Y_i, X_i, \beta_{0\tau})$ and $\Lambda_i \equiv \Lambda(Y_i, X_i, \beta_{0\tau})$,

$$\begin{aligned}
\mathrm{E}[\dot{M}_n^{(k,\cdot)}(\tau)] &\doteq \frac{1}{n}\sum_{i=1}^{n} \overbrace{\mathrm{E}\left\{ Z_i^{(k)} \tilde{I}'\big(-\Lambda(Y_i, X_i, \beta_{0\tau})/h_n\big)(-h_n^{-1}) \frac{\partial}{\partial \beta^\top} \Lambda(Y_i, X_i, \beta)\Big|_{\beta=\tilde{\beta}} \right\}}^{\text{same for all } i \text{ by A1}} \\
&= \mathrm{E}\Big[ Z_i^{(k)} D_i^\top (-h_n^{-1}) \tilde{I}'\big(-\Lambda_i/h_n\big) \Big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\Big\{ Z_i^{(k)} D_i^\top \, \mathrm{E}\Big[ (-h_n^{-1}) \tilde{I}'\big(-\Lambda_i/h_n\big) \mid Z_i, D_i \Big] \Big\} \\
&= \mathrm{E}\Bigg\{ Z_i^{(k)} D_i^\top \int (-h_n^{-1}) \tilde{I}'\big(-L/h_n\big) f_{\Lambda|Z,D}(L \mid Z_i, D_i)\, dL \Bigg\} \\
&= \mathrm{E}\Bigg\{ Z_i^{(k)} D_i^\top \int_{-1}^{1} -\tilde{I}'(v) f_{\Lambda|Z,D}(-h_n v \mid Z_i, D_i)\, dv \Bigg\} \\
&= -\mathrm{E}\Bigg\{ Z_i^{(k)} D_i^\top \int_{-1}^{1} \tilde{I}'(v) \big[ f_{\Lambda|Z,D}(0 \mid Z_i, D_i) \\
&\qquad\qquad - f'_{\Lambda|Z,D}(-h_n v \mid Z_i, D_i) h_n v + \cdots \big]\, dv \Bigg\} \\
&= -\mathrm{E}\Big[ Z_i^{(k)} D_i^\top f_{\Lambda|Z,D}(0 \mid Z_i, D_i) \Big] \\
&\quad - h_n^r \, \mathrm{E}\Bigg\{ Z_i^{(k)} D_i^\top \int_{-1}^{1} \tilde{I}'(v) (v^r/r!) f^{(r)}_{\Lambda|Z,D}(-\tilde{h} v \mid Z_i, D_i)\, dv \Bigg\} \\
&= -\mathrm{E}\Big[ Z_i^{(k)} D_i^\top f_{\Lambda|Z,D}(0 \mid Z_i, D_i) \Big] + O(h_n^r).
\end{aligned}
$$

Finally, note that Kato (2012) does not require the conditional quantile regression model to be true, i.e., his results hold under misspecification; see his remark on p. 263 and Assumptions 8–13.

# C Computational details

## C.1 Simulation DGP details

**DGP 1** With iid sampling, there is a randomized treatment offer (instrument) $Z_i = 1$ with probability 1/2 and $Z_i = 0$ otherwise; $U_i \sim \mathrm{Unif}(0,1)$ (the unobservable); $D_i = 1$ if $i$ is treated and $D_i = 0$ otherwise, with $\mathrm{P}(D_i = 1 \mid Z_i, U_i) = Z_i \min\{1, (4/3)U_i\}$ so that there is imperfect compliance and endogenous self-selection into treatment; and $Y_i = \beta(U_i) + D_i\gamma(U_i)$, where the function $\beta(\tau) = 60 + Q(\tau)$ with $Q(\cdot)$ the quantile function of the $\chi^2_3$ distribution, and $\gamma(\tau) = 100(\tau - 0.5)$, so there is heterogeneity of the quantile treatment effects.

**DGP 2** This DGP is a stationary time series regression with measurement error. The latent explanatory variable is $Z_t$, with $Z_0 \sim \mathrm{N}(0,1)$, $Z_t = \rho_Z Z_{t-1} + \sqrt{1-\rho_Z^2}\nu_t$, and $\nu_t \overset{iid}{\sim} \mathrm{N}(0,1)$, so $\mathrm{Var}(Z_t) = 1$ for all $t$; we use $\rho_Z = 0.5$. The measurement error is $\eta_t \overset{iid}{\sim} \mathrm{N}(0,1)$, and $X_t = Z_t + \eta_t$ is the observed (mismeasured) explanatory variable. Since the $\eta_t$ are independent, the lagged $X_{t-1}$ provides a valid instrument. The outcome is $Y_t = \gamma Z_t + \epsilon_t$ with $\gamma = 1$ and $\epsilon_t$ unobserved. Finally, $\epsilon_t = \rho_\epsilon \epsilon_{t-1} + \sqrt{1-\rho_\epsilon^2} V_t$, $\epsilon_0 \sim \mathrm{N}(0,1)$, $V_t \overset{iid}{\sim} \mathrm{N}(0,1)$, so the marginal distribution is $\epsilon_t \sim \mathrm{N}(0,1)$ for all $t$, and the series $\{\epsilon_t\}$, $\{\eta_t\}$, and $\{\nu_t\}$ are mutually

independent. Letting $\beta(\tau)$ be the $\tau$-quantile of $\epsilon_t - \gamma\eta_t$, since $Y_t = \gamma Z_t + \epsilon_t = \gamma X_t + \epsilon_t - \gamma\eta_t$, we have the quantile restrictions $\mathrm{P}(Y_t - \gamma X_t - \beta(\tau) \leq 0) = \mathrm{P}(\epsilon_t - \gamma\eta_t \leq \beta(\tau)) = \tau$, and $\mathrm{P}(\epsilon_t - \gamma\eta_t \leq \beta(\tau) \mid X_{t-1}) = \mathrm{P}(\epsilon_t - \gamma\eta_t \leq \beta(\tau)) = \tau$ since $X_{t-1} \perp\!\!\!\perp \eta_t, \epsilon_t$, so the IVQR intercept and slope parameters are $\beta(\tau)$ and $\gamma = 1$, respectively.

**DGP 3** This DGP is identical to DGP 2 except that $\epsilon_t = \rho_\epsilon \epsilon_{t-1} + (1-\rho_\epsilon)V_t$, $\epsilon_0 \sim$ Cauchy, $V_t \overset{iid}{\sim}$ Cauchy, so the marginal distribution is $\epsilon_t \sim$ Cauchy for all $t$.

**DGP 4** This DGP is for a log-linearized quantile Euler equation with overidentification, similar to the empirical application in Section 6. The discount factor is $\beta_\tau = 0.99$, the the EIS is $1/\gamma_\tau = 0.2$. There are four excluded instruments, $Z_{j,t} = \rho_j Z_{j,t-1} + \eta_{j,t}$ for $j = 1, 2, 3, 4$, with $\rho_j = 0.2j$. The vector $(\eta_{1,t}, \eta_{2,t}, \eta_{3,t}, \eta_{4,t})$ is multivariate normal with unit variances and $\mathrm{Corr}(\eta_{j,t}, \eta_{k,t}) = 0.1$ for $j \neq k$. Then, $X_t = \delta_0 + \delta_1 Z_{1,t} + V_t(\delta_2 Z_{2,t} + \delta_3 Z_{3,t} + \delta_4 Z_{4,t}) + V_t$, and $Y_t = \ln(\beta_\tau)/\gamma_\tau + X_t/\gamma_\tau + U_t$. Error term vector $(U_t, V_t)$ is bivariate normal with standard deviations $\sigma_U = 2$ and $\sigma_V = 2$, and covariance $\sigma_{UV} = 0.8$. Linking to Section 6, $Y_t$ represents the log consumption ratio, and $X_t$ represents the log real interest rate.

## C.2 Bandwidth selection

For consistency, only $h_n \to 0$ is required, so just picking the smallest possible bandwidth seems reasonable and performs well in simulations. In our experience, the "optimal" bandwidth given in Kaplan and Sun (2017) for the linear iid setting is usually much larger than the smallest numerically feasible value. For example, this is suggested by the figures in Section 7.3 of Kaplan and Sun (2017), where the plug-in optimal bandwidth is clearly well above the smallest fixed bandwidth they used. The same is true in their empirical example (Section 6), and it is again true in our empirical example, where the optimal bandwidth is 10–50 times larger than the bandwidth of 0.0001 that we use.

For the estimator to attain asymptotic normality, with dependent data, it seems $nh_n^2 \to \infty$ is required for Assumption A11, which implies $h_n$ must be larger than $n^{-1/2}$. The smallest possible bandwidth is too small. Additionally, in the linear IVQR case with iid sampling, Kaplan and Sun (2017) find the (approximate) MSE-optimal bandwidth rate to be $n^{-1/(2r-1)}$, where $r$ is the order of the kernel function $\tilde{I}'(\cdot)$, like $r = 4$ for the function used in the code. To try to obtain the rate $n^{-1/7}$, we first find the smallest possible bandwidth $h_0$ and the corresponding number of smoothed observations $n_0$, i.e., for how many $i$ does $-h \leq \Lambda(Y_i, X_i, \hat{\beta}) \leq h$. Then, we take $h = h_0(n^{6/7}/n_0)$. This is ad hoc, but seems to perform

reasonably.

Alternatively, one could experiment with a variation of the AMSE-optimal bandwidth (for estimating $G$) proposed in Kato (2012). With a linear model, this is easily done by replacing one of the $X$ in the rule-of-thumb formula in Kato (2012) with our $Z$, leaving the other $X$ as the regressor vector that in our notation includes the endogenous regressors in $Y$ and exogenous regressors in $X$. This is what we used for our simulations, where it seemed to work better than the more ad hoc procedure above.

# References


Amemiya, T. (1982): "Two Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689–711.

Andrews, D. W. K. (1987): "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," *Econometrica*, 55, 1465–1471.

——— (1988): "Laws of Large Numbers for Dependent Non-identically Distributed Random Variables," *Econometric Theory*, 4, 458–467.

——— (1991a): "An Empirical Process Central Limit Theorem for Dependent Non-identically Distributed Random Variables," *Journal of Multivariate Analysis*, 38, 187–203.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

——— (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

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.

Borchers, H. W. (2015): *pracma: Practical Numerical Math Functions*, R package version 1.8.3.

Bruins, M., J. A. Duffy, M. P. Keane, and A. A. Smith, Jr. (2015): "Generalized Indirect Inference for Discrete Choice Models," Working paper, available at `https://arxiv.org/abs/1507.06115`.

Buchinsky, M. (1998): "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *Journal of Human Resources*, 33, 88–126.

Campbell, J. Y. (2003): "Consumption-Based Asset Pricing," in *Handbook of the Economics of Finance: Financial Markets and Asset Pricing*, ed. by G. M. Constantinides, M. Harris, and R. M. Stulz, North Holland, vol. 1, Part B, chap. 13, 803–887.

Campbell, J. Y. and N. G. Mankiw (1989): "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence," in *NBER Macroeconomics Annual 1989*, ed. by O. J. Blanchard and S. Fischer, Cambridge, MA: MIT Press, 185–216.

Campbell, J. Y. and L. M. Viceira (1999): "Consumption and Portfolio Decisions When Expected Returns Are Time Varying," *Quarterly Journal of Economics*, 114, 433–495.

Chambers, C. P. (2009): "An Axiomatization of Quantiles on the Domain of Distribution Functions," *Mathematical Finance*, 19, 335–342.

Chen, L.-Y. and S. Lee (2017): "Exact Computation of GMM Estimators for Instrumental Variable Quantile Regression Models," Working paper, available at `https://arxiv.org/abs/1703.09382`.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): "Local Identification of Nonparametric and Semiparametric Models," *Econometrica*, 82, 785–809.

Chen, X. and Z. Liao (2015): "Sieve Semiparametric Two-Step GMM under Weak Dependence," *Journal of Econometrics*, 189, 163–186.

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608.

Chen, X. and D. Pouzo (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152, 46–60.

——— (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments," *Econometrica*, 80, 277–322.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132, 491–525.

——— (2008): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398.

Chernozhukov, V., C. Hansen, and K. Wüthrich (2017): "Instrumental Variable Quantile Regression," in *Handbook of Quantile Regression*, ed. by R. Koenker, V. Chernozhukov, X. He, and L. Peng, CRC/Chapman-Hall, forthcoming.

Chernozhukov, V. and H. Hong (2003): "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346.

Cochrane, J. H. (2005): *Asset Pricing*, Princeton, NJ: Princeton University Press.

de Castro, L. and A. F. Galvao (2017): "Dynamic Quantile Models of Rational Behavior," University of Iowa, mimeo.

de Jong, R. M. (1998): "Weak Laws of Large Numbers for Dependent Random Variables," *Annales d'Économie et de Statistique*, 209–225.

de Jong, R. M. and J. Davidson (2000): "Consistency of Kernel Estimators of Heteroscedastic and Autocorrelated Covariance Matrices," *Econometrica*, 68, 407–423.

Fernandes, M., E. Guerre, and E. Horta (2017): "Smoothing Quantile Regressions," Mimeo, available at `http://bibliotecadigital.fgv.br/dspace/handle/10438/18390`.

Galvao, A. F. and K. Kato (2016): "Smoothed Quantile Regression for Panel Data," *Journal of Econometrics*, 193, 92–112.

Giovannetti, B. C. (2013): "Asset Pricing under Quantile Utility Maximization," *Review of Financial Economics*, 22, 169–179.

Hall, R. E. (1988): "Intertemporal Substitution in Consumption," *Journal of Political Economy*, 96, 339–357.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P. and K. J. Singleton (1983): "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, 92, 249–265.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

——— (1998): "Bootstrap Methods for Median Regression Models," *Econometrica*, 66, 1327–1351.

Hwang, J. and Y. Sun (2015): "Should We Go One Step Further? An Accurate Comparison of One-step and Two-step Procedures in a Generalized Method of Moments Framework," Working paper, available at `https://hwang.econ.uconn.edu/research`.

Kaplan, D. M. and Y. Sun (2017): "Smoothed Estimating Equations for Instrumental Variables Quantile Regression," *Econometric Theory*, 33, 105–157.

Kato, K. (2012): "Asymptotic Normality of Powell's Kernel Estimator," *Annals of the Institute of Statistical Mathematics*, 64, 255–273.

Kinal, T. W. (1980): "The Existence of Moments of k-Class Estimators," *Econometrica*, 48, 241–249.

Koenker, R. and G. Bassett, Jr. (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Lancaster, T. and S. J. Jun (2010): "Bayesian quantile regression methods," *Journal of Applied Econometrics*, 25, 287–307.

Ljungqvist, L. and T. J. Sargent (2012): *Recursive Macroeconomic Theory*, Cambridge, Massachusetts: MIT Press, 3rd ed.

Lucas, R. E. (1978): "Asset Prices in an Exchange Economy," *Econometrica*, 46, 1429–1446.

MaCurdy, T. (2007): "A Practitioner's Approach to Estimating Intertemporal Relationships Using Longitudinal Data: Lessons from Applications in Wage Dynamics," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6A, chap. 62, 4057–4167.

MaCurdy, T. and H. Hong (1999): "Smoothed Quantile Regression in Generalized Method of Moments," Mimeo.

MaCurdy, T. and C. Timmins (2001): "Bounding the Influence of Attrition on Intertemporal Wage Variation in the NLSY," Mimeo, available at `http://public.econ.duke.edu/~timmins/bounds.pdf`.

Manski, C. F. (1988): "Ordinal Utility Models of Decision Making under Uncertainty," *Theory and Decision*, 25, 79–104.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4, chap. 36, 2111–2245.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Oberhofer, W. and H. Haupt (2016): "Asymptotic Theory for Nonlinear Quantile Regression under Weak Dependence," *Econometric Theory*, 32, 686–713.

Ogaki, M. and C. M. Reinhart (1998): "Measuring Intertemporal Substitution: The Role of Durable Goods," *Journal of Political Economy*, 106, 1078–1098.

Otsu, T. (2008): "Conditional Empirical Likelihood Estimation and Inference for Quantile

Regression Models," *Journal of Econometrics*, 142, 508–538.

Pötscher, B. M. and I. R. Prucha (1994): "Generic Uniform Convergence and Equicontinuity Concepts for Random Functions: An Exploration of the Basic Structure," *Journal of Econometrics*, 60, 23–63.

Powell, J. L. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303–325.

——— (1991): "Estimation of Monotonic Regression Models under Quantile Restrictions," in *Nonparametric and semiparametric methods in econometrics and statistics: proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, Cambridge University Press, 357–384.

——— (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4, chap. 41, 2443–2521.

R Core Team (2013): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rostek, M. (2010): "Quantile Maximization in Decision Theory," *Review of Economic Studies*, 77, 339–371.

Schennach, S. M. (2005): "Bayesian exponentially tilted empirical likelihood," *Biometrika*, 92, 31–46.

——— (2007): "Point estimation with exponentially tilted empirical likelihood," *Annals of Statistics*, 35, 634–672.

Su, L. and Z. Yang (2011): "Instrumental Variable Quantile Estimation of Spatial Autoregressive Models," Working paper, available at `http://www.mysmu.edu/faculty/ljsu/Publications/ivqr_sar20110505.pdf`.

Toda, A. A. and K. Walsh (2015): "The Double Power Law in Consumption and Implications for Testing Euler Equations," *Journal of Political Economy*, 123, 1177–1200.

Toda, A. A. and K. J. Walsh (2017): "Fat tails and spurious estimation of consumption-based asset pricing models," *Journal of Applied Econometrics*, 32, 1156–1177.

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

Whang, Y.-J. (2006): "Smoothed Empirical Likelihood Methods for Quantile Regression Models," *Econometric Theory*, 22, 173–205.

Wooldridge, J. M. (1986): "Asymptotic Properties of Econometric Estimators," Ph.D. thesis, Department of Economics, University of California, San Diego.

Wüthrich, K. (2016): "A comparison of two quantile models with endogeneity," Working paper.

——— (2017): "A closed-form estimator for quantile treatment effects with endogeneity," Working paper.

Xiang, Y., S. Gubian, B. Suomela, and J. Hoeng (2013): "Generalized Simulated Annealing for Global Optimization: The GenSA Package." *The R Journal*, 5, 13–28.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak," *Review of Economics and Statistics*, 86, 797–810.