SMOOTHED ESTIMATING EQUATIONS FOR INSTRUMENTAL VARIABLES QUANTILE REGRESSION

DAVID M. KAPLAN AND YIXIAO SUN

Abstract. The moment conditions or estimating equations for instrumental variables quantile regression involves the discontinuous indicator function. We instead use smoothed estimating equations, with bandwidth h. This is known to allow higher-order expansions that justify bootstrap refinements for inference. Computation of the estimator also becomes simpler and more reliable, especially with (more) endogenous regressors. We show that the mean squared error of the vector of estimating equations is minimized for some $h > 0$, which also reduces the mean squared error of the parameter estimators. The same h also minimizes higher-order type I error for a χ^2 test, leading to improved size-adjusted power. Our plug-in bandwidth consistently reproduces all of these properties in simulations.

Keywords: Edgeworth expansion, Instrumental variable, Optimal smoothing parameter choice, Quantile regression, Smoothed estimating equation.

JEL Classification Number: C13, C21.

1. Introduction

Many econometric models are specified by moment conditions or estimating equations. An advantage of this approach is that the full distribution of the data does not have to be parameterized. In this paper, we consider estimating equations that are not smooth in the parameter of interest. We focus on the instrumental variables quantile regression (IV-QR), which includes the usual quantile regression as a special case. Instead of using the estimating equations that involve the nonsmooth indicator function, we propose to smooth the indicator function, leading to our smoothed estimating equations (SEE) and SEE estimator.

Our SEE estimator has several advantages. First, from a computational point of view, the SEE estimator can be computed using any standard iterative algorithm that

---

Date: First: January 27, 2012; this: November 29, 2012.

Department of Economics, University of California, San Diego, La Jolla, CA 92093-0508, USA. Email: {dkaplan,yisun}@ucsd.edu. Thanks to Xiaohong Chen, Brendan Beare, Andres Santos, and other UCSD seminar participants for insightful questions and comments.

requires smoothness. This is especially attractive in IV-QR where simplex methods for the usual QR are not applicable. Second, from a technical point of view, smoothing the estimating equations enables us to establish high-order properties of the estimator. This motivated, for instance, Horowitz (1998) to examine a smoothed objective function for median regression, to show high-order bootstrap refinement. Instead of smoothing the objective function, we show that there is an advantage of smoothing the estimating equations. For QR estimation and inference via empirical likelihood, Otsu (2008) and Whang (2006) also examined smoothed estimators. To the best of our knowledge, nobody has examined smoothing the estimating equations for the usual QR estimator. Third, from a statistical point of view, the SEE estimator is a flexible class of estimators that includes the IV/OLS mean regression estimators and median and quantile regression estimators as special cases. Depending on the smoothing parameter, the SEE estimator can have different degrees of robustness in the sense of Huber (1964). By selecting the smoothing parameter appropriately, we can harness the advantages of both the mean regression estimator and the median/quantile regression estimator. Fourth and most importantly, from an econometric point of view, smoothing can reduce the mean squared error (MSE) of the SEE, which in turn leads to a smaller asymptotic MSE of the parameter estimator and to tests with steeper power curves. This advantage has not been discussed in the literature.

In addition to investigating the asymptotic properties of the SEE estimator, we provide a smoothing parameter choice that minimizes different criteria: the MSE of the SEE, the type I error of a chi-square test subject to exact asymptotic size, and the approximate MSE of the parameter estimator. We show that the first two criteria produce the same optimal smoothing parameter, which is also optimal under a variant of the third criterion. With the data-driven smoothing parameter choice, we show that the statistical and econometric advantages of the SEE estimator are reflected clearly in our simulation results.

The rest of the paper is organized as follows. Section 2 describes our setup and discusses some illuminating connections with other estimators. Sections 3, 4, and 5 calculate the MSE of the SEE, the type I error of a χ^2 test, and the approximate MSE of the parameter estimator, respectively; optimal bandwidth is discussed for each case. Section 6 presents simulation results before we conclude. Longer proofs and calculations are gathered in the appendix.

2. Smoothed Estimating Equations

2.1. **Setup.** We are interested in estimating the instrumental variables quantile regression (IV-QR) model

$$Y_j = X_j'\beta_0 + U_j$$

where $P(U_j < 0 | Z_j) = q$ almost surely for instrument vector $Z_j \in \mathbb{R}^d$. Instruments are taken as given; this does not preclude first determining the efficient set of instruments as in Newey (2004) or Newey and Powell (1990), for example. We restrict attention to the "just identified" case $X_j \in \mathbb{R}^d$ and iid data for simpler exposition; for the overidentified case, see (1) below.

A special case of this model is exogenous QR with $Z_j = X_j$, which is typically estimated by minimizing a criterion function:

$$\hat{\beta}_Q \equiv \arg\min_{\beta} \frac{1}{n} \sum_{j=1}^{n} \rho_q(Y_j - X_j'\beta),$$

where $\rho_q(u) \equiv [q - 1(u < 0)]u$ is the check function. Since the objective function is not smooth, it is not easy to obtain a high-order approximation to the sampling distribution of $\hat{\beta}_Q$. To avoid this technical difficulty, Horowitz (1998) proposes to smooth the objective function to obtain

$$\hat{\beta}_H = \arg\min_{\beta} \frac{1}{n} \sum_{j=1}^{n} \rho_q^H(Y_j - X_j'\beta), \quad \rho_q^H(u) \equiv [q - G(-u/h)]u,$$

where $G(\cdot)$ is a smooth function and h is the smoothing parameter or bandwidth. Instead of smoothing the objective function, we smooth the underlying moment condition[1] and define $\hat{\beta}$ to be the solution of the vector of smoothed estimating equations (SEE) $m_n(\hat{\beta}) = 0$, where[2]

$$m_n(\beta) \equiv \frac{1}{\sqrt{n}} \sum_{j=1}^{n} W_j(\beta) \text{ and } W_j(\beta) \equiv Z_j \left[G\left(\frac{X_j'\beta - Y_j}{h} \right) - q \right].$$

The SEE is analogous to that employed by Whang (2006) with $Z_j = X_j$.

[1] This is only one of an infinite number of unconditional moment conditions implied by the conditional moment condition, albeit the most popular one. Other transformations of the instrument vector Z_i could be used, which may improve efficiency. Alternatively, Otsu (2008) argues for using the conditional version of the moment condition to improve efficiency of a smoothed conditional empirical likelihood estimator.

[2] It suffices to have $m_n(\hat{\beta}) = o_p(1)$, which allows for a small error when $\hat{\beta}$ is not the exact solution to $m_n(\hat{\beta}) = 0$.

Our approach is related to kernel-based nonparametric conditional quantile estimators. The moment condition there is $E[1\{X = x\}(1\{Y < \beta\} - q)] = 0$. Usually the $1\{X = x\}$ indicator function is "smoothed" with a kernel, while the latter term is not. This yields the nonparametric conditional quantile estimator $\hat{\beta}_q(x) = \arg\min_b \sum_{i=1}^n \rho_q(Y_i - b)K[(x - X_i)/h]$ for the conditional q-quantile at $X = x$, estimated with kernel $K(\cdot)$ and bandwidth h. Our approach is different in that we smooth the indicator $1\{Y < \beta\}$ rather than $1\{X = x\}$. Smoothing both terms may help but is beyond the scope of this paper.

Estimating $\hat{\beta}$ from the SEE is computationally easy: d equations for d parameters, and a known, analytic Jacobian. Even when the model is overidentified with $\dim(Z_j) > \dim(X_j)$, we can transform the original moment conditions $E\{Z_j[q - 1\{Y_j < X_j'\beta\}]\} = 0$ into

$$E\Big(\tilde{Z}_j\big[q - 1\{Y_j < X_j'\beta\}\big]\Big) = 0, \quad \tilde{Z}_j = X_j Z_j'(Z_j Z_j')^{-1} Z_j \in \mathbb{R}^{\dim(X_j)}. \tag{1}$$

Then we have an exactly identified model with transformed instrument vector $\tilde{Z}_j$, and our asymptotic analysis can be applied to (1). Computationally, solving our problem is faster and more reliable than the IV-QR method in Chernozhukov and Hansen (2006), which requires specification of a grid of endogenous coefficient values to search over, computing a conventional QR estimator for each grid point. This advantage is important particularly when there are more endogenous variables.

2.2. **Comparison with other estimators.**

Smoothed criterion function. For the special case $Z_j = X_j$, we compare the SEE with that derived from smoothing the criterion function as in Horowitz (1998). The first order condition of the smoothed criterion function, evaluated at the true β_0, is

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\bigg|_{\beta=\beta_0} n^{-1}\sum_{i=1}^n \left[q - G\left(\frac{X_i'\beta - Y_i}{h}\right)\right](Y_i - X_i'\beta) \\
&= n^{-1}\sum_{i=1}^n \Big[- qX_i - G'(-U_i/h)(X_i/h)Y_i + G'(-U_i/h)(X_i/h)X_i'\beta_0 + G(-U_i/h)X_i\Big] \\
&= n^{-1}\sum_{i=1}^n X_i[G(-U_i/h) - q] + n^{-1}\sum_{i=1}^n G'(-U_i/h)[(X_i/h)X_i'\beta_0 - (X_i/h)Y_i] \\
&= n^{-1}\sum_{i=1}^n X_i[G(-U_i/h) - q] + n^{-1}\sum_{i=1}^n (1/h)G'(-U_i/h)[-X_iU_i].
\end{aligned}$$

Technically, it should be easier to establish high-order results for our SEE estimator since it has fewer terms than this. Later we show that the absolute bias of our SEE estimator is smaller, too.

IV mean regression. When $h \to \infty$, $G(\cdot)$ only takes arguments near zero and thus can be approximated well linearly. For example, with the $G(\cdot)$ from Whang (2006) and Horowitz (1998), $G(v) = 0.5 + (105/64)v + O(v^3)$ as $v \to 0$. Ignoring the $O(v^3)$, the corresponding estimator is defined as

$$\begin{aligned}
0 &= \sum_{i=1}^{n} Z_i \left[G\left(\frac{X_i'b - Y_i}{h} \right) - q \right] \\
&\doteq \sum_{i=1}^{n} Z_i \left[\left(0.5 + (105/64) \frac{X_i'b - Y_i}{h} \right) - q \right] \\
&= (105b/64h) \sum Z_i X_i' - (105/64h) \sum Z_i Y_i + (0.5 - q) \sum Z_i, \\
Z'Xb &= Z'Y + (64h/105)(q - 0.5) Z' \mathbf{1}_{n,1}, \\
\hat{\beta} &= (Z'X)^{-1} Z'Y + (Z'X)^{-1} Z'(Xe_1)(64h/105)(q - 0.5) \\
&= \hat{\beta}_{IV} + ((64h/105)(q - 0.5), 0, \dots, 0)',
\end{aligned}$$

where $e_1 = (1, 0, \dots, 0)'$ is $d \times 1$, $\mathbf{1}_{n,1} = (1, 1, \dots, 1)'$ is $n \times 1$, X and Z are $n \times d$ with respective rows X_i' and Z_i', and using the fact that the first column of X is $\mathbf{1}_{n,1}$ so that $Xe_1 = \mathbf{1}_{n,1}$. As h grows large, the smoothed QR estimator approaches the IV estimator plus an adjustment to the intercept term that depends on q, the bandwidth, and the slope of $G(\cdot)$ at zero. In the special case $Z_j = X_j$, the IV estimator is the OLS estimator.[3]

The intercept is often not of interest, and when $q = 0.5$, the adjustment is zero anyway. The class of SEE estimators is a continuum (indexed by h) with two well-known special cases at the extremes: unsmoothed IV-QR and mean IV. For $q = 0.5$ and $Z_j = X_j$, this is median regression and mean regression (OLS). Well known are the relative efficiency advantages of unsmoothed QR and OLS for different error distributions. Our estimator with a data-driven bandwidth can harness the advantages of both, without requiring the practitioner to guess assumptions about the unknown error distribution.

[3]This is different from Zhou et al. (2011), who add the d OLS moment conditions to the d median regression moment conditions before estimation; our connection to IV/OLS emerges naturally from smoothing the (IV)QR estimating equations.

Robust estimation. With $Z_j = X_j$, the result that our SEE can yield OLS when $h \to \infty$ or median regression when $h = 0$ calls to mind robust estimators like the trimmed or Winsorized mean (and corresponding regression estimators). Setting the trimming/Winsorization parameter to zero generates the mean while the other extreme generates the median. However, our SEE mechanism is different and more general/flexible; trimming/Winsorization is not directly applicable to $q \neq 0.5$; our method to select the smoothing parameter is novel; and the motivations for QR extend beyond (though include) robustness.

With $X_i = 1$ and $q = 0.5$ (population median estimation), our SEE becomes

$$0 = n^{-1} \sum_{i=1}^{n} [2G((\beta - Y_i)/h) - 1].$$

Note that $H(u) \equiv 2G(u) - 1$ takes value 1 for $u \geq 1$ and -1 for $u \leq -1$. Our estimator is then an M-estimator with $\psi(Y_i) = H[(\beta - Y_i)/h]$, defined by $\sum_{i=1}^{n} \psi(Y_i) = 0$. If $H(u)$ is piecewise linear with $H(u) = u$ for $u \in [-1, 1]$, then we have a Winsorized mean estimator of the type in Huber (1964, example (iii) on page 79).[4] In our framework, this is choosing $G'(\cdot)$ to be the uniform kernel ($r = 2$).

Further theoretical comparison of our SEE-QR with trimmed/Winsorized mean regression (and the IV versions) would be interesting but is beyond the scope of this paper. For more on robust location and regression estimators, see for example Huber (1964), Koenker and Bassett (1978), and Ruppert and Carroll (1980).

3. MSE of the SEE

Since statistical inference can be made based on the estimating equations (EEs), we examine the mean squared error (MSE) of the SEE. The MSE of the SEE is related to the estimator MSE and inference properties both intuitively and (as we will show) theoretically. Such a result may provide helpful guidance in contexts where the SEE MSE is easier to compute than the estimator MSE, and it provides insight into how smoothing works in the QR model as well as results that will be used in subsequent sections.

We maintain different subsets of the following assumptions for different results. They are stated together and in this order to parallel Assumptions 1–6 in both Horowitz (1998) and Whang (2006). We write $f_{U|Z}(\cdot|z)$ as the conditional PDF of U given $Z = z$, and similarly $F_{U|Z}(\cdot|z)$ for the conditional CDF.

[4] For a strict mapping, multiply by h to get $\psi(Y_i) = hH[(\beta - Y_i)/h]$. The solution is equivalent since $\sum h\psi(Y_i) = 0$ is the same as $\sum \psi(Y_i) = 0$ for any nonzero constant h.

Assumption 1. (X_j', Z_j', Y_j) is iid across $j = 1, 2, \ldots, n$, where $Y_j = X_j'\beta_0 + U_j$, X_j is an observed $d \times 1$ vector of stochastic regressors that can include a constant, β_0 is an unknown $d \times 1$ constant vector, U_j is an unobserved random scalar, and Z_j is an observed $d \times 1$ vector of instruments (see assumptions below).

Assumption 2. $\beta = \beta_0$ uniquely solves $E\big(Z_j\big[q - 1\{Y_j < X_j'\beta\}\big]\big) = 0$ over $\beta \in \mathcal{B}$.

Assumption 3. Z_j has bounded support and EZ_jZ_j' is nonsingular.

Assumption 4. (i) $P(U_j < 0 | Z_j = z) = q$ for almost all $z \in \mathcal{Z}$, the support of Z. (ii) For all u in a neighborhood of zero and almost all $z \in \mathcal{Z}$, $f_{U|Z}(u|z)$ exists, is bounded away from zero, and is r times continuously differentiable with $r \geq 2$. (iii) There exists a function $C(z)$ such that $\left|f_{U|Z}^{(s)}(u|z)\right| \leq C(z)$ for $s = 0, 2, \ldots, r$, almost all $z \in \mathcal{Z}$ and u in a neighborhood of zero, and $E\big[C(Z)\|Z\|^2\big] < \infty$.

Assumption 5. (i) $G(v)$ is a bounded function satisfying $G(v) = 0$ for $v \leq -1$ and $G(v) = 1$ for $v \geq 1$. (ii) $G'(\cdot)$ is a symmetric and bounded rth order kernel with $r \geq 2$ so that $\int_{-1}^{1} G'(v)dv = 1$, $\int_{-1}^{1} v^k G'(v)dv = 0$ for $k = 1, 2, \ldots, r-1$, $\int_{-1}^{1} |v^r G'(v)|dv < \infty$, and $\int_{-1}^{1} v^r G'(v)dv \neq 0$. (iii) Let $\tilde{G}(u) = \big(G(u), [G(u)]^2, \ldots, [G(u)]^{L+1}\big)'$ for some $L \geq 1$. For any $\theta \in \mathbb{R}^{L+1}$ satisfying $\|\theta\| = 1$, there is a partition of $[-1, 1]$ given by $-1 = a_0 < a_1 < \cdots < a_{L+1} = 1$ such that $\theta'\tilde{G}(u)$ is either strictly positive or strictly negative on the intervals (a_{i-1}, a_i) for $i = 1, 2, \ldots, L+1$.

Assumption 6. $h \propto n^{-\kappa}$ for $1/(2r) < \kappa < 1$ where $r \geq 2$.

Assumption 1 describes the sampling process. Assumption 2 ensures that β_0 is identified. See Theorem 2 of Chernozhukov and Hansen (2006) for more primitive conditions. Assumption 3 is analogous to Assumption 3 in both Horowitz (1998) and Whang (2006). As discussed in these two papers, the boundedness assumption for Z_j is made only for convenience and can be dropped at the cost of more complicated proofs.

Assumption 4(ii) is critical. If we are not willing to make such an assumption, then smoothing will be of no benefit. Inversely, with some small degree of smoothness of the conditional error density, smoothing can leverage this into the advantages described here. Also note that Horowitz (1998) assumes $r \geq 4$, which is sufficient for the estimator MSE result in §5.

Assumptions 5 and 6 are needed for the Edgeworth expansion. As Horowitz (1998) and Whang (2006) discuss, Assumption 5(iii) is a technical assumption that (along with Assumption 6) leads to a form of Cramér's condition, which is needed to justify the Edgeworth expansion used in §4.

Define

$$W_j \equiv W_j(\beta_0) = Z_j[G(-U_j/h) - q]$$

and abbreviate $m_n \equiv m_n(\beta_0) = n^{-1/2}\sum_{j=1}^n W_j$. The theorem below gives the first two moments of W_j and the first-order asymptotic distribution of m_n.

Theorem 1. *Let Assumptions 3, 4, and 5(i–ii) hold. Then*

$$E(W_j) = \frac{(-h)^r}{r!}\left(\int_{-1}^{1} G'(v)v^r dv\right) E\left[f_{U|Z}^{(r-1)}(0|Z_j)Z_j\right] + o(h^r), \tag{2}$$

$$E(W_j'W_j) = q(1-q)E\{Z_j'Z_j\} - h\left[1 - \int_{-1}^{1} G^2(u)du\right] E\{f_{U|Z}(0|Z_j)Z_j'Z_j\} + O(h^2), \tag{3}$$

$$E(W_jW_j') = q(1-q)E\{Z_jZ_j'\} - h\left[1 - \int_{-1}^{1} G^2(u)du\right] E\{f_{U|Z}(0|Z_j)Z_jZ_j'\} + O(h^2).$$

If additionally Assumptions 1 and 6 hold, then

$$m_n \xrightarrow{d} N(0,V), \quad V \equiv \lim_{n\to\infty} E\{(W_j - EW_j)(W_j - EW_j)'\} = q(1-q)E(Z_jZ_j').$$

Theorem 1 shows that $E(W_j) = O(h^r)$. This bias is smaller than that of the SEE derived from smoothing the criterion function as in Horowitz (1998). The bias of the EE derived from the smoothed criterion function is

$$\begin{aligned} &E\{[G(-U_j/h) - q]Z_j\} - \frac{1}{h}E\{U_jG'(-U_j/h)Z_j\} \\ &\quad = (-h)^r\left(\frac{1}{r!} + \frac{1}{(r-1)!}\right)\left(\int G'(v)v^r dv\right) E\left\{f_{U|Z}^{(r-1)}(0|Z_j)Z_j\right\} + o(h^r), \end{aligned}$$

as calculated in the appendix. The dominating term of the bias of our SEE is smaller in absolute value than that of the EE derived from a smoothed criterion function. A larger bias can lead to less accurate confidence regions if the same variance estimator is used.

The first-order asymptotic variance V is exactly the same as the asymptotic variance of $n^{-1/2}\sum_{j=1}^n Z_j[1(U_j<0) - q]$, the scaled EE of the unsmoothed IV-QR. The effect of smoothing on the variance is captured by the term of order h (up to a smaller-order remainder). To determine the sign of this term, we need to sign

$$\begin{aligned} 1 - \int_{-1}^{1} G^2(u)du &= 2\int_{-1}^{1} uG(u)G'(u)du \\ &= 2\int_{0}^{1} uG(u)G'(u)du + 2\int_{-1}^{0} uG(u)G'(u)du \end{aligned}$$

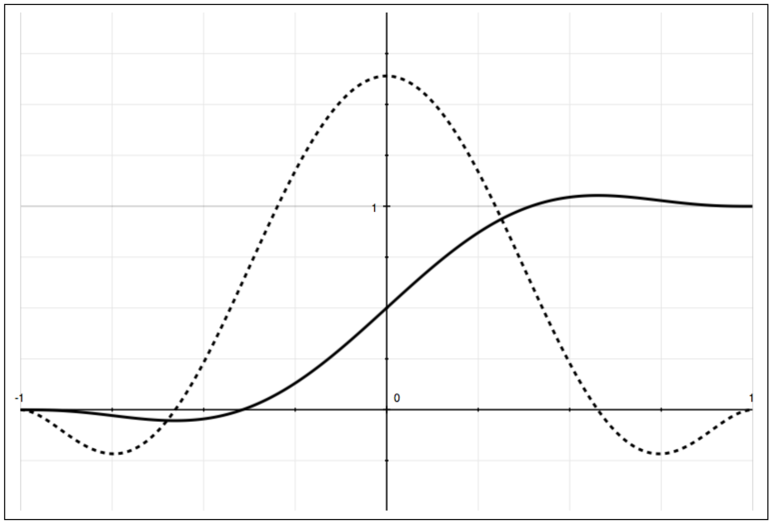

FIGURE 1. Graph of $G(u) = 0.5 + \frac{105}{64}\left(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\right)$ (solid line) and its derivative (broken).

$$= 2\int_0^1 uG(u)G'(u)du - 2\int_0^1 vG(-v)G'(-v)dv$$
$$= 2\int_0^1 uG'(u)[G(u) - G(-u)]du,$$

using the evenness of $G'(u)$.

When $r = 2$, we can use $G(u)$ such that $G'(u) > 0$ and $G(u) > G(-u)$. That is, we can take $G'(u)$ to be any symmetric PDF on $[-1, 1]$. In this case, $1 - \int_{-1}^1 G^2(u)du > 0$, so smoothing reduces the variance by order $O(h)$. This is not surprising. Replacing the discontinuous indicator function $1\{U < 0\}$ by a smooth function $G(-U/h)$ pushes the dichotomous values of zero and one into some values in between, leading to a smaller variance. The idea is similar to Breiman's (1994) bagging (bootstrap aggregating), among others.

When $r > 2$, $G'(u) < 0$ for some u, and $G(u)$ is not monotonic. It is not easy to sign $1 - \int_{-1}^1 G^2(u)du$ generally, but it is simple to calculate for any chosen $G(\cdot)$. For example, consider $r = 4$ and the $G(\cdot)$ function in Horowitz (1998) and Whang (2006) shown in Figure 1,

$$G(u) = 0.5 + \frac{105}{64}\left(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\right) \text{ for } u \in [-1, 1].$$

The range of the function falls outside $[0,1]$. Some calculations show that $1-\int_{-1}^{1} G^2(u)du > 0$. Like the case of $r=2$, smoothing reduces the variance while introducing some bias into the EE. We will assume that $1-\int_{-1}^{1} G^2(u)du > 0$ throughout the rest of the paper.

Define the MSE of the SEE to be $E\{m_n' V^{-1} m_n\}$. Building off of (2) and (3), and using $W_i \perp\!\!\!\perp W_j$ for $i \neq j$, we have:

$$\begin{aligned}
&E\{m_n' V^{-1} m_n\} \\
&= \frac{1}{n}\sum_{j=1}^{n} E\{W_j' V^{-1} W_j\} + \frac{1}{n}\sum_{j=1}^{n}\sum_{i\neq j} E\big(W_i' V^{-1} W_j\big) \\
&= \frac{1}{n}\sum_{j=1}^{n} E\{W_j' V^{-1} W_j\} + \frac{1}{n} n(n-1)(EW_j')V^{-1}(EW_j) \\
&= q(1-q)E\{Z_j' V^{-1} Z_j\} + nh^{2r}(EB)'(EB) - h\mathrm{tr}\{E(AA')\} + o(h+nh^{2r}), \qquad (4)
\end{aligned}$$

where

$$\begin{aligned}
A &\equiv \left(1-\int_{-1}^{1} G^2(u)du\right)^{1/2} \big[f_{U|Z}(0|Z_j)\big]^{1/2} V^{-1/2} Z_j, \\
B &\equiv \left(\frac{1}{r!}\int_{-1}^{1} G'(v)v^r dv\right) f_{U|Z}^{(r-1)}(0|Z_j) V^{-1/2} Z_j.
\end{aligned}$$

Ignoring the $o(\cdot)$ term, we obtain the asymptotic MSE of the SEE. We select the smoothing parameter to minimize the asymptotic MSE, leading to

$$h_{\mathrm{SEE}}^* \equiv \arg\min_h nh^{2r}(EB)'(EB) - h\mathrm{tr}\{E(AA')\}. \qquad (5)$$

The proposition below gives the optimal smoothing parameter h_{SEE}^*.

Proposition 2. *Let Assumptions 1, 3, 4, and 5(i–ii) hold. The bandwidth that minimizes the asymptotic MSE of the SEE is*

$$h_{SEE}^* = \left(\frac{\mathrm{tr}\{E(AA')\}}{(EB)'(EB)}\frac{1}{2nr}\right)^{\frac{1}{2r-1}}.$$

Under the stronger assumption $U \perp\!\!\!\perp Z$, we have

$$h_{SEE}^* = \left(\frac{(r!)^2\Big[1-\int_{-1}^{1} G^2(u)du\Big] f_U(0)}{2r\Big(\int_{-1}^{1} G'(v)v^r dv\Big)^2 \Big[f_U^{(r-1)}(0)\Big]^2}\frac{d}{n}\right)^{\frac{1}{2r-1}}.$$

When $r=2$, the MSE optimal $h_{\mathrm{SEE}}^* \asymp n^{-1/(2r-1)} = n^{-1/3}$. This is smaller than $n^{-1/5}$, the rate that minimizes the MSE of estimated standard errors of the usual regression quantiles. Since nonparametric estimators of $f_U^{(r-1)}(0)$ converge slowly, we propose a parametric plug-in described in §6.

4. Type I Error of a Chi-square Test

In this section, we explore the effect of smoothing on a χ^2 test. Other alternatives for inference exist, such as the Bernoulli-based MCMC-computed method from Chernozhukov et al. (2009), empirical likelihood as in Whang (2006), and bootstrap as in Horowitz (1998), where the latter two also use smoothing. Intuitively, when we minimize the MSE, we may expect steeper power curves: it is easier to distinguish the null hypothesis from some given alternative. We would also expect lower type I error: the χ^2 critical value is from the unsmoothed distribution, and smoothing to minimize MSE makes large values (that cause the test to reject) less likely. This combination leads to improved size-adjusted power. As seen in our simulations, this is true especially for the IV case. Here, we derive the bandwidth that minimizes the type I error from the first two high-order terms while maintaining exact asymptotic size.

Using the results in §3 and under Assumption 6, we have

$$m_n' V^{-1} m_n \xrightarrow{d} \chi_d^2,$$

where we continue the notation $m_n \equiv m_n(\beta_0)$. From this asymptotic result, we can construct a hypothesis test that rejects the null hypothesis $H_0 : \beta = \beta_0$ when

$$m_n' \hat{V}^{-1} m_n > c_\alpha,$$

where

$$\hat{V} = q(1-q)\frac{1}{n}\sum_{j=1}^{n} Z_j Z_j'$$

is a consistent estimator of V and $c_\alpha \equiv \chi^2_{d,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with d degrees of freedom. As desired, the asymptotic size is

$$\lim_{n\to\infty} P\Big(m_n' \hat{V}^{-1} m_n > c_\alpha\Big) = \alpha.$$

To more precisely measure the type I error $P\Big(m_n' \hat{V}^{-1} m_n > c_\alpha\Big)$, we develop a high-order expansion of the statistic $S_n \equiv m_n' \hat{V}^{-1} m_n$. Let $V_n \equiv \mathrm{Var}(m_n)$. Following the same calculation as in (4), we have

$$\begin{aligned} V_n &= V - h\left[1 - \int_{-1}^{1} G^2(u)du\right] E[f_{U|Z}(0|Z_j) Z_j Z_j'] + O(h^2) \\ &= V^{1/2}\big[I_d - hE(AA') + O(h^2)\big]\big(V^{1/2}\big)', \end{aligned}$$

where $V^{1/2}$ is the matrix square root of V such that $V^{1/2}\big(V^{1/2}\big)' = V$. We can choose $V^{1/2}$ to be symmetric but do not have to.

Details of the following are in the appendix; here we outline our strategy and highlight key results. Letting

$$\Lambda_n = V^{1/2}\big[I_d - hE(AA') + O(h^2)\big]^{1/2} \tag{6}$$

such that $\Lambda_n\Lambda_n' = V_n$, and defining

$$\bar{W}_n^* \equiv \frac{1}{n}\sum_{j=1}^{n} W_j^* \text{ and } W_j^* = \Lambda_n^{-1} Z_j[G(-U_j/h) - q], \tag{7}$$

we can approximate the test statistic (as shown in the appendix) as

$$S_n = S_n^L + e_n$$

where

$$S_n^L = \big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big) - h\big(\sqrt{n}\bar{W}_n^*\big)'E(AA')\big(\sqrt{n}\bar{W}_n^*\big),$$

and e_n is the remainder term satisfying $P(|e_n| > O(h^2)) = O(h^2)$.

The stochastic expansion above allows us to approximate the characteristic function of S_n with that of S_n^L, which we calculate in the appendix. Taking the Fourier-Stieltjes inverse of the characteristic function yields an approximation of the distribution function, from which we can calculate the type I error by plugging in the critical value c_α.

Theorem 3. *Under Assumptions 1–6, we have*

$$P(S_n^L < x) = \mathcal{G}_d(x) - \frac{1}{d}\mathcal{G}_d'(x)x\Big\{\big\|\sqrt{n}EW_j^*\big\|^2 - h\mathrm{tr}\{E(AA')\}\Big\} + R_n,$$

$$P(S_n > c_\alpha) = \alpha + \frac{1}{d}\mathcal{G}_d'(c_\alpha)c_\alpha\Big\{\big\|\sqrt{n}EW_j^*\big\|^2 - h\mathrm{tr}\{E(AA')\}\Big\} + R_n,$$

where $R_n = O(h^2 + nh^{2r+1})$ *and* $\mathcal{G}_d(x)$ *is the CDF of the* χ_d^2 *distribution.*

It follows from Theorem 1 that

$$\big\|\sqrt{n}EW_j^*\big\|^2 = nh^{2r}(EB)'(EB)(1 + o(1)).$$

So according to Theorem 3, an approximate measure of the type I error of the chi-square test is

$$\alpha + \frac{1}{d}\mathcal{G}_d'(c_\alpha)c_\alpha\big[nh^{2r}(EB)'(EB) - h\mathrm{tr}\{E(AA')\}\big],$$

and an approximate measure of the coverage probability error (CPE) is[5]

$$\text{CPE} = \frac{1}{d}\mathcal{G}_d'(c_\alpha)c_\alpha\big[nh^{2r}(EB)'(EB) - h\text{tr}\{E(AA')\}\big],$$

which is also the error in rejection probability under the null.

Up to smaller-order terms, the term $nh^{2r}(EB)'(EB)$ characterizes the bias effect from smoothing. The (squared) bias inflation increases the type I error and reduces coverage probability. The term $h\text{tr}\{E(AA')\}$ characterizes the variance effect from smoothing. The variance reduction decreases the type I error and increases the coverage probability. When $h \to 0$, the type I error is α up to order $O(h + nh^{2r})$. For $h = 0$ as well as some $h > 0$ that makes bias and variance effects cancel, the type I error is α up to smaller-order terms in R_n.

Note that $nh^{2r}(EB)'(EB) - h\text{tr}\{E(AA')\}$ is exactly the same as the high-order term in the asymptotic MSE of the SEE as given in (4). The h^*_{CPE} that minimizes the type I error (maximizes coverage probability) is the same as h^*_{SEE}.

Proposition 4. *Let Assumptions 1–6 hold. The bandwidth that minimizes the approximate type I error of the chi-square test based on the test statistic S_n is*

$$h^*_{CPE} = h^*_{SEE} = \left(\frac{\text{tr}\{E(AA')\}}{(EB)'(EB)}\frac{1}{2nr}\right)^{\frac{1}{2r-1}}.$$

The result that $h^*_{\text{CPE}} = h^*_{\text{SEE}}$ is intuitive. Since h^*_{SEE} minimizes $E[m_n'V^{-1}m_n]$, for a test with c_α and $\hat{V}$ both invariant to h, the null rejection probability $P(m_n'\hat{V}^{-1}m_n > c_\alpha)$ should be smaller when the SEE's MSE is smaller.

It is important to point out the optimal smoothing parameter h^*_{CPE} or h^*_{SEE} is invariant to rotation and translation of the (non-constant) regressors. This may not be obvious but can be proved easily.

5. MSE of the Parameter Estimator

In this section, we examine the approximate MSE of the parameter estimator. The approximate MSE, being a Nagar-type approximation (Nagar, 1959), can be motivated from the theory of optimal estimating equations, as presented in Heyde (1997), for example.

---

[5]The CPE is defined to be nominal coverage probability minus the true coverage probability, which may be different from the usual definition. Under this definition, smaller CPE corresponds to higher coverage probability (and smaller type I error).

The SEE estimator $\hat{\beta}$ satisfies $m_n(\hat{\beta}) = 0$. In Lemma 7 in the appendix, we show that

$$\sqrt{n}\Big(\hat{\beta} - \beta_0\Big) = -\left\{E\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\right\}^{-1} m_n + O_p\left(\frac{1}{\sqrt{nh}}\right), \tag{8}$$

and

$$E\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0) = E\big[Z_j X_j' f_{U|Z,X}(0|Z_j, X_j)\big] + O(h^r). \tag{9}$$

Consequently, the approximate MSE (AMSE) of $\sqrt{n}\Big(\hat{\beta} - \beta_0\Big)$ is[6]

$$\begin{aligned}
\text{AMSE}_\beta &= \left\{E\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\right\}^{-1}(E m_n m_n')\left\{E\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\right\}^{-1\prime} \\
&= \Sigma_{ZX}^{-1} V \Sigma_{XZ}^{-1} + \Sigma_{ZX}^{-1} V^{1/2}\big[nh^{2r}(EB)(EB)' - hE(AA')\big]\big(V^{1/2}\big)'\Sigma_{XZ}^{-1} \\
&\quad + O(h^r) + o(h + nh^{2r}),
\end{aligned}$$

where

$$\Sigma_{ZX} = E\big[Z_j X_j' f_{U|Z,X}(0|Z_j, X_j)\big] \text{ and } \Sigma_{XZ} = \Sigma_{ZX}'.$$

The first term of AMSE_β is the asymptotic variance of the unsmoothed QR estimator. The second term captures the higher-order effect of smoothing on the AMSE of $\sqrt{n}(\hat{\beta} - \beta_0)$. When $nh^r \to \infty$ and $n^3 h^{4r+1} \to \infty$, we have $h^r = o(nh^{2r})$ and $1/\sqrt{nh} = o(nh^{2r})$, so the terms of order $O_p(1/\sqrt{nh})$ in (8) and of order $O(h^r)$ in (9) are of smaller order than the $O(nh^{2r})$ and $O(h)$ terms in the AMSE. If $h \asymp n^{-1/(2r-1)}$ as before, these rate conditions are satisfied when $r > 2$.

Theorem 5. *Let Assumptions 1–5(i-ii) and 6 hold. Assume that (i) $f_{U|Z,X}(u|z,x)$ is r times continuously differentiable in u in a neighborhood of zero and for almost all $x \in \mathcal{X}$ and $Z \in \mathcal{Z}$ for $r > 2$, (ii) Σ_{ZX} is nonsingular. If $nh^r \to \infty$ and $n^3 h^{4r+1} \to \infty$, then the AMSE of $\sqrt{n}(\hat{\beta} - \beta_0)$ is*

$$\Sigma_{ZX}^{-1} V^{1/2}\big[I_d + nh^{2r}(EB)(EB)' - hE(AA')\big]\big(V^{1/2}\big)'(\Sigma_{ZX}')^{-1} + O(h^r) + o(h + nh^{2r}).$$

The optimal h^* that minimizes the high-order AMSE satisfies

$$\begin{aligned}
&\Sigma_{ZX}^{-1}\big[n(h^*)^{2r}(EB)(EB)' - h^* E(AA')\big](\Sigma_{ZX}')^{-1} \\
\leq\ &\Sigma_{ZX}^{-1}\big[nh^{2r}(EB)(EB)' - hE(AA')\big](\Sigma_{ZX}')^{-1}
\end{aligned}$$

[6]Here we follow a common practice in the estimation of nonparametric and nonlinear models and define the AMSE to be the MSE of $\sqrt{n}\Big(\hat{\beta} - \beta_0\Big)$ after dropping some smaller order terms. So the asymptotic MSE we define here is a Nagar-type approximate MSE. See Nagar (1959).

in the sense that the difference between the two sides is nonpositive definite for all h. This is equivalent to

$$n(h^*)^{2r}(EB)(EB)' - h^*E(AA') \leq nh^{2r}(EB)(EB)' - hE(AA').$$

This choice of h can also be motivated from the theory of optimal estimating equations. Given the estimating equations $m_n = 0$, we follow Heyde (1997) and define the standardized version of m_n by

$$m_n^s(\beta_0, h) = -E\frac{\partial}{\partial\beta'}m_n(\beta_0)[E(m_n m_n')]^{-1}m_n.$$

We include h as an argument of m_n^s to emphasize the dependence of m_n^s on h. The standardization can be motived from the following considerations. On one hand, the estimating equations need to be close to zero when evaluated at the true parameter value. Thus we want $E(m_n m_n')$ to be as small as possible. On the other hand, we want $m_n(\beta + \delta\beta)$ to differ as much as possible from $m_n(\beta)$ when β is the true value. That is, we want $E\frac{\partial}{\partial\beta'}m_n(\beta_0)$ to be as large as possible. To meet these requirements, we choose h to maximize

$$E\{m_n^s(\beta_0, h)[m_n^s(\beta_0, h)]'\} = \left[E\frac{\partial}{\partial\beta'}m_n(\beta_0)\right][E(m_n m_n')]^{-1}\left[E\frac{\partial}{\partial\beta'}m_n(\beta_0)\right]'.$$

More specifically, h^* is optimal if

$$E\{m_n^s(\beta_0, h^*)[m_n^s(\beta_0, h^*)]'\} - E\{m_n^s(\beta_0, h)[m_n^s(\beta_0, h)]'\}$$

is nonnegative definite for all $h \in \mathbb{R}^+$. But $E\{m_n^s(m_n^s)'\} = (\mathrm{AMSE}_\beta)^{-1}$, so maximizing $E\{m_n^s(m_n^s)'\}$ is equivalent to minimizing AMSE_β.

The question is whether such an optimal h exists. If it does, then the optimal h^* satisfies

$$h^* = \arg\min_h u'\left[nh^{2r}(EB)(EB)' - hE(AA')\right]u \tag{10}$$

for all $u \in \mathbb{R}^d$, by the definition of nonpositive definite plus the fact that the above yields a unique minimizer for any u. Using unit vectors $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, etc., for u, and noting that $\mathrm{tr}\{A\} = e_1'Ae_1 + \cdots + e_d'Ae_d$ for $d \times d$ matrix A, this implies that

$$\begin{aligned} h^* &= \arg\min_h \mathrm{tr}\left\{nh^{2r}(EB)(EB)' - hE(AA')\right\} \\ &= \arg\min_h \left[nh^{2r}(EB)'(EB) - h\mathrm{tr}\{E(AA')\}\right]. \end{aligned}$$

In view of (5), $h^*_{\text{SEE}} = h^*$ if h^* exists. Unfortunately, it is easy to show that no single h can minimize the objective function in (10) for all $u \in \mathbb{R}^d$. Thus, we have to redefine the optimality with respect to the direction of u. The direction depends on which linear combination of β is the focus of interest, as $u'\big[nh^{2r}(EB)(EB)' - hE(AA')\big]u$ is the high-order AMSE of $c'\sqrt{n}(\hat{\beta} - \beta_0)$ for $c = \Sigma_{XZ}\big(V^{-1/2}\big)'u$.

Suppose we are interested in only one linear combination. Let h^*_c be the optimal h that minimizes the high-order AMSE of $c'\sqrt{n}(\hat{\beta} - \beta_0)$. Then

$$h^*_c = \left(\frac{u'E(AA')u}{u'(EB)(EB)'u}\frac{1}{2nr}\right)^{\frac{1}{2r-1}}$$

for $u = \big(V^{1/2}\big)'\Sigma^{-1}_{XZ}c$. Some algebra shows that

$$h^*_c \geq \left(\frac{1}{(EB)'(EAA')^{-1}EB}\frac{1}{2nr}\right)^{\frac{1}{2r-1}} > 0.$$

So although h^*_c depends on c via u, it is nevertheless greater than zero.

Now suppose without loss of generality we are interested in d directions $(c_1, \ldots, c_d)$ jointly where $c_i \in \mathbb{R}^d$. In this case, it is reasonable to choose $h^*_{c_1,\ldots,c_d}$ to minimize the sum of direction-wise AMSEs, i.e.

$$h^*_{c_1,\ldots,c_d} = \arg\min_h \sum_{i=1}^{d} u_i'\big[nh^{2r}(EB)(EB)' - hE(AA')\big]u_i,$$

where $u_i = \big(V^{1/2}\big)'\Sigma^{-1}_{XZ}c_i$. It is easy to show that

$$h^*_{c_1,\ldots,c_d} = \left(\frac{\sum_{i=1}^d u_i'E(AA')u_i}{\sum_{i=1}^d u_i'(EB)(EB)'u_i}\frac{1}{2nr}\right)^{\frac{1}{2r-1}}.$$

As an example, consider $u_i = e_i = (0, \ldots, 1, \ldots, 0)$, the ith unit vector in $\mathbb{R}^d$. Correspondingly

$$(\tilde{c}_1, \ldots, \tilde{c}_d) = \Sigma_{XZ}\big(V^{-1/2}\big)'(e_1, \ldots, e_d).$$

It is clear that

$$h^*_{\tilde{c}_1,\ldots,\tilde{c}_d} = h^*_{\text{SEE}} = h^*_{\text{CPE}},$$

so all three selections coincide with each other. A special case of interest is when $Z = X$, non-constant regressors are pairwise independent and normalized to mean zero and variance one, and $U \perp\!\!\!\perp X$. Then $u_i = c_i = e_i$ and the d linear combinations reduce to the individual elements of β.

The above example illustrates the relationship between $h^*_{c_1,\ldots,c_d}$ and h^*_{SEE}. While $h^*_{c_1,\ldots,c_d}$ is tailored toward the flexible linear combinations $(c_1, \ldots, c_d)$ of the parameter vector, h^*_{SEE} is tailored toward the fixed $(\tilde{c}_1, \ldots, \tilde{c}_d)$. While $h^*_{c_1,\ldots,c_d}$ and h^*_{SEE} are of the same order of magnitude, in general there is no analytic relationship between $h^*_{c_1,\ldots,c_d}$ and h^*_{SEE}.

To shed further light on the relationship between $h^*_{c_1,\ldots,c_d}$ and h^*_{SEE}, let $\{\lambda_k, k = 1, \ldots, d\}$ be the eigenvalues of $nh^{2r}(EB)(EB)' - hE(AA')$ with the corresponding orthonormal eigenvectors $\{\ell_k, k = 1, \ldots, d\}$. Then we have $nh^{2r}(EB)(EB)' - hE(AA') = \sum_{k=1}^{d} \lambda_k \ell_k \ell_k'$ and $u_i = \sum_{j=1}^{d} u_{ij} \ell_j$ for $u_{ij} = u_i' \ell_j$. Using these representations, the objective function underlying $h^*_{c_1,\ldots,c_d}$ becomes

$$\begin{aligned}
&\sum_{i=1}^{d} u_i' \big[nh^{2r}(EB)(EB)' - hE(AA') \big] u_i \\
&= \sum_{i=1}^{d} \left(\sum_{j=1}^{d} u_{ij} \ell_j' \right) \left(\sum_{k=1}^{d} \lambda_k \ell_k \ell_k' \right) \left(\sum_{\tilde{j}=1}^{d} u_{i\tilde{j}} \ell_{\tilde{j}} \right) \\
&= \sum_{j=1}^{d} \left(\sum_{i=1}^{d} u_{ij}^2 \right) \lambda_j .
\end{aligned}$$

That is, $h^*_{c_1,\ldots,c_d}$ minimizes a weighted sum of the eigenvalues of $nh^{2r}(EB)(EB)' - hE(AA')$ with weights depending on $c_1, \ldots, c_d$. By definition, h^*_{SEE} minimizes the simple unweighted sum of the eigenvalues, viz. $\sum_{j=1}^{d} \lambda_j$. While h^*_{SEE} may not be ideal if we know the linear combination(s) of interest, it is a reasonable choice otherwise.

In empirical applications, we can estimate $h^*_{c_1,\ldots,c_d}$ using a parametric plug-in approach similar to our plug-in implementation of h^*_{SEE}. If we want to be agnostic about the directional vectors $c_1, \ldots, c_d$, we can simply use h^*_{SEE}.

6. Simulations

For our simulation study,[7] we use

$$G(u) = 0.5 + \frac{105}{64} \left(u - \frac{5}{3} u^3 + \frac{7}{5} u^5 - \frac{3}{7} u^7 \right)$$

as in Horowitz (1998), so $r = 4$. Using (the integral of) an Epanečnikov kernel with $r = 2$ also worked well in the cases where we tried it, though never better than $r = 4$. Note that our error distributions always have at least four derivatives, so $r = 4$ working somewhat better is expected. Selection of optimal r and $G(\cdot)$, and the quantitative impact thereof, remain open questions.

[7] MATLAB functions for public use are available on the first author's website. MATLAB code for the simulations is available upon request.

We implement a plug-in version of the infeasible $h^* := h^*_{\mathrm{SEE}}$. We make the plug-in assumption $U \perp\!\!\!\perp Z$ and parameterize the distribution of U. Our current method, which has proven quite accurate and stable, fits the residuals from an initial $h = (2nr)^{-1/(2r-1)}$ IV-QR to Gaussian, t, gamma, and generalized extreme value distributions via maximum likelihood. With the distribution parameter estimates, $f_U(0)$ and $f_U^{(r-1)}(0)$ can be computed and plugged in to calculate $\hat{h}$. Since the biggest risk is taking an h that is too large, we separately calculate $\hat{h}$ for each of the four distributions and take the smallest. Note that this particular plug-in approach works well even under heteroskedasticity and/or misspecification of the error distribution: settings 3.1-3.6 have error distributions other than these four, and settings 1.3, 2.2, 3.3-3.6 are heteroskedastic. For the infeasible h^*, if the PDF derivative in the denominator is zero, it is replaced by 0.01 to avoid $h^* = \infty$.

Motivated by the connection with mean IV regression in §2.2, we tried restricting (after computation) our estimators to be within a five-fold expansion of the d-dimensional rectangle with the unsmoothed IV-QR estimator at one corner and the IV (mean) estimator at the other, when our instruments Z are not simply X. This had a marginally (less than one percent) beneficial impact on MSE in some cases, but a detrimental impact in other cases, so we do not recommend this option or show it in our results.

For the unsmoothed IV-QR estimator, we use code based on Chernozhukov and Hansen (2006) from the latter author's website, for reasons given in their §3.3. We use the option to let their code determine the grid of possible endogenous coefficient values from the data. This code in turn uses the interior point method in `rq.m` (developed by Roger Koenker, Daniel Morillo, and Paul Eilers) to solve exogenous QR linear programs, as do we.

We tried data generating processes (DGPs) with homoskedasticity and heteroskedasticity, and with a variety of error distributions like Gaussian, Cauchy, exponential, and beta (of various shapes). Using $\hat{h}$ appears to consistently reduce the MSE of all estimator components compared with $h = 0$ and with IV ($h = \infty$). Almost always, the exception is cases where MSE is monotonically decreasing with h (IV is more efficient), in which $\hat{h}$ is much better than $h = 0$ but not quite large enough to match $h = \infty$. The range of $\hat{h}$ values over the simulation replications is usually less than a factor of 10, and the range from 0.05 to 0.95 empirical quantiles is around a factor of two. This is a very small impact—note the log transformation in the x-axis in the graphs.

For "size-adjusted" power of a test with nominal size α, the critical value is picked as the $(1-\alpha)$-quantile of the empirical test statistic distribution. This is for demonstration, not practice. The size adjustment fixes the left endpoint of the size-adjusted power curve to the null rejection probability α. The resulting size-adjusted power curve is one way to try to visualize a combination of type I and type II errors, in the absence of an explicit loss function. One shortcoming is that it does not reflect the variability/uniformity of size and power over the space of parameter values and DGPs.

Regarding notation in the figures, the y-axis in the size-adjusted power figures shows the simulated rejection probability. The x-axis shows the magnitude of deviation from the null hypothesis, where a randomized alternative is generated in each simulation iteration as that magnitude times a random point on the unit sphere in $\mathbb{R}^k$, where $\beta \in \mathbb{R}^k$. As the legend shows, the dashed line corresponds to the unsmoothed estimator ($h = 0$), the dotted line to the infeasible h^*_{SEE}, and the solid line to the plug-in $\hat{h}$.

For the MSE graphs, the flat horizontal solid and dashed lines are the MSE of the intercept and slope estimators (respectively) using feasible plug-in $\hat{h}$ (recomputed each replication). The other solid and dashed lines (that vary with h) are the MSE when using the value of h from the x-axis. The left y-axis shows the MSE values for the intercept parameter; the right y-axis shows the MSE for slope parameter(s); and the x-axis shows a log transformation of the bandwidth, $\log_{10}(1 + h)$.

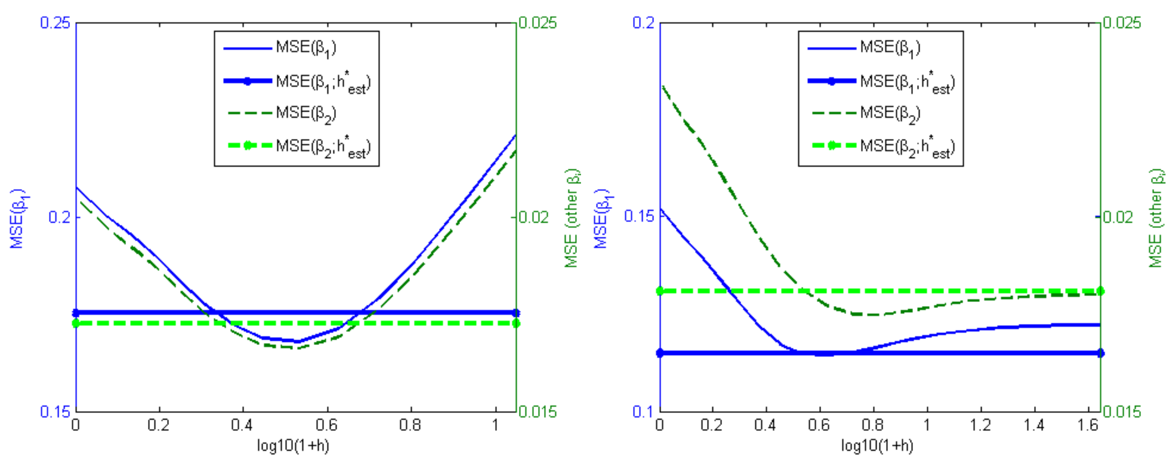

FIGURE 2. MSE for DGPs 1.1 (left) and 1.3 (right).

To save space, we report the following representative DGPs with 10,000 simulation replications each. Others produced very similar results. DGPs 1.* are the three from Horowitz (1998); 2.* are similar but with Cauchy errors and $q \neq 0.5$; 3.* include more error distributions; and 4.* are IV-QR. A qualitative description of the results is provided for each, and corresponding figures are noted when reproduced here. "SAP" below is "size-adjusted power." "Better" means $\hat{h}$ is better than $h = 0$; "worse" means $\hat{h}$ is worse than $h = 0$. "Percentage point(s)" is abbreviated "pp".

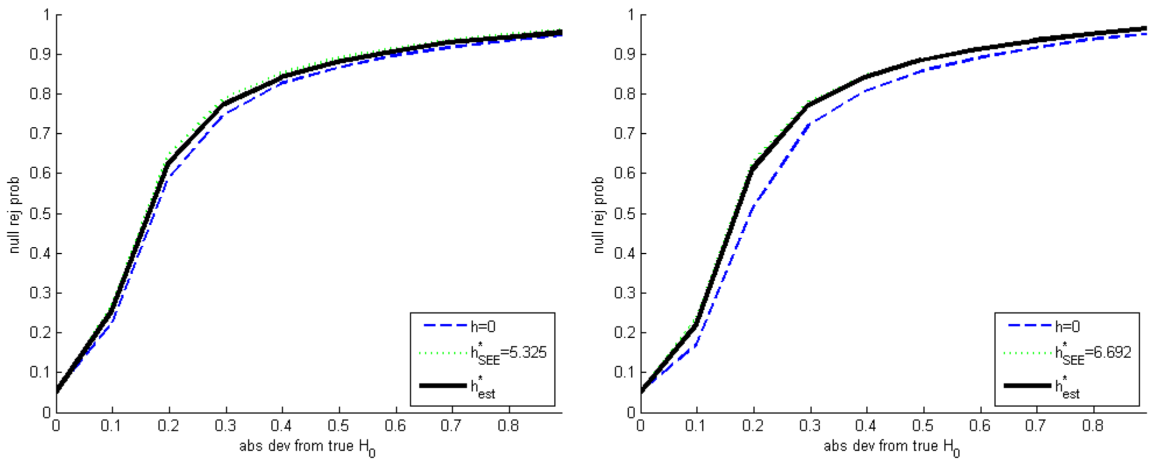

FIGURE 3. Size-adjusted power for DGPs 1.1 (left) and 1.3 (right).

1.1 DGP: (homoskedastic, thicker-tailed, $Z = X$) $q = 0.5$, $n = 50$, $\beta_0 = (1, 1)'$, errors from t_3 scaled to have variance two, non-constant regressor is Uniform$(1, 5)$. From Horowitz (1998). MSE: better than $h = 0$ and OLS for both intercept and slope; Figure 2. SAP: almost identical; Figure 3.

1.2 DGP: (homoskedastic, EV1, $Z = X$) $q = 0.5$, $n = 50$, $\beta_0 = (1, 1)'$, errors from Type I Extreme Value scaled/centered to have median zero and variance two, non-constant regressor is Uniform$(1, 5)$. From Horowitz (1998). MSE: better than $h = 0$ and OLS for both intercept and slope. SAP: a few pp better.

1.3 DGP: (heteroskedastic, thin-tailed, $Z = X$) $q = 0.5$, $n = 50$, $\beta_0 = (1, 1)'$, errors $U = 0.25(1 + x)V$ where $V \sim N(0, 1)$ and $x \sim$ Uniform$(1, 5)$ is the non-constant regressor. From Horowitz (1998). MSE: better than $h = 0$ for both intercept and slope; better than OLS for intercept, same for slope; Figure 2. SAP: a few pp better; Figure 3.

2.1 DGP: (homoskedastic, thick-tailed, $Z = X$) $q = 0.3$, $n = 50$, $\beta_0 = (1, 1)'$, Cauchy errors, non-constant regressor is Uniform$(0, 1)$. MSE: better than $h = 0$ and OLS; Figure 4. SAP: almost identical; Figure 5.

2.2 DGP: (heteroskedastic, thick-tailed, $Z = X$) $q = 0.35$, $n = 50$, $\beta_0 = (1, 1)'$, error $U = (1+x)V$ where V is a Cauchy (shifted to have 0.35-quantile equal to zero) and $x \sim$ Uniform$(0, 1)$ is the non-constant regressor. MSE: better than $h = 0$ and OLS; Figure 4. SAP: a few pp better; Figure 5.

3.1 DGP: (homoskedastic, uniform, $Z = X$, $d = 3$) $q = 0.5$, $n = 50$, $\beta_0 = (1, 1, 1)'$, uniform errors, $X = (1, x_1, x_2)'$ where $x_1 \sim$ Unif$(-5, 5)$, $x_2 \sim$ Unif$(5, 15)$. MSE: better than $h = 0$, worse than OLS. SAP: up to ten pp better.

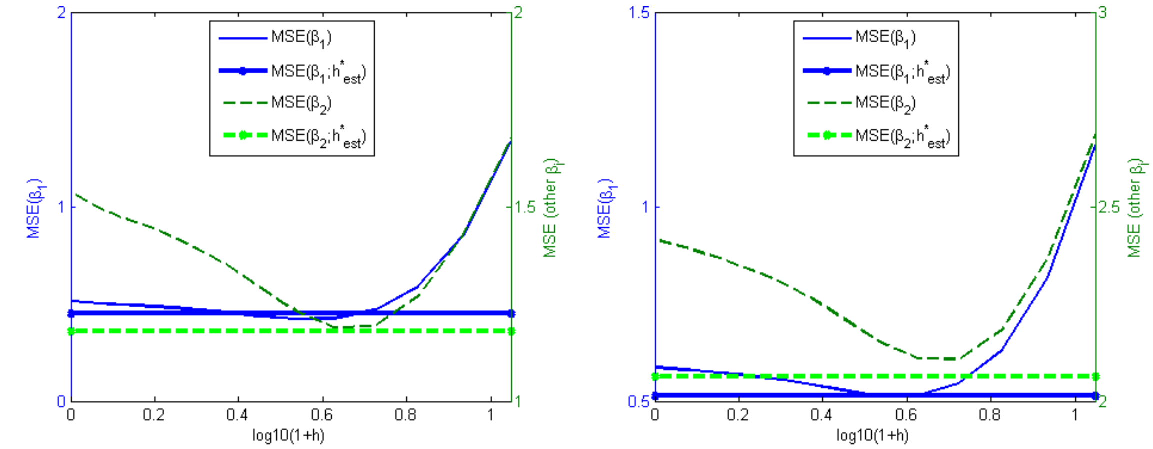

FIGURE 4. MSE for DGPs 2.1 (left) and 2.2 (right).

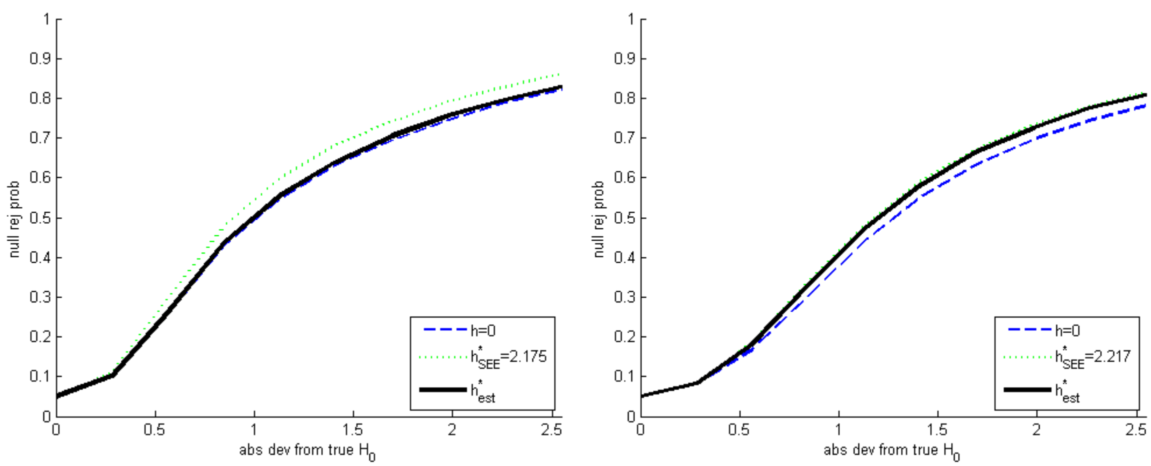

FIGURE 5. Size-adjusted power for DGPs 2.1 (left) and 2.2 (right).

3.2 DGP: Same as 3.1 but lognormal errors. MSE: $\hat{h}$ better than $h = 0$ and OLS. SAP: a few pp better over small range.
3.3 DGP: (heteroskedastic, uniform, $Z = X$) $q = 0.25$, $n = 50$, $\beta_0 = (1, 1)'$, $U = (1+x)V$ where V is uniform and $x \sim \text{Unif}(0, 1)$ is non-constant regressor. MSE: better than $h = 0$ and OLS; Figure 6. SAP: a few pp better.
3.4 DGP: (heteroskedastic, beta, $Z = X$) $q = 0.35$, $n = 50$, $\beta_0 = (1, 1)'$, $U = (1 + x)V$ where V is a (shifted) $\beta(2, 2)$ and $x \sim \text{Uniform}(0, 1)$ is the non-constant regressor. MSE: better than $h = 0$; better than OLS for intercept, same for slope. SAP: around 5pp better.
3.5 DGP: same as 3.4 but $\beta(1/2, 1/2)$ (U-shaped PDF), $n = 25$. MSE: better than $h = 0$; better than OLS for intercept, worse for slope. SAP: a few pp better.

3.6 DGP: same as 3.5 but $\beta(2,5)$ (skewed right). MSE: better than $h = 0$ and OLS; Figure 6. SAP: a few pp better.

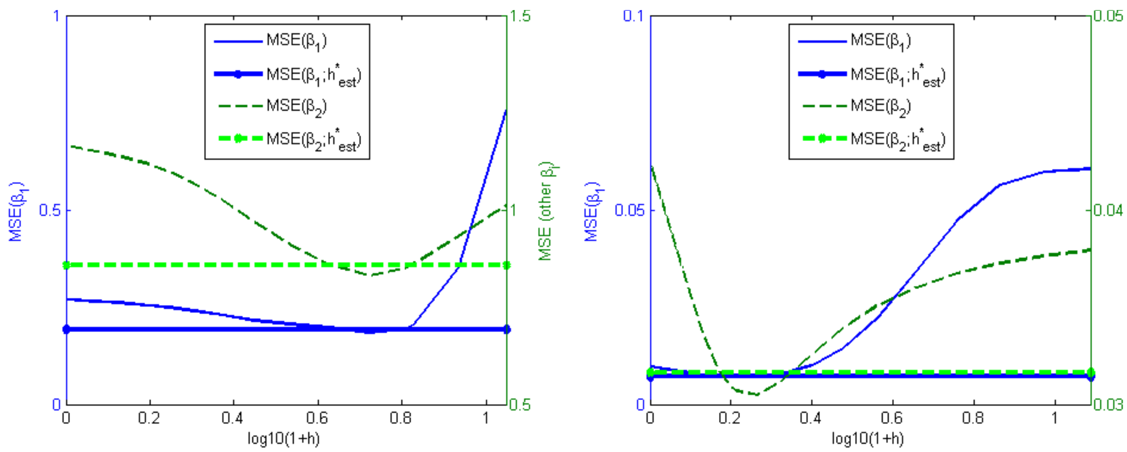

FIGURE 6. MSE for DGPs 3.3 (left) and 3.6 (right).

4.1 DGP: (normal, IV) $q = 0.5$, $n = 20$, $\beta_0 = (0, 1)'$. Simulated using reduced form equations in Cattaneo et al. (2012, equation 2) with $\gamma_1 = \gamma_2 = 1$, $x_i = 1$, $z_i \sim N(0, 1)$, and $\pi = 0.5$. Similar to their simulations, we set $\rho = 0.5$, $(\tilde{v}_{1i}, \tilde{v}_{2i})$ iid $N(0, 1)$, and $(v_{1i}, v_{2i})' = (\tilde{v}_{1i}, \sqrt{1-\rho^2}\tilde{v}_{2i} + \rho\tilde{v}_{1i})'$. MSE: regular MSE (not shown) is around 10 for $\hat{h}$ and around 100 for $h = 0$ and IV, due to "outlier" draws in a few percent of the simulation replications where both $h = 0$ and IV yield estimates far from β_0. In those draws, $\hat{h}$ is very large (but not to the point of equivalence with IV), in a range where there is significant bias but small enough variance to keep the estimates relatively close to β_0. With the "robust MSE" described in the figure caption, $\hat{h}$ performs better than $h = 0$ and worse than IV; Figure 7. SAP: up to 10pp better; Figure 9.

4.2 DGP: (Cauchy, IV) Similar to 4.1 but with $n = 250$, $(\tilde{v}_{1i}, \tilde{v}_{2i})'$ iid Cauchy, $\beta_0^{(2)} = [\rho - \sqrt{1-\rho^2}]^{-1}$ so that the structural error $u_i = v_{1i} - v_{2i}\beta$ is Cauchy with standard deviation $2[1-\rho/(\rho-\sqrt{1-\rho^2})]$. MSE: regular MSE (not shown) is in the hundred thousands for $\hat{h}$, similar for IV, and around 10^{11} for $h = 0$, due to "outlier" draws similar to discussed for DGP 4.1. With the "robust MSE" described in the figure caption, $\hat{h}$ is better than $h = 0$ and IV; Figure 7. SAP: a few pp worse.

4.3 DGP: (normal, IV) Same as 4.1 but $q = 0.35$ (and consequent re-centering of error term), $n = 30$. MSE: better than $h = 0$ for slope, same for intercept;

better than IV for intercept, worse for slope; Figure 8. SAP: over 5pp better; Figure 9.

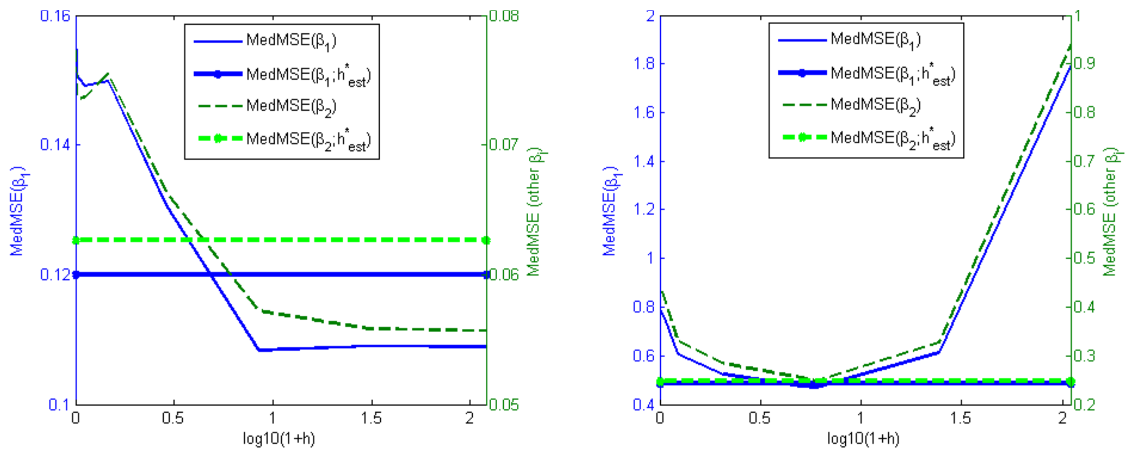

Figure 7. For DGPs 4.1 (left) and 4.2 (right), "robust MSE": squared median-bias plus the square of the interquartile range divided by 1.349, $\text{Bias}^2_{\text{median}} + (\text{IQR}/1.349)^2$.

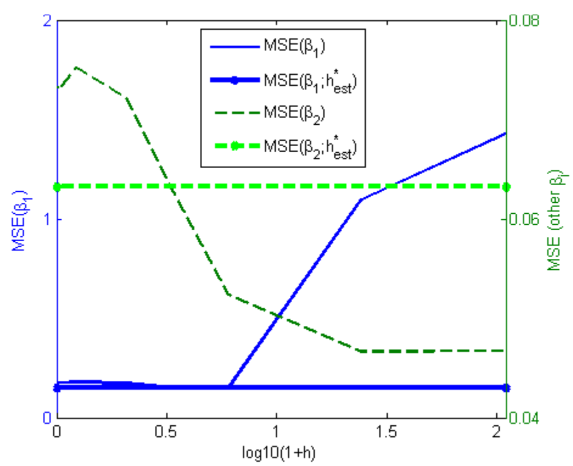

Figure 8. MSE for DGP 4.3.

With the infeasible h^*, there is usually some gain in size-adjusted power because the estimator is more precise. With a feasible h^*, this gain is in the 1–10 percentage point range for exogenous models ($Z = X$), though can be larger for IV setups. Depending on one's loss function of type I and type II error, this test may be preferred or not.

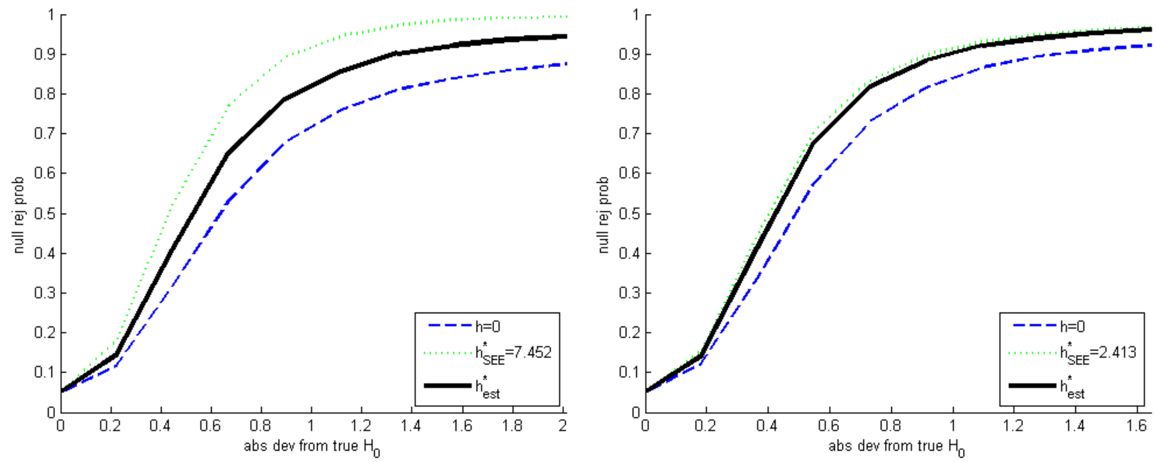

FIGURE 9. Size-adjusted power for DGPs 4.1 (left) and 4.3 (right).

7. Conclusion

We have presented a new estimator for quantile regression with or without instrumental variables. Smoothing the estimating equations (moment conditions) has multiple advantages beyond the known advantage of allowing higher-order expansions. It can reduce the MSE of both the estimating equations and the estimator, minimize type I error and improve size-adjusted power of a χ^2 test, and allow more reliable computation of the instrumental variables quantile regression estimator especially when the number of endogenous regressors is larger. We have given the theoretical bandwidth that optimizes these properties, and simulations show our plug-in bandwidth to reproduce all these advantages over the unsmoothed estimator. Links to mean instrumental variables regression and robust estimation are insightful and of practical use.

The strategy of smoothing the estimating equations can be applied to any model with nonsmooth estimating equations; there is nothing peculiar to the quantile regression model that we have exploited. For example, this strategy could be applied to censored quantile regression, or to select the optimal smoothing parameter in Horowitz's (2002) smoothed maximum score estimator. The present paper has focused on parametric and linear IV quantile regression; extensions to nonlinear IV quantile regression and nonparametric IV quantile regression along the lines of Chen and Pouzo (2009) would be interesting future topics.

References

Breiman, L. (1994). Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2012). Optimal inference for instrumental variables regression with non-Gaussian errors. *Journal of Econometrics*, 167:1–15.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525.

Chernozhukov, V., Hansen, C., and Jansson, M. (2009). Finite sample inference for quantile regression models. *Journal of Econometrics*, 152:93–103.

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

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer, New York.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, 66(6):1327–1351.

Horowitz, J. L. (2002). Bootstrap critical values for tests based on the smoothed maximum score estimator. *Journal of Econometrics*, 111:141–167.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Koenker, R. and Bassett, Jr., G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4):573–595.

Newey, W. K. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica*, 72(6):1877–1897.

Newey, W. K. and Powell, J. L. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory*, 6(3):295–317.

Otsu, T. (2008). Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, 142(1):508–538.

Phillips, P. C. B. (1982). Small sample distribution theory in econometric models of simultaneous models of simultaneous equations. Cowles Foundation Discussion Paper 617, Yale University.

Phillips, P. C. B. and Park, J. Y. (1988). On the formulation of wald tests of nonlinear restrictions. *Econometrica*, 56(5):1065–1083.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75(372):828–838.

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

Whang, Y.-J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, 22(2):173–205.

Zhou, Y., Wan, A. T. K., and Yuan, Y. (2011). Combining least-squares and quantile regressions. *Journal of Statistical Planning and Inference*, 141:3814–3828.

Appendix A. Appendix of Proofs

Proof of Theorem 1.

First moment of W_j. Let $[U_L(z), U_H(z)]$ be the support of the conditional PDF of U given $Z = z$. Since $P(U_j < 0|Z_j) = q$ for almost all Z_j and $h \to 0$, we can assume without loss of generality that $U_L(Z_j) \le -h$ and $U_H(Z_j) \ge h$ for almost all Z_j. For some $\tilde{h} \in [0, h]$, we have

$$
\begin{aligned}
E(W_j) &= E\{Z_j[G(-U_j/h) - q]\} = E\left\{\left(\int_{U_L(Z_j)}^{U_H(Z_j)} [G(-u/h) - q] dF_{U|Z}(u|Z_j)\right) Z_j\right\} \\
&= E\left\{\left([G(-u/h) - q]\, F_{U|Z}(u|Z_j)\Big|_{U_L(Z_j)}^{U_H(Z_j)} + \frac{1}{h}\int_{U_L(Z_j)}^{U_H(Z_j)} F_{U|Z}(u|Z_j) G'(-u/h) du\right) Z_j\right\} \\
&= E\left\{\left(-q + \int_{-1}^{1} F_{U|Z}(-hv|Z_j) G'(v) dv\right) Z_j\right\} \quad (\text{since } G'(v) = 0 \text{ for } v \notin [-1, 1]) \\
&= E\left\{\left[-q + F_{U|Z}(0|Z_j) + \int_{-1}^{1}\left(\sum_{k=1}^{r} f_{U|Z}^{(k-1)}(0|Z_j)\frac{(-h)^k v^k}{k!}\right) G'(v) dv\right] Z_j\right\} \\
&\quad + E\left\{\left[\int_{-1}^{1} f_{U|Z}^{(r)}(-\tilde{h}v|Z_j) v^r G'(v) dv\right] Z_j\right\}\frac{(-h)^{r+1}}{(r+1)!} \\
&= \frac{(-h)^r}{r!}\left(\int_{-1}^{1} G'(v) v^r dv\right) E\left[f_{U|Z}^{(r-1)}(0|Z_j) Z_j\right] \\
&\quad + E\left\{\left[\int_{-1}^{1} f_{U|Z}^{(r)}(-\tilde{h}v|Z_j) v^r G'(v) dv\right] Z_j\right\} O\big(h^{r+1}\big).
\end{aligned}
$$

Under Assumption 4, for some bounded $C(\cdot)$ we have

$$\begin{aligned}&\left\| E\left\{ \left[\int_{-1}^{1} f_{U|Z}^{(r)}(-\tilde{h}v|Z)v^r G'(v)dv \right] Z \right\} \right\| \\ &\quad \le E\left\{ \int_{-1}^{1} C(Z)\|Z\||v^r G'(v)|dv \right\} = O(1).\end{aligned}$$

Hence

$$E(W_j) = \frac{(-h)^r}{r!}\left(\int_{-1}^{1} G'(v)v^r dv \right) E\left[f_{U|Z}^{(r-1)}(0|Z_j)Z_j \right] + o(h^r).$$

Second moment of W_j. For the second moment,

$$EW_j'W_j = E\{[G(-U_j/h) - q]^2 Z_j'Z_j\} = E\left\{ \left(\int_{U_L(Z_j)}^{U_H(Z_j)} [G(-u/h) - q]^2 dF_{U|Z}(u|Z_j) \right) Z_j'Z_j \right\}.$$

Integrating by parts and using Assumption 4(i) in the last line yields:

$$\begin{aligned}&\int_{U_L(Z_j)}^{U_H(Z_j)} [G(-u/h) - q]^2 dF_{U|Z}(u|Z_j) \\ &= [G(-u/h) - q]^2 F_{U|Z}(u|Z_j)\Big|_{U_L(Z_j)}^{U_H(Z_j)} + \frac{2}{h}\int_{U_L(Z_j)}^{U_H(Z_j)} F_{U|Z}(u|Z_j)[G(-u/h) - q]G'(-u/h)du \\ &= q^2 + 2\int_{-1}^{1} F_{U|Z}(hv|Z_j)[G(-v) - q]G'(-v)dv \quad (\text{since } G'(v) = 0 \text{ for } v \notin [-1,1]) \\ &= q^2 + 2q\left\{ \int_{-1}^{1} [G(-v) - q]G'(-v)dv \right\} + 2hf_{U|Z}(0|Z_j)\left\{ \int_{-1}^{1} v[G(-v) - q]G'(-v)dv \right\} \\ &\quad + \left\{ \int_{-1}^{1} v^2 f_{U|Z}'(\tilde{h}v|Z_j)[G(-v) - q]G'(-v)dv \right\} h^2.\end{aligned}$$

But

$$\begin{aligned}&q^2 + 2q\int_{-1}^{1} [G(-v) - q]G'(-v)dv \\ &\qquad = q^2 + q\int_{-1}^{1} 2[G(u) - q]G'(u)du \\ &\qquad = q^2 + q\left[G^2(u) - 2qG(u) \right]\Big|_{-1}^{1} = q^2 + q(1 - 2q) = q(1 - q), \\ &2\int_{-1}^{1} v[G(-v) - q]G'(-v)dv \\ &\qquad = -2\int_{-1}^{1} u[G(u) - q]G'(u)du = -2\int_{-1}^{1} uG(u)G'(u)du \quad (\text{by Assumption 5(ii)})\end{aligned}$$

and

$$\left|\int_{-1}^{1} v^2 f'_{U|Z}(\tilde{h}v|Z_j)[G(-v)-q]G'(-v)dv\right| \leq \int_{-1}^{1} C(Z_j)\left|v^2 G'(v)\right|dv$$

for some function $C(\cdot)$. So

$$\begin{aligned}
&E(W_j'W_j)\\
&\quad= E\left(\left\{q^2 + q(1-2q) - hf_{U|Z}(0|Z_j)\left[1-\int_{-1}^{1} G^2(u)du\right]\right\}Z_j'Z_j\right) + O(h^2)\\
&\quad= q(1-q)E\left[Z_j'Z_j\right] - h\left[1-\int_{-1}^{1} G^2(u)du\right]E\left[f_{U|Z}(0|Z_j)Z_j'Z_j\right] + O(h^2).
\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}
&E(W_jW_j')\\
&\quad= q(1-q)E(Z_jZ_j') - h\left[1-\int_{-1}^{1} G^2(u)du\right]E\left[f_{U|Z}(0|Z_j)Z_jZ_j'\right] + O(h^2).
\end{aligned}$$

First-order asymptotic distribution of m_n. We can write m_n as

$$m_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_j = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(W_j - EW_j) + \sqrt{n}EW_j. \tag{11}$$

In view of the mean of W_j, we have $\sqrt{n}EW_j = O(h^r\sqrt{n}) = o(1)$ by Assumption 6. So the bias is asymptotically (first-order) negligible. Consequently, the variance of W_j is $E(W_jW_j') + o(1)$, so the first-order term from the second moment calculation above can be used for the asymptotic variance.

Next, we apply the Lindeberg-Feller central limit theorem to the first term in (11), which is a scaled sum of a triangular array since the bandwidth in W_j depends on n. We consider the case when W_j is a scalar as vector cases can be handled using the Cramér-Wold device. Note that

$$\begin{aligned}
\sigma_W^2 &\equiv \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(W_j - EW_j)\right] = n\frac{1}{n}\text{Var}(W_j - E(W_j)) \quad \text{(by iid Assumption 1)}\\
&= EW_j^2 - (EW_j)^2 = q(1-q)E\left[Z_j^2\right](1+o(1)).
\end{aligned}$$

For any $\varepsilon > 0$,

$$\lim_{n\to\infty}\sum_{j=1}^{n} E\left(\frac{W_j - EW_j}{\sqrt{n}\sigma_W}\right)^2\left\{\frac{|W_j - EW_j|}{\sqrt{n}\sigma_W} \geq \varepsilon\right\}$$

$$= \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} E\frac{(W_j - EW_j)^2}{\sigma_W^2}\left\{\frac{|W_j - EW_j|}{\sigma_W} \ge \sqrt{n}\varepsilon\right\}$$

$$= \lim_{n\to\infty} E\frac{(W_j - EW_j)^2}{\sigma_W^2}\left\{\frac{|W_j - EW_j|}{\sigma_W} \ge \sqrt{n}\varepsilon\right\} = 0,$$

where the last equality follows from the dominated convergence theorem, as

$$\frac{(W_j - EW_j)^2}{\sigma_W^2}\left\{\frac{|W_j - EW_j|}{\sigma_W} \ge \sqrt{n}\varepsilon\right\} \le C\frac{Z_j^2 + EZ_j^2}{\sigma_W^2}$$

for some constant C and $EZ_j^2 < \infty$. So the Lindeberg condition holds and $m_n \xrightarrow{d} N(0, V)$.

Bias of estimating equations derived from smoothed criterion function. With the EE derived from smoothing the criterion function,

$$E\left[G\left(-\frac{U_j}{h}\right) - q\right]Z_j - \frac{1}{h}EU_jG'\left(-\frac{U_j}{h}\right)Z_j$$

$$= \frac{(-h)^r}{r!}\left(\int G'(v)v^r dv\right)Ef_{U|Z}^{(r-1)}(0|Z_j)Z_j + o(h^r) - h\int vG'(v)f_{U_i|Z_i}(-hv)dvZ_j$$

$$= \frac{(-h)^r}{r!}\left(\int G'(v)v^r dv\right)Ef_{U|Z}^{(r-1)}(0|Z_j)Z_j + o(h^r)$$

$$- h\frac{(-h)^{r-1}}{(r-1)!}\left(\int G'(v)v^r dv\right)Ef_{U|Z}^{(r-1)}(0|Z_j)Z_j + o(h^r)$$

$$= (-h)^r\left(\frac{1}{r!} + \frac{1}{(r-1)!}\right)\left(\int G'(v)v^r dv\right)Ef_{U|Z}^{(r-1)}(0|Z_j)Z_j + o(h^r).$$

Proof of Proposition 2. The first expression comes directly from the FOC. The simplified h^*_{SEE} is calculated using the following lemma.

Lemma 6. *If $Z \in \mathbb{R}^d$ is a random vector with first element equal to one and $V \equiv E(ZZ')$ is nonsingular, then*

$$E(Z'V^{-1}Z)/[(EZ')V^{-1}(EZ)] = d.$$

Proof. For the numerator, rearrange using the trace:

$$E(Z'V^{-1}Z) = E\left[\text{tr}\{Z'V^{-1}Z\}\right] = E\left[\text{tr}\{V^{-1}ZZ'\}\right] = \text{tr}\{V^{-1}E(ZZ')\} = \text{tr}\{I_d\} = d.$$

For the denominator, let $E(Z') = (1, t')$ for some $t \in \mathbb{R}^{d-1}$. Since the first element of Z is one, the first row and first column of V are $E(Z')$ and $E(Z)$. Writing the other

$(d-1)\times(d-1)$ part of the matrix as Ω,

$$V = E(ZZ') = \begin{pmatrix} 1 & t' \\ t & \Omega \end{pmatrix}.$$

We can read off $V^{-1}E(Z) = (1, 0, \ldots, 0)'$ from the first column of the identity matrix since

$$V^{-1}\begin{pmatrix} 1 & v' \\ v & \Omega \end{pmatrix} = V^{-1}V = I_d = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Thus,

$$(EZ')V^{-1}(EZ) = (1, t')(1, 0, \ldots, 0)' = 1. \qquad \square$$

Proof of Theorem 3. Adding to the variables already defined in the main text, let

$$Z_j^* \equiv \left(EZ_jZ_j'\right)^{-1/2} Z_j \text{ and } D_n \equiv n^{-1}\sum_{j=1}^{n}\left(Z_j^* Z_j^{*\prime} - EZ_j^* Z_j^{*\prime}\right) = \frac{1}{n}\sum_{j=1}^{n} Z_j^* Z_j^{*\prime} - I_d.$$

Then using the definition of Λ_n in (6), we have

$$\begin{aligned}
\Lambda_n^{-1}\hat{V}\left(\Lambda_n^{-1}\right)' &= n^{-1}\sum_{j=1}^{n}\Lambda_n^{-1}Z_j\left(\Lambda_n^{-1}Z_j\right)' q(1-q) \\
&= \left(I_d - E(AA')h + O(h^2)\right)^{-1/2}\left[\frac{1}{n}\sum_{j=1}^{n} Z_j^* Z_j^{*\prime}\right]\left(I_d - E(AA')h + O(h^2)\right)^{-1/2} \\
&= \left(I_d - E(AA')h + O(h^2)\right)^{-1/2}[I_d + D_n]\left(I_d - E(AA')h + O(h^2)\right)^{-1/2} \\
&= \left[I_d + (1/2)E(AA')h + O(h^2)\right][I_d + D_n]\left[I_d + (1/2)E(AA')h + O(h^2)\right].
\end{aligned}$$

Let $\xi_n = (I_d + D_n)^{-1} - (I_d - D_n) = (I_d + D_n)^{-1}D_n^2$, then

$$\begin{aligned}
&\left[\Lambda_n^{-1}\hat{V}\left(\Lambda_n^{-1}\right)'\right]^{-1} \\
= &\left[I_d - \frac{1}{2}E(AA')h + O(h^2)\right][I_d - D_n + \xi_n]\left[I_d - \frac{1}{2}E(AA')h + O(h^2)\right] \\
= &\ I_d - E(AA')h + \eta_n, \qquad (12)
\end{aligned}$$

where $\eta_n = -D_n + D_nO(h) + \xi_n + O(h^2) + \xi_nO(h)$ collects the remainder terms. To evaluate the order of η_n, we start by noting that $E\left(\|D_n\|^2\right) = O(1/n)$. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the smallest and largest eigenvalues of a matrix, then for any constant

$C > 2\sqrt{d} > 0$:

$$\begin{aligned} & P\big\{\big\|(I_d + D_n)^{-1}\big\| \geq C\big\} \leq P\Big\{\lambda_{\max}((I_d + D_n)^{-1}) \geq C/\sqrt{d}\Big\} \\ = \; & P\Big\{\lambda_{\min}(I_d + D_n) \leq \sqrt{d}/C\Big\} = P\Big\{1 + \lambda_{\min}(D_n) \leq \sqrt{d}/C\Big\} \\ = \; & P(\lambda_{\min}(D_n) \leq -1/2) \leq P\big(\lambda^2_{\min}(D_n) > 1/4\big) \\ \leq \; & P\big(\|D_n\|^2 > 1/4\big) = O\left(\frac{1}{n}\right) \end{aligned}$$

by the Markov inequality. Using this probability bound and the Chernoff bound, we have for any $\epsilon > 0$,

$$\begin{aligned} & P\left\{\frac{n}{\log n}\|\xi_n\| > \epsilon\right\} \leq P\left\{\frac{n}{\log n}\big\|(I_d + D_n)^{-1}\big\| \times \|D_n\|^2 > \epsilon\right\} \\ = \; & P\left\{n\|D_n\|^2 > \frac{\epsilon \log n}{C}\right\} + P\big\{\big\|(I_d + D_n)^{-1}\big\| > C\big\} = O\left(\frac{1}{n}\right). \end{aligned}$$

It then follows that

$$P\left\{\|\eta_n\| \geq C \max\left(h^2, \sqrt{\frac{\log n}{n}}, h\sqrt{\frac{\log n}{n}}, \frac{\log n}{n}, \frac{h \log n}{n}\right)\right\} = O\left(\frac{1}{n} + h^2\right).$$

Under Assumption 6, we can rewrite the above as

$$P\big\{\|\eta_n\| \geq Ch^2/\log n\big\} = O\big(h^2\big) \tag{13}$$

for any constant $C > 0$.

Using (12) and defining $W_j^* \equiv \Lambda_n^{-1} Z_j [G(-U_j/h) - q]$, we have

$$\begin{aligned} S_n &= \big(\Lambda_n^{-1} m_n\big)' \Lambda_n' \hat{V}^{-1} \Lambda_n \big(\Lambda_n^{-1} m_n\big) \\ &= \big(\Lambda_n^{-1} m_n\big)' \Big[\Lambda_n^{-1} \hat{V} \big(\Lambda_n^{-1}\big)'\Big]^{-1} \big(\Lambda_n^{-1} m_n\big) = S_n^L + e_n \end{aligned}$$

where

$$\begin{aligned} S_n^L &= \big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big) - h\big(\sqrt{n}\bar{W}_n^*\big)' E(AA')\big(\sqrt{n}\bar{W}_n^*\big), \\ e_n &= \big(\sqrt{n}\bar{W}_n^*\big)' \eta_n \big(\sqrt{n}\bar{W}_n^*\big), \end{aligned}$$

and $\bar{W}_n^* = n^{-1}\sum_{j=1}^n W_j^*$ as defined in (7). Using the Chernoff bound on $\sqrt{n}\bar{W}_n^*$ and the result in (13), we can show that $P(|e_n| > Ch^2) = O(h^2)$. This ensures that we can ignore e_n to the order of $O(h^2)$ in approximating the distribution of S_n.

The characteristic function of S_n^L is

$$E\big\{\exp(itS_n^L)\big\} = C_0(t) - hC_1(t) + O\big(h^2\big) \text{ where}$$

$$C_0(t) \equiv E\Big\{\exp\Big[it\big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big)\Big]\Big\},$$
$$C_1(t) \equiv E\Big\{it\big(\sqrt{n}\bar{W}_n^*\big)'(EAA')\big(\sqrt{n}\bar{W}_n^*\big)\exp\Big[it\big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big)\Big]\Big\}.$$

Following Phillips (1982) and using arguments similar to those in Horowitz (1998) and Whang (2006), we can establish an asymptotic expansion of the density of $n^{-1/2}\sum_{j=1}^{n}\big(W_j^* - EW_j^*\big)$ of the form

$$pdf(x) = (2\pi)^{-d/2}\exp(-x'x/2)[1 + n^{-1/2}p(x)] + O(n^{-1}),$$

where $p(x)$ is an odd polynomial in the elements of x of degree 3. When $d = 1$, we know from Hall (1992, §2.8) that

$$p(x) = -\frac{\kappa_3}{6}\frac{1}{\phi(x)}\frac{d}{dx}\phi(x)(x^2-1) \quad \text{for} \quad \kappa_3 = \frac{E\big(W_j^* - EW_j^*\big)^3}{V_n^{3/2}} = O(1).$$

We use this expansion to compute the dominating terms in $C_j(t)$ for $j = 0, 1$.

First,

$$\begin{aligned}
C_0(t) &= E\Big\{\exp\Big[it\big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big)\Big]\Big\} \\
&= (2\pi)^{-d/2}\int \exp\Big\{it\big[x + \sqrt{n}EW_j^*\big]'\big[x + \sqrt{n}EW_j^*\big]\Big\}\exp\Big(-\frac{1}{2}x'x\Big)dx + O(n^{-1}) \\
&\quad + \frac{1}{\sqrt{n}}(2\pi)^{-d/2}\int \exp\Big\{it\big[x + \sqrt{n}EW_j^*\big]'\big[x + \sqrt{n}EW_j^*\big]\Big\}p(x)\exp\Big(-\frac{1}{2}x'x\Big)dx \\
&= (1-2it)^{-d/2}\exp\left(\frac{i\big\|\sqrt{n}EW_j^*\big\|^2 t}{1-2it}\right) + O(n^{-1}) \\
&\quad + \frac{1}{\sqrt{n}}(2\pi)^{-d/2}\int p(x)\exp\Big\{-\frac{1}{2}x'x\Big\}\big(1 + it2x'\sqrt{n}EW_j^* + O(n\|EW_j^*\|^2)\big)dx \\
&= (1-2it)^{-d/2}\exp\left(\frac{i\big\|\sqrt{n}EW_j^*\big\|^2 t}{1-2it}\right) + O\big(\big\|EW_j^*\big\| + \sqrt{n}h^{2r} + n^{-1}\big) \\
&= (1-2it)^{-d/2}\exp\left(\frac{i\big\|\sqrt{n}EW_j^*\big\|^2 t}{1-2it}\right) + O(h^r),
\end{aligned}$$

where the third equality follows from the characteristic function of a noncentral chi-square distribution.

Second, for $C_1(t)$ we can put any $o(1)$ term into the remainder since $hC_1(t)$ will then have remainder $o(h)$. Noting that x is an odd function (of x) and so integrates to zero against any symmetric PDF,

$$
\begin{aligned}
C_1(t) &= E\Big\{it\big(\sqrt{n}\bar{W}_n^*\big)'E(AA')\big(\sqrt{n}\bar{W}_n^*\big)\exp\Big[it\big(\sqrt{n}\bar{W}_n^*\big)'\big(\sqrt{n}\bar{W}_n^*\big)\Big]\Big\}\\
&= (2\pi)^{-d/2}\int it\big(x+\sqrt{n}EW_j^*\big)'E(AA')\big(x+\sqrt{n}EW_j^*\big)\\
&\quad\times\exp it\big[x+\sqrt{n}EW_j^*\big]'\big[x+\sqrt{n}EW_j^*\big]\exp\Big(-\frac{1}{2}x'x\Big)dx\\
&\quad\times\Big(1+O\Big(\frac{1}{\sqrt{n}}\Big)\Big)\\
&= (2\pi)^{-d/2}\int itx'E(AA')x\exp\Big[-\frac{1}{2}x'x(1-2it)\Big]dx+O\Big(\big(\sqrt{n}EW_j^*\big)^2\Big)+O\big(EW_j^*\big)\\
&= (1-2it)^{-d/2}it(\mathrm{tr}E(AA')E\mathbb{X}\mathbb{X}')+O\Big(\big\|\sqrt{n}EW_j^*\big\|^2\Big)+O\big(\big\|EW_j^*\big\|\big)\\
&= (1-2it)^{-d/2-1}it(\mathrm{tr}E(AA'))+O\Big(\big\|\sqrt{n}EW_j^*\big\|^2\Big)+O\big(\big\|EW_j^*\big\|\big),
\end{aligned}
$$

where $\mathbb{X}\sim N(0,diag(1-2it)^{-1})$.

Combining the above steps, we have, for $r\geq 2$,

$$
\begin{aligned}
E\big\{\exp(itS_n^L)\big\} &= \overbrace{(1-2it)^{-d/2}\exp\left(\frac{i\big\|\sqrt{n}EW_j^*\big\|^2t}{1-2it}\right)}^{C_0(t)} - h\,\overbrace{(1-2it)^{-d/2-1}it(\mathrm{tr}E(AA'))}^{O(1)\text{ term in }C_1(t)}\\
&\quad+\overbrace{O\big(nh^{2r+1}\big)+O\big(h^{r+1}\big)}^{\text{Remainder from }hC_1(t)}+O(h^2)\\
&= (1-2it)^{-d/2}+(1-2it)^{-d/2-1}it\Big\{\big\|\sqrt{n}EW_j^*\big\|^2-h(\mathrm{tr}E(AA'))\Big\}\\
&\quad+O(h^2+nh^{2r+1}).
\end{aligned}
$$

Taking a Fourier-Stieltjes inversion, we have

$$
P\big(S_n^L<x\big)=\mathcal{G}_d(x)-\frac{1}{d}\mathcal{G}_d'(x)x\Big\{\big\|\sqrt{n}EW_j^*\big\|^2-h(\mathrm{tr}E(AA'))\Big\}+O(h^2+nh^{2r+1}).
$$

See Phillips and Park (1988, Theorem 2.4).

A direct implication is that type I error is

$$
P\Big(m_n'\hat{V}^{-1}m_n>c_\alpha\Big)=\alpha+\frac{1}{d}\mathcal{G}_d'(c_\alpha)c_\alpha\Big\{\big\|\sqrt{n}EW_j^*\big\|^2-h(\mathrm{tr}E(AA'))\Big\}+O(h^2+nh^{2r+1}).\quad\square
$$

Lemma 7. *Let the assumptions in Theorem 5 hold. Then*

$$
\sqrt{n}(\hat{\beta}-\beta_0)=-\Big\{E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\Big\}^{-1}m_n+O_p\Big(\frac{1}{\sqrt{nh}}\Big)+O_p\Big(\frac{1}{\sqrt{n}}\Big),
$$

and

$$E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)=\Sigma_{ZX}+O(h^r).$$

Proof. We first prove that $\hat{\beta}$ is consistent. Using the Markov inequality, we can show that when $E(\|Z_j\|^2)<\infty$,

$$\frac{1}{\sqrt{n}}m_n(\beta)=\frac{1}{\sqrt{n}}Em_n(\beta)+o_p(1)$$

for each $\beta\in\mathcal{B}$. It is easy to show that the above $o_p(1)$ term also holds uniformly over $\beta\in\mathcal{B}$. But

$$\begin{aligned}&\lim_{h\to0}\sup_{\beta\in\mathcal{B}}\left\|\frac{1}{\sqrt{n}}Em_n(\beta)-E(Z[1\{Y<X'\beta\}-q])\right\|\\ =\;&\lim_{h\to0}\max_{\beta\in\mathcal{B}}\left\|EZ\left[G\left(\frac{X'\beta-Y}{h}\right)-1\{Y<X'\beta\}\right]\right\|\\ =\;&\lim_{h\to0}\left\|EZ\left[G\left(\frac{X'\beta^*-Y}{h}\right)-1\{Y<X'\beta^*\}\right]\right\|=0\end{aligned}$$

by the dominated convergence theorem, where β^* is the value of β that achieves the maximum. Hence

$$\frac{1}{\sqrt{n}}m_n(\beta)=E(Z[1\{Y<X'\beta\}-q])+o_p(1)$$

uniformly over $\beta\in\mathcal{B}$. Given the uniform convergence and the identification condition in Assumption 2, we can invoke Theorem 5.9 of van der Vaart (1998) to obtain that $\hat{\beta}\to\beta_0$.

Next we prove the first result of the lemma. Under Assumption 5(i–ii), we can use the elementwise mean value theorem to obtain

$$\sqrt{n}(\hat{\beta}-\beta_0)=-\left[\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\left(\tilde{\beta}\right)\right]^{-1}m_n$$

where

$$\frac{\partial}{\partial\beta'}m_n(\tilde{\beta})=\left[\frac{\partial}{\partial\beta}m_{n,1}(\tilde{\beta}_1),\ldots,\frac{\partial}{\partial\beta}m_{n,d}(\tilde{\beta}_d)\right]'$$

and each $\tilde{\beta}_i$ is a point between $\hat{\beta}$ and β_0. Under Assumptions 1 and 5(i–ii) and that $E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta)$ is continuous at $\beta=\beta_0$, we have, using standard textbook arguments, that $\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\left(\tilde{\beta}\right)=\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)+o_p(1)$. But

$$\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)=\frac{1}{nh}\sum_{j=1}^{n}Z_jX_j'G'\left(-\frac{U_j}{h}\right)\xrightarrow{p}\Sigma_{ZX}.$$

Hence, under the additional Assumption 6 and nonsingularity of Σ_{ZX}, we have $\sqrt{n}(\hat{\beta}-\beta_0) = O_p(1)$. With this rate of convergence, we can focus on a $\sqrt{n}$ neighborhood $\mathcal{N}_0$ of β_0. We write

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) = -\left\{\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0) + \left[\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}\left[m_n\left(\tilde{\beta}\right) - m_n(\beta_0)\right]\right]\right\}^{-1} m_n.$$

Using standard arguments again, we can obtain the following stochastic equicontinuity result:

$$\sup_{\beta\in\mathcal{N}_0}\left\|\left[\frac{\partial}{\partial\beta'}m_n(\beta) - E\frac{\partial}{\partial\beta'}m_n(\beta)\right] - \left[\frac{\partial}{\partial\beta'}m_n(\beta_0) - E\frac{\partial}{\partial\beta'}m_n(\beta_0)\right]\right\| = o_p(1),$$

which, combined with the continuity of $E\frac{\partial}{\partial\beta'}m_n(\beta)$, implies that

$$\left[\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}\left[m_n\left(\tilde{\beta}\right) - m_n(\beta_0)\right]\right] = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Therefore

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right) &= -\left\{\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0) + O_p\left(\frac{1}{\sqrt{n}}\right)\right\}^{-1} m_n \\ &= -\left\{\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\right\}^{-1} m_n + O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Now

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{vec}\left[\frac{\partial}{\partial\beta'}m_n/\sqrt{n}\right]\right) \\ &= n^{-1}\mathrm{Var}\left[\mathrm{vec}\left(Z_jX_j'\right)h^{-1}G'(-U_j/h)\right] \\ &\le n^{-1}E\left[\mathrm{vec}\left(Z_jX_j'\right)\left[\mathrm{vec}\left(Z_jX_j'\right)\right]'h^{-2}[G'(-U_j/h)]^2\right] \\ &= n^{-1}E\left\{\mathrm{vec}\left(Z_jX_j'\right)\left[\mathrm{vec}\left(Z_jX_j'\right)\right]'\int h^{-2}[G'(-u/h)]^2 f_{U|Z,X}(u|Z_j,X_j)du\right\} \\ &= (nh)^{-1}E\left\{\mathrm{vec}\left(Z_jX_j'\right)\left[\mathrm{vec}\left(Z_jX_j'\right)\right]'\int [G'(v)]^2 f_{U|Z,X}(-hv|Z_j,X_j)dv\right\} \\ &= O\left(\frac{1}{nh}\right),\end{aligned}$$

so

$$\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n = E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n + O_p\left(\frac{1}{\sqrt{nh}}\right).$$

As a result,

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) = -\left\{E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n + O_p\left(\frac{1}{\sqrt{nh}}\right)\right\}^{-1} m_n + O_p\left(\frac{1}{\sqrt{n}}\right)$$

$$= \quad -\left\{E\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\right\}^{-1} m_n + O_p\left(\frac{1}{\sqrt{nh}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right).$$

For the second result of the lemma, we use the same technique as in the proof of Theorem 1. We have

$$\begin{aligned}
E\left[\frac{\partial}{\partial\beta'}m_n/\sqrt{n}\right] &= E\left[\frac{1}{nh}\sum_{j=1}^{n} Z_j X_j' G'(-U_j/h)\right] = E\big[E\{Z_j X_j' h^{-1} G'(-U_j/h)|Z_j, X_j\}\big] \\
&= E\left[Z_j X_j' \int G'(-u/h) f_{U|Z,X}(u|Z_j, X_j) d(u/h)\right] \\
&= E\left[Z_j X_j' \int G'(v) f_{U|Z,X}(-hv|Z_j, X_j) dv\right] \\
&= E\big[Z_j X_j' f_{U|Z,X}(0|Z_j, X_j)\big] + O(h^r),
\end{aligned}$$

as desired. □