Econometric Theory, **33**, 2017, 105–157.
doi:10.1017/S0266466615000407

SMOOTHED ESTIMATING EQUATIONS FOR INSTRUMENTAL VARIABLES QUANTILE REGRESSION

DAVID M. KAPLAN
University of Missouri

YIXIAO SUN
University of California, San Diego

The moment conditions or estimating equations for instrumental variables quantile regression involve the discontinuous indicator function. We instead use smoothed estimating equations (SEE), with bandwidth h. We show that the mean squared error (MSE) of the vector of the SEE is minimized for some $h > 0$, leading to smaller asymptotic MSE of the estimating equations and associated parameter estimators. The same MSE-optimal h also minimizes the higher-order type I error of a SEE-based χ^2 test and increases size-adjusted power in large samples. Computation of the SEE estimator also becomes simpler and more reliable, especially with (more) endogenous regressors. Monte Carlo simulations demonstrate all of these superior properties in finite samples, and we apply our estimator to JTPA data. Smoothing the estimating equations is not just a technical operation for establishing Edgeworth expansions and bootstrap refinements; it also brings the real benefits of having more precise estimators and more powerful tests.

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 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

Thanks to Victor Chernozhukov (co-editor) and an anonymous referee for insightful comments and references, and thanks to Peter C. B. Phillips (editor) for additional editorial help. Thanks to Xiaohong Chen, Brendan Beare, Andres Santos, and active seminar and conference participants for insightful questions and comments. Address correspondence to David M. Kaplan, Department of Economics, University of Missouri, 118 Professional Bldg, 909 University Ave, Columbia, MO 65211-6040; e-mail: kaplandm@missouri.edu.

© Cambridge University Press 2016

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

that requires smoothness. This is especially attractive in IV-QR where simplex methods for the usual QR are not applicable. In fact, the SEE approach has been used in Chen and Pouzo (2009, 2012) for computing their nonparametric sieve estimators in the presence of nonsmooth moments or generalized residuals. However, a rigorous investigation is currently lacking. Our paper can be regarded as a first step towards justifying the SEE approach in nonparametric settings. Relatedly, Fan and Liao (2014, Sect. 7.1) have employed the same strategy of smoothing the indicator function to reduce the computational burden of their focused GMM approach. Second, from a technical point of view, smoothing the estimating equations enables us to establish high-order properties of the estimator. This motivates Horowitz (1998), for instance, 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 to smoothing the estimating equations. This point has not been recognized and emphasized in the literature. For QR estimation and inference via empirical likelihood, Otsu (2008) and Whang (2006) also examine smoothed estimators. To the best of our knowledge, no paper has examined smoothing the estimating equations for the usual QR estimator, let alone IV-QR. 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 more powerful tests. We seem to be the first to establish these advantages.

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 control, 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.

There is a growing literature on IV-QR. For a recent review, see Chernozhukov and Hansen (2013). Our paper is built upon Chernozhukov and Hansen (2005), which establishes a structural framework for IV-QR and provides primitive identification conditions. Within this framework, Chernozhukov and Hansen (2006) and Chernozhukov, Hansen, and Jansson (2009) develop estimation and inference procedures under strong identification. For inference procedures that are robust to weak identification, see Chernozhukov and Hansen (2008) and Jun (2008), for example. IV-QR can also reduce bias for dynamic panel fixed effects estimation

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

as in Galvao (2011). None of these papers considers smoothing the IV-QR estimating equations; that idea (along with minimal first-order theory) seemingly first appeared in an unpublished draft by MaCurdy and Hong (1999), although the idea of smoothing the indicator function in general appears even earlier, as in Horowitz (1992) for the smoothed maximum score estimator. An alternative approach to overcome the computational obstacles in the presence of a nonsmooth objective function is to explore the asymptotic equivalence of the Bayesian and classical methods for regular models and use the MCMC approach to obtain the classical extremum estimator; see Chernozhukov and Hong (2003), whose Example 3 is IV-QR. As a complement, our approach deals with the computation problem in the classical framework directly.

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 and type II errors of a chi-square test, and the approximate MSE of the parameter estimator, respectively. Section 6 applies our estimator to JTPA data, and Section 7 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 $\mathbb{E}\big[Z_j\big(1\{U_j < 0\} - q\big)\big] = 0$ for instrument vector $Z_j \in \mathbb{R}^d$ and $1\{\cdot\}$ is the indicator function. 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:

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

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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 and define $\hat{\beta}$ to be the solution of the vector of smoothed estimating equations (SEE) $m_n(\hat{\beta}) = 0$, where[1]

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

Our approach is related to kernel-based nonparametric conditional quantile estimators. The moment condition there is $\mathbb{E}[1\{X = x\}(1\{Y < \beta\} - q)] = 0$. 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. 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.

If the model is overidentified with $\dim(Z_j) > \dim(X_j)$, we can use a $\dim(X_j) \times \dim(Z_j)$ matrix $\mathbb{W}$ to transform the original moment conditions $\mathbb{E}\left[Z_j\left(q - 1\left\{Y_j < X_j'\beta\right\}\right)\right] = 0$ into

$$\mathbb{E}\left[\tilde{Z}_j\left(q - 1\left\{Y_j < X_j'\beta\right\}\right)\right] = 0, \text{ for } \tilde{Z}_j = \mathbb{W}Z_j \in \mathbb{R}^{\dim(X_j)}. \tag{1}$$

Then we have an exactly identified model with transformed instrument vector $\tilde{Z}_j$, and our asymptotic analysis can be applied to (1).

By the theory of optimal estimating equations or efficient two-step GMM, the optimal $\mathbb{W}$ takes the following form:

$$\begin{aligned}\mathbb{W} &= \frac{\partial}{\partial\beta}\mathbb{E}\left[Z'\left(q - 1\left\{Y < X'\beta\right\}\right)\right]\Big|_{\beta=\beta_0} \operatorname{Var}\left[Z\left(q - 1\{Y < X'\beta_0\}\right)\right]^{-1} \\ &= \mathbb{E}\left[XZ' f_{U|Z,X}(0)\right]\left\{\mathbb{E}\left[ZZ'\sigma^2(Z)\right]\right\}^{-1},\end{aligned}$$

where $f_{U|Z,X}(0)$ is the conditional PDF of U evaluated at $U = 0$ given (Z, X) and $\sigma^2(Z) = \operatorname{Var}(1\{U < 0\} \mid Z)$. The standard two-step approach requires an initial estimator of β_0 and nonparametric estimators of $f_{U|Z,X}(0)$ and $\sigma^2(Z)$. The underlying nonparametric estimation error may outweigh the benefit of having an optimal weighting matrix. This is especially a concern when the dimensions of X and Z are large. The problem is similar to what Hwang and Sun (2015)

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

consider in a time series GMM framework where the optimal weighting matrix is estimated using a nonparametric HAC approach. Under the alternative and more accurate asymptotics that captures the estimation error of the weighting matrix, they show that the conventionally optimal two-step approach does not necessarily outperform a first-step approach that does not employ a nonparametric weighting matrix estimator. While we expect a similar qualitative message here, we leave a rigorous analysis to future research.

In practice, a simple procedure is to ignore $f_{U|Z,X}(0)$ and $\sigma^2(Z)$ (or assume that they are constants) and employ the following empirical weighting matrix,

$$\mathbb{W}_n = \left(\frac{1}{n}\sum_{j=1}^{n} X_j Z_j'\right)\left(\frac{1}{n}\sum_{j=1}^{n} Z_j Z_j'\right)^{-1}.$$

This choice of $\mathbb{W}_n$ is in the spirit of the influential work of Liang and Zeger (1986) who advocate the use of a working correlation matrix in constructing the weighting matrix. Given the above choice of $\mathbb{W}_n$, $\tilde{Z}_j$ is the least squares projection of X_j on Z_j. It is easy to show that with some notational changes our asymptotic results remain valid in this case.

An example of an overidentified model is the conditional moment model

$$\mathbb{E}\big[\big(1\{U_j < 0\} - q\big) \mid Z_j\big] = 0.$$

In this case, any measurable function of Z_j can be used as an instrument. As a result, the model could be overidentified. According to Chamberlain (1987) and Newey (1990), the optimal set of instruments in our setting is given by

$$\left[\frac{\partial}{\partial\beta}\mathbb{E}\Big(1\{Y_j - X_j'\beta < 0\} \mid Z_j\Big)\right]\Big|_{\beta=\beta_0}.$$

Let $F_{U|Z,X}(u \mid z, x)$ and $f_{U|Z,X}(u \mid z, x)$ be the conditional distribution function and density function of U given $(Z, X) = (z, x)$. Then under some regularity conditions,

$$\begin{aligned}
&\left[\frac{\partial}{\partial\beta}\mathbb{E}\Big(1\Big\{Y_j - X_j'\beta < 0\Big\} \mid Z_j\Big)\right]\Big|_{\beta=\beta_0} \\
&\quad= \left\{\frac{\partial}{\partial\beta}\mathbb{E}\Big[\mathbb{E}\Big(1\Big\{Y_j - X_j'\beta < 0\Big\} \mid Z_j, X_j\Big) \mid Z_j\Big]\right\}\Big|_{\beta=\beta_0} \\
&\quad= \mathbb{E}\left\{\left[\frac{\partial}{\partial\beta}F_{U_j|Z_j,X_j}\big(X_j(\beta-\beta_0) \mid Z_j, X_j\big)\right]\Big|_{\beta=\beta_0} \;\Big|\; Z_j\right\} \\
&\quad= \mathbb{E}\big[f_{U_j|Z_j,X_j}(0 \mid Z_j, X_j)X_j \mid Z_j\big].
\end{aligned}$$

The optimal instruments involve the conditional density $f_{U|Z,X}(u \mid z, x)$ and a conditional expectation. In principle, these objects can be estimated nonparametrically. However, the nonparametric estimation uncertainty can be very high, adversely affecting the reliability of inference. A simple and practical strategy[2]

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

is to construct the optimal instruments as the OLS projection of each X_j onto some sieve basis functions $\Phi^K(Z_j) \equiv [\Phi_1(Z_j), \ldots, \Phi_K(Z_j)]'$, leading to

$$\tilde{Z}_j = \left[\frac{1}{n}\sum_{j=1}^{n} X_j \Phi^K(Z_j)'\right]\left[\frac{1}{n}\sum_{j=1}^{n} \Phi^K(Z_j)\Phi^K(Z_j)'\right]^{-1} \Phi^K(Z_j) \in \mathbb{R}^{\dim(X_j)}$$

as the instruments. Here $\{\Phi_i(\cdot)\}$ are the basis functions such as power functions. Since the dimension of $\tilde{Z}_j$ is the same as the dimension of X_j, our asymptotic analysis can be applied for any fixed value of K.[3]

2.2. Comparison with other estimators

Smoothed criterion function. For the special case $Z_j = X_j$, we compare the SEE with the estimating equations 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

$$\begin{aligned}
0 &= \left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} n^{-1}\sum_{i=1}^{n}\left[q - G\left(\frac{X_i'\beta - Y_i}{h}\right)\right](Y_i - X_i'\beta) \\
&= n^{-1}\sum_{i=1}^{n}\left[-qX_i - G'(-U_i/h)(X_i/h)Y_i + G'(-U_i/h)(X_i/h)X_i'\beta_0 + G(-U_i/h)X_i\right] \\
&= n^{-1}\sum_{i=1}^{n} X_i[G(-U_i/h) - q] + n^{-1}\sum_{i=1}^{n} G'(-U_i/h)\left[(X_i/h)X_i'\beta_0 - (X_i/h)Y_i\right] \\
&= n^{-1}\sum_{i=1}^{n} X_i[G(-U_i/h) - q] + n^{-1}\sum_{i=1}^{n}(1/h)G'(-U_i/h)[-X_iU_i]. \qquad (2)
\end{aligned}$$

The first term agrees with our proposed SEE. Technically, it should be easier to establish high-order results for our SEE estimator since it has one fewer term. Later we show that the absolute bias of our SEE estimator is smaller, too. Another subtle point is that our SEE requires only the estimating equation $\mathbb{E}[X_j(1\{U_j < 0\} - q)] = 0$, whereas Horowitz (1998) has to impose an additional condition to ensure that the second term in the FOC is approximately mean zero.

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 $\hat{\beta}_\infty$ is defined by

$$\begin{aligned}
0 &= \sum_{i=1}^{n} Z_i\left[G\left(\frac{X_i'\hat{\beta}_\infty - Y_i}{h}\right) - q\right] \\
&\doteq \sum_{i=1}^{n} Z_i\left[\left(0.5 + (105/64)\frac{X_i'\hat{\beta}_\infty - Y_i}{h}\right) - q\right]
\end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$$= (105/64h)Z'X\hat{\beta}_\infty - (105/64h)Z'Y + (0.5-q)Z'\mathbf{1}_{n,1}$$
$$= (105/64h)Z'X\hat{\beta}_\infty - (105/64h)Z'Y + (0.5-q)Z'(Xe_1),$$

where $e_1 = (1, 0, \ldots, 0)'$ is $d \times 1$, $\mathbf{1}_{n,1} = (1, 1, \ldots, 1)'$ 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 $Xe_1 = \mathbf{1}_{n,1}$. It then follows that

$$\hat{\beta}_\infty = \hat{\beta}_{IV} + ((64h/105)(q-0.5), 0, \ldots, 0)'.$$

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.[4]

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 the median and the mean for different error distributions. Our estimator with a data-driven bandwidth can harness the advantages of both, without requiring the practitioner to make guesses about the unknown error distribution.

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

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

If $G'(u) = 1\{-1 \leq u \leq 1\}/2$ (the uniform kernel), then $H(u) \equiv 2G(u) - 1 = u$ for $u \in [-1, 1]$, $H(u) = 1$ for $u > 1$, and $H(u) = -1$ for $u < -1$. The SEE is then $0 = \sum_{i=1}^{n} \psi(Y_i; \beta)$ with $\psi(Y_i; \beta) = H((\beta - Y_i)/h)$. This produces the Winsorized mean estimator of the type in Huber (1964, Example (iii), p. 79).[5]

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).

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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. An advantage of using EEs directly is that inference can be made robust to the strength of identification. Our focus on the EEs is also in the same spirit of the large literature on optimal estimating equations. For the historical developments of EEs and their applications in econometrics, see Bera, Bilias, and Simlai (2006). The MSE of the SEE is also related to the estimator MSE and inference properties both intuitively and (as we will show) theoretically. Such results 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. We write $f_{U|Z}(\cdot \mid z)$ and $F_{U|Z}(\cdot \mid z)$ as the conditional PDF and CDF of U given $Z = z$. We define $f_{U|Z,X}(\cdot \mid z, x)$ and $F_{U|Z,X}(\cdot \mid z, x)$ similarly.

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 such that $\mathbb{E}\big[Z_j\big(1\{U_j < 0\} - q\big)\big] = 0$.

Assumption 2. (i) Z_j has bounded support. (ii) $\mathbb{E}\big(Z_j Z_j'\big)$ is nonsingular.

Assumption 3. (i) $P(U_j < 0 \mid Z_j = z) = q$ for almost all $z \in \mathcal{Z}$, the support of Z. (ii) For all u in a neighborhood of zero and almost all $z \in \mathcal{Z}$, $f_{U|Z}(u \mid z)$ 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 $\left|f_{U|Z}^{(s)}(u \mid z)\right| \leq C(z)$ for $s = 0, 2, \ldots, r$, almost all $z \in \mathcal{Z}$ and u in a neighborhood of zero, and $\mathbb{E}\big[C(Z)\|Z\|^2\big] < \infty$.

Assumption 4. (i) $G(v)$ is a bounded function satisfying $G(v) = 0$ for $v \leq -1$, $G(v) = 1$ for $v \geq 1$, and $1 - \int_{-1}^{1} G^2(u)du > 0$. (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 $k = 1, 2, \ldots, r-1$, $\int_{-1}^{1} \left|v^r G'(v)\right| dv < \infty$, 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 $-1 = a_0 < a_1 < \cdots < a_{\tilde{L}} = 1$ for some finite $\tilde{L}$ such that $\theta'\tilde{G}(u)$ is either strictly positive or strictly negative on the intervals (a_{i-1}, a_i) for $i = 1, 2, \ldots, \tilde{L}$.

Assumption 5. $h \propto n^{-\kappa}$ for $1/(2r) < \kappa < 1$.

Assumption 6. $\beta = \beta_0$ uniquely solves $\mathbb{E}\big[Z_j\big(q - 1\{Y_j < X_j'\beta\}\big)\big] = 0$ over $\beta \in \mathcal{B}$.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Assumption 7. (i) $f_{U|Z,X}(u \mid z, x)$ is r times continuously differentiable in u in a neighborhood of zero for almost all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ for $r > 2$. (ii) $\Sigma_{ZX} \equiv \mathbb{E}\Big[Z_j X_j' f_{U|Z,X}(0 \mid Z_j, X_j)\Big]$ is nonsingular.

Assumption 1 describes the sampling process. Assumption 2 is analogous to Assumption 3 in both Horowitz (1998) and Whang (2006). As discussed in these two papers, the boundedness assumption for Z_j, which is a technical condition, is made only for convenience and can be dropped at the cost of more complicated proofs.

Assumption 3(i) allows us to use the law of iterated expectations to simplify the asymptotic variance. Our qualitative conclusions do not rely on this assumption. Assumption 3(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 Section 5.

Assumptions 4(i–ii) are analogous to the standard high-order kernel conditions in the kernel smoothing literature. The integral condition in (i) ensures that smoothing reduces (rather than increases) variance. Note that

$$\begin{aligned}
1 - \int_{-1}^{1} G^2(u)du &= 2\int_{-1}^{1} uG(u)G'(u)du \\
&= 2\int_{0}^{1} uG(u)G'(u)du + 2\int_{-1}^{0} uG(u)G'(u)du \\
&= 2\int_{0}^{1} uG(u)G'(u)du - 2\int_{0}^{1} vG(-v)G'(-v)dv \\
&= 2\int_{0}^{1} uG'(u)[G(u) - G(-u)]du,
\end{aligned}$$

using the evenness of $G'(u)$. When $r = 2$, we can use any $G(u)$ such that $G'(u)$ is a symmetric PDF on $[-1, 1]$. In this case, $1 - \int_{-1}^{1} G^2(u)du > 0$ holds automatically. 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 this quantity 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:

$$G(u) = \begin{cases} 0, & u \leq -1 \\ 0.5 + \frac{105}{64}\left(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\right), & u \in [-1, 1] \\ 1 & u \geq 1 \end{cases} \tag{3}$$

The range of the function is outside $[0, 1]$. Simple calculations show that $1 - \int_{-1}^{1} G^2(u)du > 0$.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

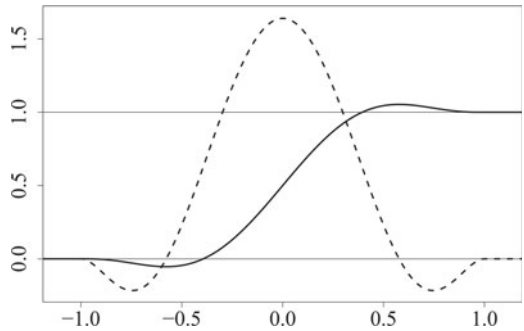


FIGURE 1. Graph of $G(u) = 0.5 + \frac{105}{64}\left(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\right)$ (solid line) and its derivative (broken).

Assumption 4(iii) is needed for the Edgeworth expansion. As Horowitz (1998) and Whang (2006) discuss, Assumption 4(iii) is a technical assumption that (along with Assumption 5) leads to a form of Cramér's condition, which is needed to justify the Edgeworth expansion used in Section 4. Any $G(u)$ constructed by integrating polynomial kernels in Müller (1984) satisfies Assumption 4(iii). In fact, $G(u)$ in (3) is obtained by integrating a fourth-order kernel given in Table 1 of Müller (1984). Assumption 5 ensures that the bias of the SEE is of smaller order than its variance. It is needed for the asymptotic normality of the SEE as well as the Edgeworth expansion.

Assumption 6 is an identification assumption. See Theorem 2 of Chernozhukov and Hansen (2006) for more primitive conditions. It ensures the consistency of the SEE estimator. Assumption 7 is necessary for the $\sqrt{n}$-consistency and asymptotic normality of the SEE estimator.

Define

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

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 2(i), 3, and 4(i–ii) hold. Then*

$$\mathbb{E}(W_j) = \frac{(-h)^r}{r!}\left[\int_{-1}^{1} G'(v)v^r dv\right]\mathbb{E}\Big[f_{U|Z}^{(r-1)}\big(0 \mid Z_j\big)Z_j\Big] + o\big(h^r\big), \tag{4}$$

$$\mathbb{E}(W_j'W_j) = q(1-q)\mathbb{E}\Big(Z_j'Z_j\Big) - h\left[1 - \int_{-1}^{1} G^2(u)du\right]\mathbb{E}\Big[f_{U|Z}\big(0 \mid Z_j\big)Z_j'Z_j\Big] + O\Big(h^2\Big), \tag{5}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

If additionally Assumptions 1 and 5 hold, then

$$m_n \xrightarrow{d} N(0, V), \quad V \equiv \lim_{n\to\infty} \mathbb{E}\Big\{\big[W_j - \mathbb{E}(W_j)\big]\big[W_j - \mathbb{E}(W_j)\big]'\Big\} = q(1-q)\mathbb{E}\Big(Z_j Z_j'\Big).$$

Compared with the EE derived from smoothing the criterion function as in Horowitz (1998), our SEE has smaller bias and variance, and these differences affect the bias and variance of the parameter estimator. The former approach only applies to exogenous QR with $Z_j = X_j$. The EE derived from smoothing the criterion function in (2) for $Z_j = X_j$ can be written

$$0 = n^{-1}\sum_{j=1}^{n} W_j, \quad W_j \equiv X_j\big[G\big(-U_j/h\big) - q\big] + (1/h)G'\big(-U_j/h\big)\big(-X_j U_j\big). \quad \textbf{(6)}$$

Consequently, as calculated in the appendix,

$$\mathbb{E}\big(W_j\big) = (r+1)\frac{(-h)^r}{r!}\left[\int G'(v)v^r dv\right]\mathbb{E}\Big[f_{U|Z}^{(r-1)}\big(0 \mid Z_j\big)Z_j\Big] + o\big(h^r\big), \quad \textbf{(7)}$$

$$\mathbb{E}\Big(W_j W_j'\Big) = q(1-q)\mathbb{E}\Big(X_j X_j'\Big) + h\int_{-1}^{1}\big[G'(v)v\big]^2 dv\,\mathbb{E}\Big[f_{U|X}\big(0 \mid X_j\big)X_j X_j'\Big] + O\Big(h^2\Big), \quad \textbf{(8)}$$

$$\mathbb{E}\left[\frac{\partial}{\partial\beta'}n^{-1/2}m_n(\beta_0)\right] = \mathbb{E}\Big[f_{U|X}\big(0 \mid X_j\big)X_j X_j'\Big] - h\mathbb{E}\Big[f_{U|X}'\big(0 \mid X_j\big)X_j X_j'\Big] + O\Big(h^2\Big). \quad \textbf{(9)}$$

The dominating term of the bias of our SEE in (4) is $r+1$ times smaller in absolute value than that of the EE derived from a smoothed criterion function in (7). A larger bias can lead to less accurate confidence regions if the same variance estimator is used. Additionally, the smoothed criterion function analog of $\mathbb{E}\big(W_j W_j'\big)$ in (8) has a positive $O(h)$ term instead of the negative $O(h)$ term for SEE. The connection between these terms and the estimator's asymptotic mean squared error (AMSE) is shown in Section 5 to rely on the inverse of the matrix in equation (9). Here, though, the sign of the $O(h)$ term is indeterminant since it depends on a PDF derivative. (A negative $O(h)$ term implies higher AMSE since this matrix is inverted in the AMSE expression, and positive implies lower.) If $U = 0$ is a mode of the conditional (on X) distribution, then the $O(h)$ term is zero and the AMSE comparison is driven by $\mathbb{E}\big(W_j\big)$ and $\mathbb{E}\big(W_j W_j'\big)$.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Since SEE yields smaller $\mathbb{E}(W_j W_j')$ and smaller absolute $\mathbb{E}(W_j)$, it will have smaller estimator AMSE in such cases. Simulation results in Section 7 add evidence that the SEE estimator usually has smaller MSE in practice.

The first-order asymptotic variance V is the same as the asymptotic variance of

$$n^{-1/2}\sum_{j=1}^{n} Z_j\big(1\{U_j < 0\} - q\big),$$

the scaled EE of the unsmoothed IV-QR. The effect of smoothing to reduce variance is captured by the term of order h, where $1 - \int_{-1}^{1} G^2(u)du > 0$ by Assumption 4(i). This reduction in variance 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.

Define the MSE of the SEE to be $\mathbb{E}\big(m_n' V^{-1} m_n\big)$. Building upon (4) and (5), and using $W_i \perp\!\!\!\perp W_j$ for $i \neq j$, we have:

$$\begin{aligned}
&\mathbb{E}\Big(m_n' V^{-1} m_n\Big) \\
&\quad = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(W_j' V^{-1} W_j\Big) + \frac{1}{n}\sum_{j=1}^{n}\sum_{i\neq j}\mathbb{E}\Big(W_i' V^{-1} W_j\Big) \\
&\quad = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(W_j' V^{-1} W_j\Big) + \frac{1}{n}n(n-1)\mathbb{E}(W_j')V^{-1}\mathbb{E}(W_j) \\
&\quad = q(1-q)\mathbb{E}\Big(Z_j' V^{-1} Z_j\Big) + nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big] + o\Big(h + nh^{2r}\Big), \\
&\quad = d + nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big] + o\Big(h + nh^{2r}\Big), \qquad \textbf{(10)}
\end{aligned}$$

where

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

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

$$h_{\mathrm{SEE}}^{*} \equiv \arg\min_{h} nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]. \qquad \textbf{(11)}$$

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

$$h^*_{SEE} = \left(\frac{\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]}{\mathbb{E}(B)'\mathbb{E}(B)} \frac{1}{2nr} \right)^{\frac{1}{2r-1}}.$$

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

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

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 Section 7.

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

For the unsmoothed IV-QR, let

$$\tilde{m}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j \left(1\left\{Y_j \le X_j'\beta\right\} - q\right),$$

then the MSE of the estimating equations is $\mathbb{E}\left(\tilde{m}_n' V^{-1} \tilde{m}_n\right) = d$. Comparing this to the MSE of the SEE given in (10), we find that the SEE has a smaller MSE when $h = h^*_{\mathrm{SEE}}$ because

$$n(h^*_{\mathrm{SEE}})^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h^*_{\mathrm{SEE}}\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right] = -h^*_{\mathrm{SEE}}\left(1 - \frac{1}{2r}\right)\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right] < 0.$$

In terms of MSE, it is advantageous to smooth the estimating equations. To the best of our knowledge, this point has never been discussed before in the literature.

4. TYPE I AND TYPE II ERRORS OF A CHI-SQUARE TEST

In this section, we explore the effect of smoothing on a chi-square 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 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. The reduced

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

MSE also makes it easier to distinguish the null hypothesis from some given alternative. This combination leads to improved size-adjusted power. As seen in our simulations, this is true especially for the IV case.

Using the results in Section 3 and under Assumption 5, we have

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

where we continue to use 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

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

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

$$\lim_{n\to\infty} P(S_n > c_\alpha) = \alpha.$$

Here $P \equiv P_{\beta_0}$ is the probability measure under the true model parameter β_0. We suppress the subscript β_0 when there is no confusion.

It is important to point out that the above result does not rely on the strong identification of β_0. It still holds if β_0 is weakly identified or even unidentified. This is an advantage of focusing on the estimating equations instead of the parameter estimator. When a direct inference method based on the asymptotic normality of $\hat{\beta}$ is used, we have to impose Assumptions 6 and 7.

4.1. Type I error and the associated optimal bandwidth

To more precisely measure the type I error $P(S_n > c_\alpha)$, we first develop a high-order stochastic expansion of S_n. Let $V_n \equiv \mathrm{Var}(m_n)$. Following the same calculation as in (10), we have

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

$$\Lambda_n = V^{1/2}\Big[I_d - h\mathbb{E}\big(AA'\big) + O\Big(h^2\Big)\Big]^{1/2} \tag{12}$$

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

$$\bar{W}_n^* \equiv \frac{1}{n}\sum_{j=1}^{n} W_j^* \text{ and } W_j^* = \Lambda_n^{-1} Z_j\big[G(-U_j/h) - q\big], \tag{13}$$

we can approximate the test statistic as $S_n = S_n^L + e_n$, where

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

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

The stochastic expansion above allows us to approximate the characteristic function of S_n with that of S_n^L. 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 1–5, we have*

$$P\Big(S_n^L < x\Big) = \mathcal{G}_d(x) - \mathcal{G}'_{d+2}(x)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\} + R_n,$$
$$P(S_n > c_\alpha) = \alpha + \mathcal{G}'_{d+2}(c_\alpha)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\} + R_n,$$

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

From Theorem 3, an approximate measure of the type I error of the SEE-based chi-square test is

$$\alpha + \mathcal{G}'_{d+2}(c_\alpha)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\},$$

and an approximate measure of the coverage probability error (CPE) is[6]

$$\mathrm{CPE} = \mathcal{G}'_{d+2}(c_\alpha)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\},$$

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

Up to smaller-order terms, the term $nh^{2r}\mathbb{E}(B)'\mathbb{E}(B)$ characterizes the bias effect from smoothing. The bias increases type I error and reduces coverage probability. The term $h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]$ characterizes the variance effect from smoothing. The variance reduction decreases type I error and increases coverage probability. The type I error is α up to order $O\big(h + nh^{2r}\big)$. There exists some $h > 0$ that makes bias and variance effects cancel, leaving type I error equal to α up to smaller-order terms in R_n.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Note that $nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\text{tr}\left[\mathbb{E}\left(AA'\right)\right]$ is identical to the high-order term in the asymptotic MSE of the SEE in (10). The h^*_{CPE} that minimizes type I error is the same as h^*_{SEE}.

PROPOSITION 4. *Let Assumptions 1–5 hold. The bandwidth that minimizes the approximate type I error of the chi-square test based on the test statistic S_n is*

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

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

When $h = h^*_{\text{CPE}}$,

$$P(S_n > c_\alpha) = \alpha - C^+\mathcal{G}'_{d+2}(c_\alpha)h^*_{\text{CPE}}[1+o(1)],$$

where $C^+ = \left(1-\frac{1}{2r}\right)\text{tr}\left[\mathbb{E}\left(AA'\right)\right] > 0$. If instead we construct the test statistic based on the unsmoothed estimating equations, $\tilde{S}_n = \tilde{m}_n'\hat{V}^{-1}\tilde{m}_n$, then it can be shown that

$$P\left(\tilde{S}_n > c_\alpha\right) = \alpha + Cn^{-1/2}[1+o(1)]$$

for some constant C, which is in general not equal to zero. Given that $n^{-1/2} = o(h^*_{\text{CPE}})$ and $C^+ > 0$, we can expect the SEE-based chi-square test to have a smaller type I error in large samples.

4.2. Type II error and local asymptotic power

To obtain the local asymptotic power of the S_n test, we let the true parameter value be $\beta_n = \beta_0 - \delta/\sqrt{n}$, where β_0 is the parameter value that satisfies the null hypothesis H_0. In this case,

$$m_n(\beta_0) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Z_j\left[G\left(\frac{X_j'\delta/\sqrt{n} - U_j}{h}\right) - q\right].$$

In the proof of Theorem 5, we show that

$$\mathbb{E}[m_n(\beta_0)] = \Sigma_{ZX}\delta + \sqrt{n}(-h)^r V^{1/2}\mathbb{E}(B) + O\left(n^{-1/2} + \sqrt{n}h^{r+1}\right),$$

$$V_n = \text{Var}[m_n(\beta_0)] = V - hV^{1/2}\left[\mathbb{E}\left(AA'\right)\right]\left(V^{1/2}\right)' + O\left(n^{-1/2} + h^2\right).$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

THEOREM 5. *Let Assumptions 1–5 and 7(i) hold. Define* $\Delta \equiv \mathbb{E}\big[V_n^{-1/2} m_n(\beta_0)\big]$ *and* $\tilde{\delta} \equiv V^{-1/2}\Sigma_{ZX}\delta$. *We have*

$$\begin{aligned} P_{\beta_n}(S_n < x) &= \mathcal{G}_d\Big(x; \|\Delta\|^2\Big) + \mathcal{G}'_{d+2}\Big(x; \|\Delta\|^2\Big) h \mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big] \\ &\quad + \mathcal{G}'_{d+4}\Big(x; \|\Delta\|^2\Big) h\big[\Delta'\mathbb{E}\big(AA'\big)\Delta\big] + O\Big(h^2 + n^{-1/2}\Big) \\ &= \mathcal{G}_d\Big(x; \|\tilde{\delta}\|^2\Big) - \mathcal{G}'_{d+2}\Big(x; \|\tilde{\delta}\|^2\Big)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\} \\ &\quad + \Big[\mathcal{G}'_{d+4}\Big(x; \|\tilde{\delta}\|^2\Big) - \mathcal{G}'_{d+2}\Big(x; \|\tilde{\delta}\|^2\Big)\Big] h\big[\tilde{\delta}'\mathbb{E}\big(AA'\big)\tilde{\delta}\big] \\ &\quad - \mathcal{G}'_{d+2}\Big(x; \|\tilde{\delta}\|^2\Big) 2\tilde{\delta}'\sqrt{n}(-h)^r\mathbb{E}(B) + O\Big(h^2 + n^{-1/2}\Big), \end{aligned}$$

where $\mathcal{G}_d(x;\lambda)$ *is the CDF of the noncentral chi-square distribution with degrees of freedom* d *and noncentrality parameter* λ. *If we further assume that* $\tilde{\delta}$ *is uniformly distributed on the sphere* $\mathcal{S}_d(\tau) = \big\{\tilde{\delta} \in \mathbb{R}^d : \|\tilde{\delta}\| = \tau\big\}$, *then*

$$\begin{aligned} &\mathbb{E}_{\tilde{\delta}}\big[P_{\beta_n}(S_n > c_\alpha)\big] \\ &\quad = 1 - \mathcal{G}_d\Big(c_\alpha; \tau^2\Big) + \mathcal{G}'_{d+2}\Big(c_\alpha; \tau^2\Big)\Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\} \\ &\qquad - \Big[\mathcal{G}'_{d+4}\Big(c_\alpha; \tau^2\Big) - \mathcal{G}'_{d+2}\Big(c_\alpha; \tau^2\Big)\Big]\frac{\tau^2}{d} h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big] + O\Big(h^2 + n^{-1/2}\Big) \end{aligned}$$

where $\mathbb{E}_{\tilde{\delta}}$ *takes the average uniformly over the sphere* $\mathcal{S}_d(\tau)$.

When $\delta = 0$, which implies $\tau = 0$, the expansion in Theorem 5 reduces to that in Theorem 3.

When $h = h^*_{\mathrm{SEE}}$, it follows from Theorem 3 that

$$\begin{aligned} P_{\beta_0}(S_n > c_\alpha) &= 1 - \mathcal{G}_d(c_\alpha) - C^+\mathcal{G}'_{d+2}(c_\alpha)h^*_{\mathrm{SEE}} + o\big(h^*_{\mathrm{SEE}}\big) \\ &= \alpha - C^+\mathcal{G}'_{d+2}(c_\alpha)h^*_{\mathrm{SEE}} + o\big(h^*_{\mathrm{SEE}}\big). \end{aligned}$$

To remove the error in rejection probability of order h^*_{SEE}, we make a correction to the critical value c_α. Let c^*_α be a high-order corrected critical value such that $P_{\beta_0}\big(S_n > c^*_\alpha\big) = \alpha + o\big(h^*_{\mathrm{SEE}}\big)$. Simple calculation shows that

$$c^*_\alpha = c_\alpha - \frac{\mathcal{G}'_{d+2}(c_\alpha)}{\mathcal{G}'_d(c_\alpha)} C^+ h^*_{\mathrm{SEE}}$$

meets the requirement.

To approximate the size-adjusted power of the S_n test, we use c^*_α rather than c_α because c^*_α leads to a more accurate test in large samples. Using Theorem 5, we can prove the following corollary.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

COROLLARY 6. *Let the assumptions in Theorem 5 hold. Then for* $h = h^*_{SEE}$,

$$\mathbb{E}_{\tilde{\delta}}\left[P_{\beta_n}\left(S_n > c^*_\alpha\right)\right] = 1 - \mathcal{G}_d\left(c_\alpha; \tau^2\right) + Q_d\left(c_\alpha, \tau^2, r\right)\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]h^*_{SEE} + O\left(h^{*2}_{SEE} + n^{-1/2}\right), \tag{14}$$

where

$$Q_d\left(c_\alpha, \tau^2, r\right) = \left(1 - \frac{1}{2r}\right)\left[\mathcal{G}'_d\left(c_\alpha; \tau^2\right)\frac{\mathcal{G}'_{d+2}(c_\alpha)}{\mathcal{G}'_d(c_\alpha)} - \mathcal{G}'_{d+2}\left(c_\alpha; \tau^2\right)\right] - \frac{1}{d}\left[\mathcal{G}'_{d+4}\left(c_\alpha; \tau^2\right) - \mathcal{G}'_{d+2}\left(c_\alpha; \tau^2\right)\right]\tau^2.$$

In the asymptotic expansion of the local power function in (14), $1 - \mathcal{G}_d\left(c_\alpha; \tau^2\right)$ is the usual first-order power of a standard chi-square test. The next term of order $O(h^*_{\mathrm{SEE}})$ captures the effect of smoothing the estimating equations. To sign this effect, we plot the function $Q_d\left(c_\alpha, \tau^2, r\right)$ against τ^2 for $r = 2$, $\alpha = 10\%$, and different values of d in Figure 2. Figures for other values of r and α are qualitatively similar. The range of τ^2 considered in Figure 2 is relevant as the first-order local asymptotic power, i.e., $1 - \mathcal{G}_d\left(c_\alpha; \tau^2\right)$, increases from 10% to about 94%, 96%, 97%, and 99%, respectively for $d = 1, 2, 3, 4$. It is clear from this figure that $Q_d\left(c_\alpha, \tau^2, r\right) > 0$ for any $\tau^2 > 0$. This indicates that smoothing leads to a test with improved power. The power improvement increases with r. The smoother the conditional PDF of U in a neighborhood of the origin is, the larger the power improvement is.

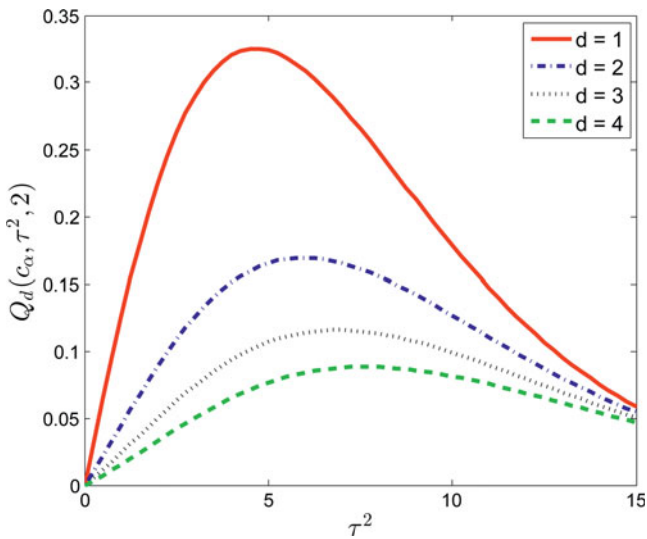


FIGURE 2. Plots of $Q_d\left(c_\alpha, \tau^2, 2\right)$ against τ^2 for different values of d with $\alpha = 10\%$.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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.

The SEE estimator $\hat{\beta}$ satisfies $m_n(\hat{\beta}) = 0$. In Lemma A.2 in the appendix, we show that

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = -\left\{\mathbb{E}\left[\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\right]\right\}^{-1} m_n + O_p\left(\frac{1}{\sqrt{nh}}\right) \tag{15}$$

and

$$\mathbb{E}\left[\frac{\partial}{\partial \beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)\right] = \mathbb{E}\left[Z_j X_j' f_{U|Z,X}(0 \mid Z_j, X_j)\right] + O\left(h^r\right). \tag{16}$$

Consequently, the approximate MSE (AMSE) of $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ is[7]

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

where

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

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 $h^r = o\left(nh^{2r}\right)$ and $1/\sqrt{nh} = o\left(nh^{2r}\right)$, so the terms of order $O_p(1/\sqrt{nh})$ in (15) and of order $O(h^r)$ in (16) 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 7. *Let Assumptions 1–4(i–ii), 6, and 7 hold. If* $nh^r \to \infty$ *and* $n^3 h^{4r+1} \to \infty$, *then the AMSE of* $\sqrt{n}(\hat{\beta} - \beta_0)$ *is*

$$\begin{aligned}&\Sigma_{ZX}^{-1} V^{1/2}\left[I_d + nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\left(AA'\right)\right]\left(V^{1/2}\right)'\left(\Sigma_{ZX}'\right)^{-1} \\ &\quad + O\left(h^r\right) + o\left(h + nh^{2r}\right).\end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

$$\Sigma_{ZX}^{-1} V^{1/2}\Big[n\big(h^*\big)^{2r}\mathbb{E}(B)\mathbb{E}(B') - h^*\mathbb{E}\big(AA'\big)\Big]\Big(V^{1/2}\Big)'\big(\Sigma_{ZX}'\big)^{-1}$$
$$\leq \Sigma_{ZX}^{-1} V^{1/2}\Big[nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)\Big]\Big(V^{1/2}\Big)'\big(\Sigma_{ZX}'\big)^{-1}$$

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

$$n\big(h^*\big)^{2r}\mathbb{E}(B)\mathbb{E}(B') - h^*\mathbb{E}\big(AA'\big) \leq nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big).$$

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

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

We include h as an argument of m_n^s to emphasize the dependence of m_n^s on h. The standardization can be motivated 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 $\mathbb{E}(m_n m_n')$ 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 $\mathbb{E}\frac{\partial}{\partial\beta'}m_n(\beta_0)$ to be as large as possible. To meet these requirements, we choose h to maximize

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

More specifically, h^* is optimal if

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

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

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

$$h^* = \arg\min_h u'\Big[nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)\Big]u \tag{17}$$

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 $e_1 = (1, 0, \ldots, 0)$,

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$e_2 = (0, 1, 0, \ldots, 0)$, 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

$$h^* = \arg\min_h \mathrm{tr}\Big[nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)\Big]$$
$$= \arg\min_h \Big\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\big[\mathbb{E}\big(AA'\big)\big]\Big\}.$$

In view of (11), $h^*_{\mathrm{SEE}} = h^*$ if h^* exists. Unfortunately, it is easy to show that no single h can minimize the objective function in (17) 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 $u'\big[nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)\big]u$ 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

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

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

$$h^*_c \geq \left(\frac{1}{\mathbb{E}(B)'[\mathbb{E}(AA')]^{-1}\mathbb{E}(B)}\frac{1}{2nr}\right)^{\frac{1}{2r-1}} > 0.$$

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.,

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

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

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

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

so all three selections coincide with each other. A special case of interest is when $Z = X$, nonconstant 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 $nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)$ with the corresponding orthonormal eigenvectors $\{\ell_k, k = 1,\ldots,d\}$. Then we have $nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big) = \sum_{k=1}^d \lambda_k \ell_k \ell_k'$ 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

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

That is, $h^*_{c_1,\ldots,c_d}$ minimizes a weighted sum of the eigenvalues of $nh^{2r}\mathbb{E}(B)\mathbb{E}(B') - h\mathbb{E}\big(AA'\big)$ 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. EMPIRICAL EXAMPLE: JTPA

We revisit the IV-QR analysis of Job Training Partnership Act (JTPA) data in Abadie, Angrist, and Imbens (2002), specifically their Table III.[8] They use 30-month earnings as the outcome, randomized offer of JTPA services as the instrument, and actual enrollment for services as the endogenous treatment variable. Of those offered services, only around 60 percent accepted, so self-selection into treatment is likely. Section 4 of Abadie et al. (2002) provides much more background and descriptive statistics.

We compare estimates from a variety of methods.[9] "AAI" is the original paper's estimator. AAI restricts X to have finite support (see condition (iii) in their

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Theorem 3.1), which is why all the regressors in their example are binary. Our fully automated plug-in estimator is "SEE ($\hat{h}$)." "CH" is Chernozhukov and Hansen (2006). Method "tiny h" uses $h = 400$ (compared with our plug-in values on the order of 10 000), while "huge h" uses $h = 5 \times 10^6$. 2SLS is the usual (mean) two-stage least squares estimator, put in the $q = 0.5$ column only for convenience of comparison.

Table 1 shows results from the sample of 5102 adult men, for a subset of the regressors used in the model. Not shown in the table are coefficient estimates

TABLE 1. IV-QR estimates of coefficients for certain regressors as in Table III of Abadie et al. (2002) for adult men

		Quantile				
Regressor	Method	0.15	0.25	0.50	0.75	0.85
Training	AAI	121	702	1544	3131	3378
Training	SEE ($\hat{h}$)	57	381	1080	2630	2744
Training	CH	−125	341	385	2557	3137
Training	tiny h	−129	500	381	2760	3114
Training	huge h	1579	1584	1593	1602	1607
Training	2SLS			1593		
HS or GED	AAI	714	1752	4024	5392	5954
HS or GED	SEE ($\hat{h}$)	812	1498	3598	6183	6753
HS or GED	CH	482	1396	3761	6127	6078
HS or GED	tiny h	463	1393	3767	6144	6085
HS or GED	huge h	4054	4062	4075	4088	4096
HS or GED	2SLS			4075		
Black	AAI	−171	−377	−2656	−4182	−3523
Black	SEE ($\hat{h}$)	−202	−546	−1954	−3273	−3653
Black	CH	−38	−109	−2083	−3233	−2934
Black	tiny h	−18	−139	−2121	−3337	−2884
Black	huge h	−2336	−2341	−2349	−2357	−2362
Black	2SLS			−2349		
Married	AAI	1564	3190	7683	9509	10 185
Married	SEE ($\hat{h}$)	1132	2357	7163	10 174	10 431
Married	CH	504	2396	7722	10 463	10 484
Married	tiny h	504	2358	7696	10 465	10 439
Married	huge h	6611	6624	6647	6670	6683
Married	2SLS			6647		
Constant	AAI	−134	1049	7689	14 901	22 412
Constant	SEE ($\hat{h}$)	−88	1268	7092	15 480	22 708
Constant	CH	242	1033	7516	14 352	22 518
Constant	tiny h	294	1000	7493	14 434	22 559
Constant	huge h	−1 157 554	−784 046	10 641	805 329	1 178 836
Constant	2SLS			10 641		

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

for dummies for Hispanic, working less than 13 weeks in the past year, five age groups, originally recommended service strategy, and whether earnings were from the second follow-up survey. CH is very close to "tiny h"; that is, simply using the smallest possible h with SEE provides a good approximation of the unsmoothed estimator in this case. Demonstrating our theoretical results in Section 2.2, "huge h" is very close to 2SLS for everything except the constant term for $q \neq 0.5$. The IVQR-SEE estimator using our plug-in bandwidth has some economically significant differences with the unsmoothed estimator. Focusing on the treatment variable ("Training"), the unsmoothed median effect estimate is below 400 (dollars), whereas SEE ($\hat{h}$) yields 1080, both of which are smaller than AAI's 1544 (AAI is the most positive at all quantiles). For the 0.15-quantile effect, the unsmoothed estimates are actually slightly negative, while SEE ($\hat{h}$) and AAI are slightly positive. For $q = 0.85$, though, the SEE ($\hat{h}$) estimate is smaller than the unsmoothed one, and the two are quite similar for $q = 0.25$ and $q = 0.75$; there is no systematic ordering.

Computationally, our code takes only one second total to calculate the plug-in bandwidths and coefficient estimates at all five quantiles. Using the fixed $h = 400$ or $h = 5 \times 10^6$, computation is immediate.

Table 2 shows estimates of the endogenous coefficient when various "continuous" control variables are added, specifically a quartic polynomial in age (replacing the age range dummies), baseline weekly hours worked, and baseline hourly wage.[10] The estimates do not change much; the biggest difference is for the unsmoothed estimate at the median. Our code again computes the plug-in bandwidth and SEE coefficient estimates at all five quantiles in one second. Using the small $h = 400$ bandwidth now takes nine seconds total (more iterations of `fsolve` are needed); $h = 5 \times 10^6$ still computes almost immediately.

7. SIMULATIONS

For our simulation study,[11] we use $G(u)$ given in (3) as in Horowitz (1998) and Whang (2006). This satisfies Assumption 4 with $r = 4$. Using (the integral of) an

TABLE 2. IV-QR estimates similar to Table 1, but replacing age dummies with a quartic polynomial in age and adding baseline measures of weekly hours worked and wage

		Quantile				
Regressor	Method	0.15	0.25	0.50	0.75	0.85
Training	SEE ($\hat{h}$)	74	398	1045	2748	2974
Training	CH	−20	451	911	2577	3415
Training	tiny h	−50	416	721	2706	3555
Training	huge h	1568	1573	1582	1590	1595
Training	2SLS			1582		

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Epanechnikov kernel with $r = 2$ also worked well in the cases we consider here, though never better than $r = 4$. 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.

We implement a plug-in version $\left(\hat{h}\right)$ of the infeasible $h^* \equiv h^*_{\text{SEE}}$. 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 $\hat{h}$. With larger n, a nonparametric kernel estimator may perform better, but nonparametric estimation of $f_U^{(r-1)}(0)$ will have high variance in smaller samples. Viewing the unsmoothed estimator as a reference point, potential regret (of using $\hat{h}$ instead of $h = 0$) is largest when $\hat{h}$ is too large, so 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: DGPs 3.1–3.6 in Section 7.3 have error distributions other than these four, and DGPs 1.3, 2.2, 3.3–3.6 are heteroskedastic, as are the JTPA-based simulations. For the infeasible h^*, if the PDF derivative in the denominator is zero, it is replaced by 0.01 to avoid $h^* = \infty$.

For the unsmoothed IV-QR estimator, we use code based on Chernozhukov and Hansen (2006) from the latter author's website. 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.

7.1. JTPA-based simulations

We use two DGPs based on the JTPA data examined in Section 6. The first DGP corresponds to the variables used in the original analysis in Abadie et al. (2002). For individual i, let Y_i be the scalar outcome (30-month earnings), X_i be the vector of exogenous regressors, D_i be the scalar endogenous training dummy, Z_i be the scalar instrument of randomized training offer, and $U_i \sim \text{Unif}(0, 1)$ be a scalar unobservable term. We draw X_i from the joint distribution estimated from the JTPA data. We randomize $Z_i = 1$ with probability 0.67 and zero otherwise. If $Z_i = 0$, then we set the endogenous training dummy $D_i = 0$ (ignoring that in reality, a few percent still got services). If $Z_i = 1$, we set $D_i = 1$ with a probability increasing in U_i. Specifically, $P(D_i = 1 \mid Z_i = 1, U_i = u) = \min\{1, u/0.75\}$, which roughly matches the $P(D_i = 1 \mid Z_i = 1) = 0.62$ in the data. This corresponds to a high degree of self-selection into treatment (and thus endogeneity). Then, $Y_i = X_i\beta_X + D_i\beta_D(U_i) + G^{-1}(U_i)$, where β_X is the IVQR-SEE $\hat{\beta}_X$ from the JTPA data (rounded to the nearest 500), the function $\beta_D(U_i) = 2000U_i$ matches $\hat{\beta}_D(0.5)$ and the increasing pattern of other $\hat{\beta}_D(q)$, and $G^{-1}(\cdot)$ is a recentered gamma distribution quantile function with parameters estimated to match

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

the distribution of residuals from the IVQR-SEE estimate with JTPA data. In each of 1000 simulation replications, we generate $n = 5102$ iid observations.

For the second DGP, we add a second endogenous regressor (and instrument) and four exogenous regressors, all with normal distributions. Including the intercept and two endogenous regressors, there are 20 regressors. The second instrument is $Z_{2i} \overset{iid}{\sim} N(0,1)$, and the second endogenous regressor is $D_{2i} = 0.8Z_{2i} + 0.2\Phi^{-1}(U_i)$. The coefficient on D_{2i} is 1000 at all quantiles. The new exogenous regressors are all standard normal and have coefficients of 500 at all quantiles. To make the asymptotic bias of 2SLS relatively more important, the sample size is increased to $n = 50\,000$.

Table 3 shows results for the first JTPA-based DGP, for three estimators of the endogenous coefficient: Chernozhukov and Hansen (2006); SEE with our data-dependent $\hat{h}$; and 2SLS. The first and third can be viewed as limits of IVQR-SEE estimators as $h \to 0$ and $h \to \infty$, respectively. We show median bias and "robust MSE," which is squared median bias plus the square of the interquartile range divided by 1.349, $\text{Bias}^2_{\text{median}} + (\text{IQR}/1.349)^2$. We report these "robust" versions of bias and MSE since the (mean) IV estimator does not even possess a first moment in finite samples (Kinal, 1980). We are unaware of an analogous result for IV-QR but remain wary of presenting bias and MSE results for IV-QR, too, especially since the IV estimator is the limit of the SEE IV-QR estimator as $h \to \infty$. At all quantiles, for all methods, the robust MSE is dominated by the IQR rather than bias. Consequently, even though the 2SLS median bias is quite large for $q = 0.15$, it has less than half the robust MSE of SEE $(\hat{h})$, which in turn has half the robust MSE of the unsmoothed estimator. With only a couple exceptions, this is the ordering among the three methods' robust MSE at all quantiles. Although the much larger bias of 2SLS than that of SEE $(\hat{h})$ or the unsmoothed estimator is expected, the smaller median bias of SEE $(\hat{h})$ than that of the unsmoothed estimator is surprising. However, the differences are not big, and they may be partly

TABLE 3. Simulation results for endogenous coefficient estimators with first JTPA-based DGP. "Robust MSE" is squared median-bias plus the square of the interquartile range divided by 1.349, $\text{Bias}^2_{\text{median}} + (\text{IQR}/1.349)^2$; it is shown in units of 10^5, so 7.8 means 7.8×10^5, for example. "Unsmoothed" is the estimator from Chernozhukov and Hansen (2006).

	Robust MSE / 10^5			Median Bias		
q	Unsmoothed	SEE $(\hat{h})$	2SLS	Unsmoothed	SEE $(\hat{h})$	2SLS
0.15	78.2	43.4	18.2	−237.6	8.7	1040.6
0.25	30.5	18.9	14.4	−122.2	16.3	840.6
0.50	9.7	7.8	8.5	24.1	−8.5	340.6
0.75	7.5	7.7	7.6	−5.8	−48.1	−159.4
0.85	11.7	9.4	8.6	49.9	−17.7	−359.4

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

due to the much larger variance of the unsmoothed estimator inflating the simulation error in the simulated median bias, especially for $q = 0.15$. The bigger difference is the reduction in variance from smoothing.

Table 4 shows results from the second JTPA-based DGP. The first estimator is now a nearly-unsmoothed SEE estimator instead of the unsmoothed Chernozhukov and Hansen (2006) estimator. Although in principle Chernozhukov and Hansen (2006) can be used with multiple endogenous coefficients, the provided code allows only one, and Tables 1 and 2 show that SEE with $h = 400$ produces very similar results in the JTPA data. For the binary endogenous regressor's coefficient, the 2SLS estimator now has the largest robust MSE since the larger sample size reduces the variance of all three estimators but does not reduce the 2SLS median bias (since it has first-order asymptotic bias). The plug-in bandwidth yields smaller robust MSE than the nearly-unsmoothed $h = 400$ at four of five quantiles. At the median, for example, compared with $h = 400$, $\hat{h}$ slightly increases the median bias but greatly reduces the dispersion, so the net effect is to reduce robust MSE. This is consistent with the theoretical results. For the continuous endogenous regressor's coefficient, the same pattern holds for the nearly-unsmoothed and $\hat{h}$-smoothed estimators. Since this coefficient is constant across quantiles, the 2SLS estimator is consistent and very similar to the SEE estimators with $q = 0.5$.

7.2. Comparison of SEE and smoothed criterion function

For exogenous QR, smoothing the criterion function (SCF) is a different approach, as discussed. The following simulations compare the MSE of our SEE estimator with that of the SCF estimator. All DGPs have $n = 50$, $X_i \overset{iid}{\sim} \text{Unif}(1, 5)$,

TABLE 4. Simulation results for endogenous coefficient estimators with second JTPA-based DGP

	Robust MSE			Median Bias		
q	$h = 400$	SEE $(\hat{h})$	2SLS	$h = 400$	SEE $(\hat{h})$	2SLS
	Estimators of binary endogenous regressor's coefficient					
0.15	780624	539542	1071377	−35.7	10.6	993.6
0.25	302562	227508	713952	−18.5	17.9	793.6
0.50	101433	96350	170390	−14.9	−22.0	293.6
0.75	85845	90785	126828	−9.8	−22.5	−206.4
0.85	147525	119810	249404	−15.7	−17.4	−406.4
	Estimators of continuous endogenous regressor's coefficient					
0.15	9360	7593	11434	−3.3	−5.0	−7.0
0.25	10641	9469	11434	−3.4	−3.5	−7.0
0.50	13991	12426	11434	−5.7	−9.8	−7.0
0.75	28114	25489	11434	−12.3	−17.9	−7.0
0.85	43890	37507	11434	−17.2	−17.8	−7.0

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$U_i \overset{iid}{\sim} N(0,1)$, $X_i \perp\!\!\!\perp U_i$, and $Y_i = 1 + X_i + \sigma(X_i)\big(U_i - \Phi^{-1}(q)\big)$. DGP 1 has $q = 0.5$ and $\sigma(X_i) = 5$. DGP 2 has $q = 0.25$ and $\sigma(X_i) = 1 + X_i$. DGP 3 has $q = 0.75$ and $\sigma(X_i) = 1 + X_i$. In addition to using our plug-in $\hat{h}$, we also compute the estimators for a much smaller bandwidth in each DGP: $h = 1$, $h = 0.8$, and $h = 0.8$, respectively. Each simulation ran 1000 replications. We compare only the slope coefficient estimators.

Table 5 shows MSE and bias for the SEE and SCF estimators, for our plug-in $\hat{h}$ as well as the small, fixed h mentioned above. The SCF estimator can have slightly lower MSE, as in the third DGP ($q = 0.75$ with heteroskedasticity), but the SEE estimator has more substantially lower MSE in more DGPs, including the homoskedastic conditional median DGP. The differences are quite small with the small h, as expected. Deriving and implementing an MSE-optimal bandwidth for the SCF estimator could shrink the differences, but based on these simulations and the theoretical comparison in Section 2, such an effort seems unlikely to yield improvement over the SEE estimator.

7.3. Additional simulations

We tried additional data generating processes (DGPs). The first three DGPs are for exogenous QR, taken directly from Horowitz (1998). In each case, $q = 0.5$, $Y_i = X_i'\beta_0 + U_i$, $\beta_0 = (1,1)'$, $X_i = (1, x_i)'$ with $x_i \overset{iid}{\sim}$ Uniform$(1,5)$, and $n = 50$. In DGP 1.1, the U_i are sampled iid from a t_3 distribution scaled to have variance two. In DGP 1.2, the U_i are iid from a type I extreme value distribution again scaled and centered to have median zero and variance two. In DGP 1.3, $U_i = (1 + x_i)V_i/4$ where $V_i \overset{iid}{\sim} N(0,1)$.

DGPs 2.1, 2.2, and 3.1–3.6 are shown in the working paper version; they include variants of the Horowitz (1998) DGPs with $q \neq 0.5$, different error distributions, and another regressor.

DGPs 4.1–4.3 have endogeneity. DGP 4.1 has $q = 0.5$, $n = 20$, and $\beta_0 = (0,1)'$. It uses the reduced form equations in Cattaneo, Crump, and Jansson (2012, eqn. 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 $(v_{1i}, v_{2i})' = \big(\tilde{v}_{1i}, \sqrt{1-\rho^2}\tilde{v}_{2i} + \rho\tilde{v}_{1i}\big)'$. DGP 4.2 is similar to DGP 4.1 but with $(\tilde{v}_{1i}, \tilde{v}_{2i})'$ iid Cauchy, $n = 250$,

TABLE 5. Simulation results comparing SEE and SCF exogenous QR estimators

	MSE				Bias			
	Plug-in $\hat{h}$		small h		Plug-in $\hat{h}$		small h	
DGP	SEE	SCF	SEE	SCF	SEE	SCF	SEE	SCF
1	0.423	0.533	0.554	0.560	−0.011	−0.013	−0.011	−0.009
2	0.342	0.433	0.424	0.430	0.092	−0.025	0.012	0.012
3	0.146	0.124	0.127	0.121	−0.292	−0.245	−0.250	−0.232

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

and $\beta_0 = \left(0, \left[\rho - \sqrt{1-\rho^2}\right]^{-1}\right)'$. DGP 4.3 is the same as DGP 4.1 but with $q = 0.35$ (and consequent re-centering of the error term) and $n = 30$.

We compare MSE for our SEE estimator using the plug-in $\hat{h}$ and estimators using different (fixed) values of h. We include $h = 0$ by using unsmoothed QR or the method in Chernozhukov and Hansen (2006) for the endogenous DGPs. We also include $h = \infty$ (although not in graphs) by using the usual IV estimator. For the endogenous DGPs, we consider both MSE and the "robust MSE" defined in Section 7.1 as $\text{Bias}^2_{\text{median}} + (\text{IQR}/1.349)^2$.

For "size-adjusted" power (SAP) of a test with nominal size α, the critical value is picked as the $(1-\alpha)$-quantile of the simulated 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 size-adjusted power figures, the vertical axis in the size-adjusted power figures shows the simulated rejection probability. The horizontal 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 $\mathbb{R}^d$, where $\beta \in \mathbb{R}^d$. 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 horizontal axis. The left vertical axis shows the MSE values for the intercept parameter; the right vertical axis shows the MSE for slope parameter(s); and the horizontal axis shows a log transformation of the bandwidth, $\log_{10}(1+h)$.

Our plug-in bandwidth is quite stable. 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 corresponds to a very small impact on MSE; note the log transformation in the x-axis in the MSE graphs.

In DGPs 1.1–1.3, SEE $\left(\hat{\hat{h}}\right)$ has smaller MSE than either the unsmoothed estimator or OLS, for both the intercept and slope coefficients. Figure 3 shows MSE for DGPs 1.1 and 1.3. It shows that the MSE of SEE $\left(\hat{\hat{h}}\right)$ is very close to that of the best estimator with a fixed h. In principle, a data-dependent $\hat{h}$ can attain MSE even lower than any fixed h. SAP for SEE $\left(\hat{\hat{h}}\right)$ is similar to that with $h = 0$; see Figure 4 for DGPs 1.1 and 1.3.

Figures 5 and 6 show MSE and "robust MSE" for two DGPs with endogeneity. Graphs for the other endogenous DGP (4.1) are similar to those for the slope estimator in DGP 4.3 but with larger MSE; they may be found in the working paper. The MSE graph for DGP 4.2 is not as informative since it is sensitive to

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

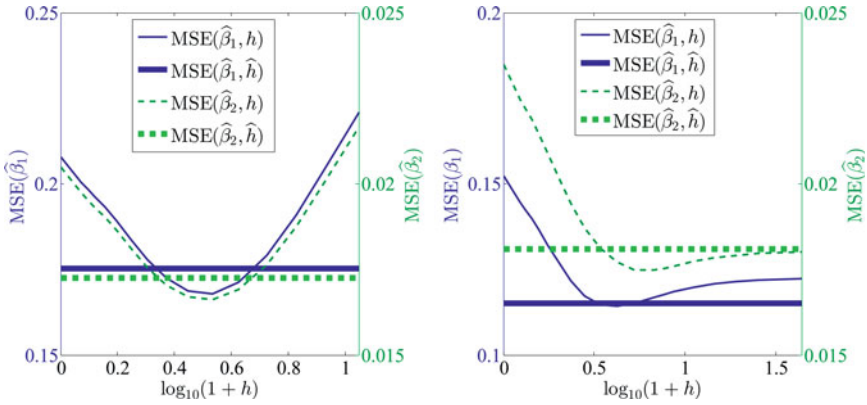


FIGURE 3. MSE for DGPs 1.1 (left) and 1.3 (right).

very large outliers that occur in only a few replications. However, as shown, the MSE for SEE $(\hat{h})$ is still better than that for the unsmoothed IV-QR estimator, and it is nearly the same as the MSE for the mean IV estimator (not shown: 1.1×10^6 for β_1, 2.1×10^5 for β_2). For robust MSE, SEE $(\hat{h})$ is again always better than the unsmoothed estimator. For DGP 4.3 with normal errors and $q = 0.35$, it is similar to the IV estimator, slightly worse for the slope coefficient and slightly better for the intercept, as expected. Also as expected, for DGP 4.2 with Cauchy errors, SEE $(\hat{h})$ is orders of magnitude better than the mean IV estimator. Overall, 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 (mean regression is more efficient), in which $\hat{h}$ is much better than $h = 0$ but not quite large enough to match $h = \infty$.

Figure 7 shows SAP for DGPs 4.1 and 4.3. The gain from smoothing is more substantial than in the exogenous DGPs, close to 10 percentage points for a range

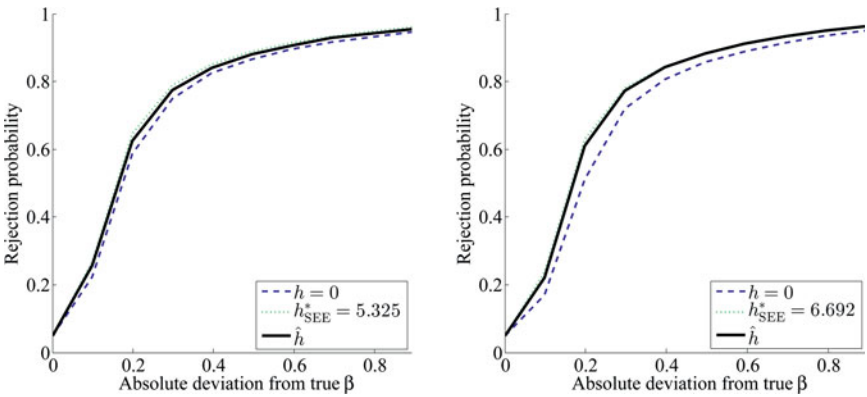


FIGURE 4. Size-adjusted power for DGPs 1.1 (left) and 1.3 (right).

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

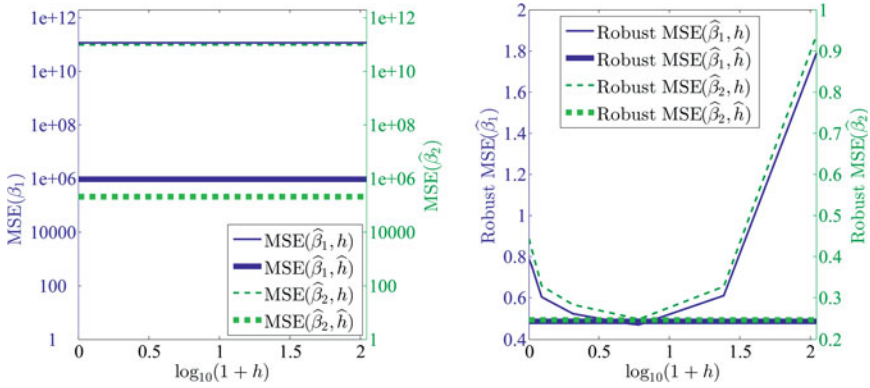


FIGURE 5. For DGP 4.2, MSE (left) and "robust MSE" (right): squared median-bias plus the square of the interquartile range divided by 1.349, $\text{Bias}^2_{\text{median}} + (\text{IQR}/1.349)^2$.

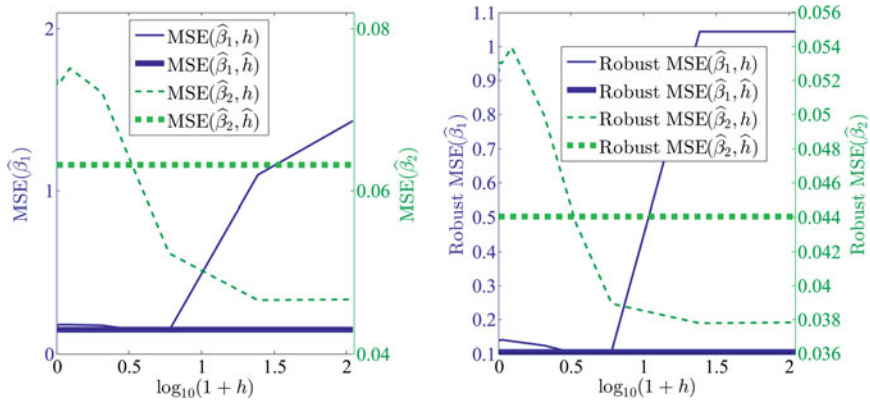


FIGURE 6. Similar to Figure 5, MSE (left) and "robust MSE" (right) for DGP 4.3.

of deviations. Here, the randomness in $\hat{h}$ is not helpful. In DGP 4.2 (not shown), the SAP for $\hat{h}$ is actually a few percentage points *below* that for $h = 0$ (which in turn is below the infeasible h^*), and in DGP 4.1, the SAP improvement from using the infeasible h^* instead of $\hat{h}$ is similar in magnitude to the improvement from using $\hat{h}$ instead of $h = 0$. Depending on one's loss function of type I and type II errors, the SEE-based test may be preferred or not.

8. 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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

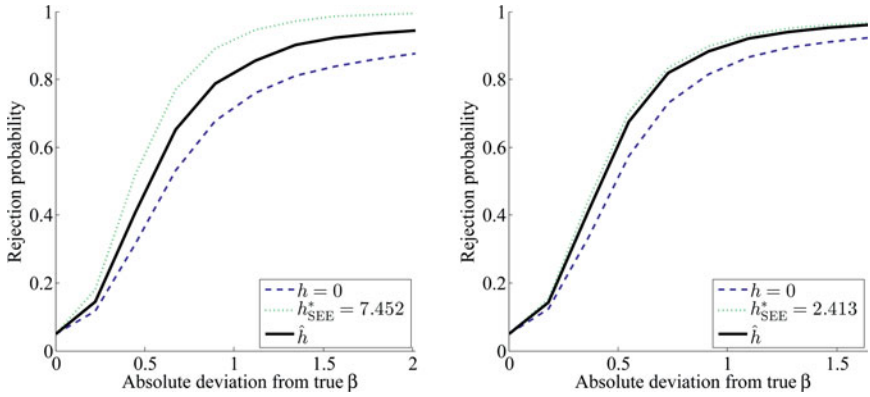


FIGURE 7. Size-adjusted power for DGPs 4.1 (left) and 4.3 (right).

expansions. It can reduce the MSE of both the estimating equations and the parameter estimator, minimize type I error and improve size-adjusted power of a chi-square 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, 2012) are currently under development.

NOTES

1. 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 $m_n(\hat{\beta}) = 0$.

2. We are not alone in recommending this simple strategy for empirical work. Chernozhukov and Hansen (2006) make the same recommendation in their Remark 5 and use this strategy in their empirical application. See also Kwak (2010).

3. A theoretically efficient estimator can be obtained using the sieve minimum distance approach. It entails first estimating the conditional expectation $\mathbb{E}\left[\left(1\{Y_j < X_j\beta\} - q\right) \mid Z_j\right]$ using $\Phi^K(Z_j)$ as the basis functions and then choosing β to minimize a weighted sum of squared conditional expectations. See, for example, Chen and Pouzo (2009, 2012). To achieve the semiparametric efficiency bound, K has to grow with the sample size at an appropriate rate. In work in progress, we consider nonparametric quantile regression with endogeneity and allow K to diverge, which is necessary for both identification and efficiency. Here we are content with a fixed K for empirical convenience at the cost of possible efficiency loss.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

4. This is different from Zhou, Wan, and Yuan (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.

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

6. The CPE is defined to be the nominal coverage 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).

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

8. Their data and Matlab code for replication are helpfully provided online in the Angrist Data Archive, `http://economics.mit.edu/faculty/angrist/data1/data/abangim02`.

9. Code and data for replication is available on the first author's website.

10. Additional JTPA data downloaded from the W.E. Upjohn Institute at http://upjohn.org/services/resources/employment-research-data-center/national-jtpa-study variables are named `age`, `bfhrswrk`, and `bfwage` in file `expbif.dta`.

11. Code to replicate our simulations is available on the first author's website.

REFERENCES

Abadie, A., J. Angrist, & G. Imbens (2002) Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70, 91–117.

Bera, A.K., Y. Bilias, & P. Simlai (2006) Estimating functions and equations: An essay on historical developments with applications to econometrics. In T.C. Mills & K. Patterson (eds.), *Palgrave Handbook of Econometrics: Volume 1 Econometric Theory*, pp. 427–476. Palgrave MacMillan.

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

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

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–334.

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

Chen, X. & D. Pouzo (2012) Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica* 80, 277–322.

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

Chernozhukov, V. & C. Hansen (2005) An IV model of quantile treatment effects. *Econometrica* 73, 245–261.

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

Chernozhukov, V. & C. Hansen (2008) Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics* 142, 379–398.

Chernozhukov, V. & C. Hansen (2013) Quantile models with endogeneity. *Annual Review of Economics* 5, 57–81.

Chernozhukov, V. & H. Hong (2003) An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293–346.

Fan, J. & Y. Liao (2014) Endogeneity in high dimensions. *Annals of Statistics* 42, 872–917.

Galvao, A.F. (2011) Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics* 164, 142–157.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Hall, P. (1992) *Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer-Verlag.
Heyde, C.C. (1997) *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer Series in Statistics. Springer.
Horowitz, J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–531.
Horowitz, J.L. (1998) Bootstrap methods for median regression models. *Econometrica* 66, 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, 73–101.
Hwang, J. & 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, Department of Economics, UC San Diego.
Jun, S.J. (2008) Weak identification robust tests in an instrumental quantile model. *Journal of Econometrics* 144, 118–138.
Kinal, T.W. (1980) The existence of moments of k-class estimators. *Econometrica* 48, 241–249.
Koenker, R. & G. Bassett, Jr. (1978) Regression quantiles. *Econometrica* 46, 33–50.
Kwak, D.W. (2010) Implementation of instrumental variable quantile regression (IVQR) methods Working paper, Michigan State University.
Liang, K.-Y. & S. Zeger (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
MaCurdy, T. & H. Hong (1999) Smoothed quantile regression in generalized method of moments. Working paper, Stanford University.
Müller, H.-G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics* 12, 766–774.
Nagar, A.L. (1959) The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27, 573–595.
Newey, W.K. (1990) Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58, 809–837.
Newey, W.K. (2004) Efficient semiparametric estimation via moment restrictions. *Econometrica* 72, 1877–1897.
Newey, W.K. & J.L. Powell (1990) Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* 6, 295–317.
Otsu, T. (2008) Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics* 142, 508–538.
Phillips, P.C.B. (1982) Small sample distribution theory in econometric models of simultaneous equations. Cowles Foundation Discussion Paper 617, Yale University.
Ruppert, D. & R.J. Carroll (1980) Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* 75, 828–838.
van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge University Press.
Whang, Y.-J. (2006) Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory* 22, 173–205.
Zhou, Y., A.T.K. Wan, & Y. Yuan (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 $P(U_j < 0 \mid Z_j) = q$ for almost all Z_j and $h \to 0$, we can assume

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

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

Under Assumption 3, for some bounded $C(\cdot)$ we have

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

Hence

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

Second moment of W_j. For the second moment,

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

Integrating by parts and using Assumption 3(i) in the last line yields:

$$\begin{aligned}
&\int_{U_L(Z_j)}^{U_H(Z_j)} [G(-u/h)-q]^2 dF_{U|Z}(u \mid Z_j) \\
&\quad = [G(-u/h)-q]^2 F_{U|Z}(u \mid Z_j)\Big|_{U_L(Z_j)}^{U_H(Z_j)}
\end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

But

$$
\begin{aligned}
2\int_{-1}^{1}[G(-v)-q]G'(-v)dv &= \int_{-1}^{1} 2[G(u)-q]G'(u)du \\
&= \left[G^2(u)-2qG(u)\right]\Big|_{-1}^{1} = 1-2q, \\
2\int_{-1}^{1} v[G(-v)-q]G'(-v)dv &= -2\int_{-1}^{1} u[G(u)-q]G'(u)du \\
&= -2\int_{-1}^{1} uG(u)G'(u)du \\
&= -\left[uG^2(u)\Big|_{-1}^{1} - \int_{-1}^{1} G^2(u)du\right] \\
&= -\left[1-\int_{-1}^{1} G^2(u)du\right] \quad \text{(by Assumption 4(ii)),}
\end{aligned}
$$

and

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

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

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

Similarly, we can show that

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

$$m_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_j = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big[W_j - \mathbb{E}(W_j)\big] + \sqrt{n}\mathbb{E}(W_j). \tag{A.1}$$

In view of the mean of W_j, we have $\sqrt{n}\mathbb{E}(W_j) = O(h^r\sqrt{n}) = o(1)$ by Assumption 5. So the bias is asymptotically (first-order) negligible. Consequently, the variance of W_j is $\mathbb{E}\big(W_j W_j'\big) + o(1)$, 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 (A.1), 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

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

For any $\varepsilon > 0$,

$$\begin{aligned}&\lim_{n\to\infty}\sum_{j=1}^{n}\mathbb{E}\left(\frac{W_j - \mathbb{E}W_j}{\sqrt{n}\sigma_W}\right)^2 1\left\{\frac{|W_j - \mathbb{E}W_j|}{\sqrt{n}\sigma_W} \geq \varepsilon\right\}\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\frac{(W_j - \mathbb{E}W_j)^2}{\sigma_W^2}1\left\{\frac{|W_j - \mathbb{E}W_j|}{\sigma_W} \geq \sqrt{n}\varepsilon\right\}\\ &= \lim_{n\to\infty}\mathbb{E}\frac{(W_j - EW_j)^2}{\sigma_W^2}1\left\{\frac{|W_j - EW_j|}{\sigma_W} \geq \sqrt{n}\varepsilon\right\} = 0,\end{aligned}$$

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

$$\frac{(W_j - \mathbb{E}W_j)^2}{\sigma_W^2}1\left\{\frac{|W_j - \mathbb{E}W_j|}{\sigma_W} \geq \sqrt{n}\varepsilon\right\} \leq C\frac{Z_j^2 + \mathbb{E}\Big(Z_j^2\Big)}{\sigma_W^2}$$

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

Properties of estimating equations derived from smoothed criterion function

The EE derived from smoothing the criterion function in (2) for $Z_j = X_j$ can be written

$$0 = n^{-1}\sum_{j=1}^{n} W_j, \quad W_j \equiv X_j\big[G\big(-U_j/h\big) - q\big] + (1/h)G'\big(-U_j/h\big)\big(-X_jU_j\big).$$

We calculate $\mathbb{E}(W_j)$ and $\mathbb{E}(W_j W_j')$, which can be compared to the results in Theorem 1.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

First moment. Using iterated expectations,

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

Second moment. Here,

$$\begin{aligned}
W_jW_j'=&\underbrace{[G(-U_j/h)-q]X_jX_j'}_{\text{Term 1}}+\underbrace{(2/h)[G(-U_j/h)-q]G'(-U_j/h)(-U_j)X_jX_j'}_{\text{Term 2}}\\
&+\underbrace{h^{-2}[G'(-U_j/h)]^2U_j^2X_jX_j'}_{\text{Term 3}}.
\end{aligned}$$

Term 1 is identical to the (only) SEE term, so its expectation is identical, too. The dominant terms of the expectations of Terms 2 and 3 are both positive and $O(h)$.

For Term 2, the expectation conditional on X_j is

$$\begin{aligned}
&(2/h)\int_{-h}^{h}[G(-u/h)-q]G'(-u/h)(-u)f_{U|X}(u\mid X_j)du\,X_jX_j'\\
&\quad=(2/h)\int_{-1}^{1}[G(-v)-q]G'(-v)(-vh)f_{U|X}(hv\mid X_j)hdv\,X_jX_j'\\
&\quad=-2\int_{-1}^{1}[1-G(v)-q]G'(v)(vh)f_{U|X}(hv\mid X_j)dv\,X_jX_j'\\
&\quad=-2(1-q)\int_{-1}^{1}G'(v)(vh)\Bigg[f_{U|X}(0\mid X_j)+(hv)f_{U|X}'(0\mid X_j)\\
&\qquad\qquad+\cdots+\frac{(hv)^{r-1}}{(r-1)!}f_{U|X}^{(r-1)}(0\mid X_j)\Bigg]dv\,X_jX_j'\\
&\qquad+2\int_{-1}^{1}G(v)G'(v)(vh)\left[f_{U|X}(0\mid X_j)+(hv)f_{U|X}'(\tilde{v}h\mid X_j)\right]dv\,X_jX_j'\\
&\quad=\left[O(h^r)+2h\underbrace{\int_{-1}^{1}G(v)G'(v)v\,dv}_{>0}f_{U|X}(0\mid X_j)+O(h^2)\right]X_jX_j'.
\end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

The largest is the $O(h)$ middle term, which is positive.

For Term 3, the expectation conditional on X_j is

$$
\begin{aligned}
&h^{-2}\int_{-h}^{h}[G'(-u/h)]^2u^2 f_{U|X}(u \mid X_j)du\, X_j X_j' \\
&\quad = h^{-2}\int_{-1}^{1}[G'(-v)]^2(hv)^2 f_{U|X}(vh \mid X_j)h dv\, X_j X_j' \\
&\quad = \left[hf_{U|X}(0 \mid X_j)\int_{-1}^{1}[G'(v)v]^2 dv + O\left(h^2\right)\right]X_j X_j'.
\end{aligned}
$$

This is $O(h)$ and positive since the integrand is positive. For the $G(\cdot)$ we use in our code, for example, $\int_{-1}^{1}[G'(v)v]^2dv = 0.061$; values are similar if $G'(\cdot)$ is a bisquare (0.065) or Epanechnikov (0.086) kernel.

Altogether,

$$
\begin{aligned}
\mathbb{E}\left(W_j W_j'\right) = &\underbrace{q(1-q)\mathbb{E}\left(X_j X_j'\right) - h\left(1-\int_{-1}^{1}[G(v)]^2dv\right)\mathbb{E}\left[f_{U|X}(0 \mid X_j)X_j X_j'\right] + O\left(h^2\right)}_{\text{from Term 1}} \\
&\underbrace{+h\int_{-1}^{1}2G(v)G'(v)v dv\,\mathbb{E}\left[f_{U|X}(0 \mid X_j)X_j X_j'\right]}_{\text{from Term 2}} \\
&\underbrace{+h\int_{-1}^{1}[G'(v)v]^2 dv\,\mathbb{E}\left[f_{U|X}(0 \mid X_j)X_j X_j'\right] + O\left(h^2\right)}_{\text{from Term 3}} \\
&= q(1-q)E\left(X_j X_j'\right) + h\int_{-1}^{1}[G'(v)v]^2 dv\,\mathbb{E}\left[f_{U|X}(0 \mid X_j)X_j X_j'\right] + O\left(h^2\right).
\end{aligned}
$$

The cancellation between the $O(h)$ parts of Terms 1 and 2 is by integration by parts of Term 2, using $\frac{\partial}{\partial v}G(v)^2 = 2G(v)G'(v)$. The remaining $O(h)$ term is positive, whereas for SEE the $O(h)$ term is negative.

AMSE of estimator. Analogous to (16), we examine the estimator's properties by way of a mean value expansion, using a similar change of variables, Taylor expansion, and properties of $G'(\cdot)$,

$$
\begin{aligned}
&\mathbb{E}\frac{\partial}{\partial\beta'}n^{-1/2}m_n(\beta_0) \\
&\quad = \mathbb{E}\left[(1/h)G'(-U_j/h)X_j X_j' + \underbrace{h^{-1}G'(-U_j/h)X_j X_j' - h^{-2}G''(-U_j/h)U_j X_j X_j'}_{\text{Not in SEE; using product rule for derivative}}\right] \\
&\quad = \mathbb{E}\left\{2\int_{-1}^{1}G'(-v)\left[f_{U|X}(0 \mid X_j) + \cdots\right]dv\, X_j X_j'\right\}
\end{aligned}
$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$$
\begin{aligned}
&+\mathbb{E}\left\{h^{-1}\int_{-1}^{1} G''(v)vh\left[f_{U|X}(0\mid X_j)+(vh)f'_{U|X}(0\mid X_j)+\cdots\right]dv\,X_jX_j'\right\}\\
&=\mathbb{E}\left[2f_{U|X}(0\mid X_j)X_jX_j'\right]+O(h^r)\\
&\quad+(-1)\mathbb{E}\left[f_{U|X}(0\mid X_j)X_jX_j'\right]+(-1)h\mathbb{E}\left[f'_{U|X}(0\mid X_j)X_jX_j'\right]+O\left(h^2\right)\\
&=\mathbb{E}\left[f_{U|X}(0\mid X_j)X_jX_j'\right]-h\mathbb{E}\left[f'_{U|X}(0\mid X_j)X_jX_j'\right]+O(h^2),
\end{aligned}
$$

where $G''(-v)=-G''(v)$, and integration by parts gives

$$\int_{-1}^{1} vdG'(v)=vG'(v)\big|_{-1}^{1}-\int_{-1}^{1}G'(v)dv=0-1=-1.$$

The dominant term is the same as for SEE. However, the $O(h)$ term will also affect the estimator's AMSE, through the first-order variance. Its sign is indeterminant since it depends on the conditional PDF derivative.

Proof of Proposition 2

The first expression comes directly from the FOC. Under the assumption $U \perp\!\!\!\perp Z$, we have

$$h^*_{\mathrm{SEE}}=\left(\frac{(r!)^2\left[1-\int_{-1}^{1}G^2(u)du\right]f_U(0)}{2r\left(\int_{-1}^{1}G'(v)v^r\,dv\right)^2\left[f_U^{(r-1)}(0)\right]^2}\frac{\mathbb{E}(Z'V^{-1}Z)}{\mathbb{E}(Z)'V^{-1}\mathbb{E}(Z)}\frac{1}{n}\right)^{\frac{1}{2r-1}}.$$

The simplified h^*_{SEE} then follows from the lemma below.

LEMMA A.1. *If $Z\in\mathbb{R}^d$ is a random vector with first element equal to one and $V\equiv\mathbb{E}(ZZ')$ is nonsingular, then*

$$\mathbb{E}(Z'V^{-1}Z)/\left[\mathbb{E}(Z')V^{-1}\mathbb{E}(Z)\right]=d.$$

Proof. For the numerator, rearrange using the trace:

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

For the denominator, let $\mathbb{E}(Z')=(1,t')$ 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 $\mathbb{E}(Z')$ and $\mathbb{E}(Z)$. Writing the other $(d-1)\times(d-1)$ part of the matrix as Ω,

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

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

$$V^{-1}\begin{pmatrix}1 & t'\\ t & \Omega\end{pmatrix}=V^{-1}V=I_d=\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}.$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Thus,

$$\mathbb{E}(Z')V^{-1}\mathbb{E}(Z) = (1,t')(1,0,\ldots,0)' = 1. \qquad \blacksquare$$

Proof of Theorem 3

Adding to the variables already defined in the main text, let

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

Then using the definition of Λ_n in (12), we have

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

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

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

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 $\mathbb{E}\left(\|D_n\|^2\right) = O(1/n)$. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the smallest and largest eigenvalues of a matrix, then for any constant $C > 2\sqrt{d} > 0$:

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

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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 5, we can rewrite the above as

$$P\left\{\|\eta_n\| \geq Ch^2/\log n\right\} = O\left(h^2\right) \tag{A.3}$$

for any large enough constant $C > 0$.

Using (A.2) and defining $W_j^* \equiv \Lambda_n^{-1} Z_j\left[G\left(-U_j/h\right) - q\right]$, we have

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

where

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

and $\bar{W}_n^* = n^{-1}\sum_{j=1}^n W_j^*$ as defined in (13). Using the Chernoff bound on $\sqrt{n}\bar{W}_n^*$ and the result in (A.3), we can show that $P\left(|e_n| > Ch^2\right) = O(h^2)$. 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

$$\begin{aligned} \mathbb{E}\left[\exp\left(it S_n^L\right)\right] &= C_0(t) - hC_1(t) + O\left(h^2\right) \text{ where} \\ C_0(t) &\equiv \mathbb{E}\left\{\exp\left[it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\sqrt{n}\bar{W}_n^*\right)\right]\right\}, \\ C_1(t) &\equiv \mathbb{E}\left\{it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\mathbb{E}AA'\right)\left(\sqrt{n}\bar{W}_n^*\right)\exp\left[it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\sqrt{n}\bar{W}_n^*\right)\right]\right\}. \end{aligned}$$

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

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

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

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

First,

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

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,

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

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

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

$$\begin{aligned}\mathbb{E}\left[\exp\left(itS_n^L\right)\right] &= \overbrace{(1-2it)^{-d/2}\exp\left(\frac{i\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2 t}{1-2it}\right)}^{C_0(t)} - h\overbrace{(1-2it)^{-d/2-1}it\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]}^{O(1)\text{ term in } C_1(t)} \\ &\quad + \overbrace{O\left(nh^{2r+1}\right)+O\left(h^{r+1}\right)}^{\text{remainder from } hC_1(t)} + O(h^2) \\ &= (1-2it)^{-d/2} + (1-2it)^{-d/2-1}it\left\{\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2 - h\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]\right\} \\ &\quad + O\left(h^2+nh^{2r+1}\right). \end{aligned} \tag{A.4}$$

The χ_d^2 characteristic function is $(1-2it)^{-d/2}$, and integrating by parts yields the Fourier–Stieltjes transform of the χ_d^2 PDF:

$$\begin{aligned}\int_0^\infty \exp(itx)d\mathcal{G}_d'(x) &= \int_0^\infty \exp(itx)\mathcal{G}_d''(x)dx \\ &= \exp(itx)\mathcal{G}_d'(x)|_0^\infty - \int_0^\infty (it)\exp(itx)\mathcal{G}_d'(x)dx \\ &= (-it)(1-2it)^{-d/2}.\end{aligned}$$

Taking a Fourier–Stieltjes inversion of (A.4) thus yields

$$\begin{aligned}P\left(S_n^L < x\right) &= \mathcal{G}_d(x) - \mathcal{G}_{d+2}'(x)\left\{\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2 - h\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]\right\} + O\left(h^2+nh^{2r+1}\right) \\ &= \mathcal{G}_d(x) - \mathcal{G}_{d+2}'(x)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]\right\} + O\left(h^2+nh^{2r+1}\right).\end{aligned}$$

A direct implication is that type I error is

$$\begin{aligned}P\left(m_n'\hat{V}^{-1}m_n > c_\alpha\right) &= \alpha + \mathcal{G}_{d+2}'(c_\alpha)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}\left(AA'\right)\right]\right\} \\ &\quad + O\left(h^2+nh^{2r+1}\right).\end{aligned}$$

■

Proof of Theorem 5

Define

$$W_j \equiv W_j(\delta) \equiv Z_j\left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{U_j}{h}\right) - q\right],$$

then

$$m_n(\beta_0) = \frac{1}{\sqrt{n}}\sum_{j=1}^n W_j = \frac{1}{\sqrt{n}}\sum_{j=1}^n W_j(\delta).$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

We first compute the mean of $m_n(\beta_0)$. Let $[U_L(Z_j, X_j), U_H(Z_j, X_j)]$ be the support of U_j conditional on Z_j and X_j. Using the same argument as in the proof of Theorem 1,

$$\begin{aligned}
\mathbb{E}[m_n(\beta_0)] &= \sqrt{n}\mathbb{E}(W_j) = \sqrt{n}\mathbb{E}\left\{ Z_j \int_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} \left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) - q \right] dF_{U|Z,X}(u \mid Z_j, X_j) \right\} \\
&= \sqrt{n}\mathbb{E}\left\{ Z_j \left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) - q \right] F_{U|Z,X}(u \mid Z_j, X_j) \Big|_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} \right\} \\
&\quad + \frac{\sqrt{n}}{h}\mathbb{E}\left\{ Z_j \int_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} F_{U|Z,X}(u \mid Z_j, X_j) G'\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) du \right\} \\
&= -\sqrt{n}\mathbb{E}(Z_j q) + \sqrt{n}\mathbb{E}\left[Z_j \int_{-1}^{1} F_{U|Z,X}\left(\frac{X_j'\delta}{\sqrt{n}} - hv \mid Z_j, X_j\right) G'(v) dv \right] \\
&= \sqrt{n}\mathbb{E}\left\{ Z_j \left[F_{U|Z,X}\left(\frac{X_j'\delta}{\sqrt{n}} \mid Z_j, X_j\right) - q \right] \right\} \\
&\quad + \sqrt{n}\mathbb{E}\left\{ Z_j \int_{-1}^{1} \left[f_{U|Z,X}^{(r-1)}\left(\frac{X_j'\delta}{\sqrt{n}} \mid Z_j, X_j\right) \frac{(-h)^r v^r}{r!} \right] G'(v) dv \right\} + O\left(\sqrt{n}h^{r+1}\right).
\end{aligned}$$

Expanding $F_{U|Z,X}\left(\frac{X_j'\delta}{\sqrt{n}} \mid Z_j, X_j\right)$ and $f_{U|Z,X}^{(r-1)}\left(\frac{X_j'\delta}{\sqrt{n}} \mid Z_j, X_j\right)$ at zero, and since r is even,

$$\begin{aligned}
\mathbb{E}[m_n(\beta_0)] &= \sqrt{n}\mathbb{E}\{Z_j[F_{U|Z,X}(0 \mid Z_j, X_j) - q]\} \\
&\quad + \mathbb{E}\left[Z_j X_j' \delta f_{U|Z,X}(0 \mid Z_j, X_j) \right] + O\left(n^{-1/2}\right) \\
&\quad + \frac{h^r}{r!}\sqrt{n}\mathbb{E}\left[Z_j f_{U|Z,X}^{(r-1)}(0 \mid Z_j, X_j) \right] \left(\int_{-1}^{1} v^r G'(v) dv \right) + O\left(\sqrt{n}h^{r+1} + h^r\right) \\
&= \mathbb{E}\left[f_{U|Z,X}(0 \mid Z_j, X_j) Z_j X_j' \delta \right] + \sqrt{n}h^r V^{1/2}\mathbb{E}(B) + O\left(n^{-1/2} + \sqrt{n}h^{r+1} + h^r\right).
\end{aligned}$$

Here we have used the following extensions of the law of iterated expectation:

$$\begin{aligned}
&\mathbb{E}\{Z_j[F_{U|Z,X}(0 \mid Z_j, X_j) - q]\} = \mathbb{E}\{Z_j \mathbb{E}[\mathbb{E}(1\{U_j < 0\} \mid Z_j, X_j) - q \mid Z_j]\} \\
&\quad = \mathbb{E}\{Z_j[F_{U|Z}(0 \mid Z_j) - q]\} = 0, \\
&\mathbb{E}[f_{U|Z,X}(u \mid Z, X) \mid Z = z] = \int_{\mathcal{X}} f_{U|Z,X}(u \mid z, x) f_{X|Z}(x \mid z) dx \\
&\quad = \int_{\mathcal{X}} \frac{f_{U,Z,X}(u,z,x)}{f_{Z,X}(z,x)} \frac{f_{Z,X}(z,x)}{f_Z(z)} dx = [f_Z(z)]^{-1} \int_{\mathcal{X}} f_{U,Z,X}(u,z,x) dx \\
&\quad = f_{U,Z}(u,z)/f_Z(z) = f_{U|Z}(u \mid z), \\
&\mathbb{E}[f_{U|Z,X}(0 \mid Z, X) g(Z)] = \mathbb{E}\{\mathbb{E}[f_{U|Z,X}(0 \mid Z, X) g(Z) \mid Z]\} \\
&\quad = \mathbb{E}\{\mathbb{E}[f_{U|Z,X}(0 \mid Z, X) \mid Z] g(Z)\} = \mathbb{E}\{f_{U|Z}(0 \mid Z) g(Z)\},
\end{aligned}$$

and similarly for derivatives of the PDF by exchanging the order of differentiation and integration.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

Next, we compute the variance V_n of $m_n(\beta_0)$:

$$
\begin{aligned}
V_n &= \mathrm{Var}\big[m_n(\beta_0)\big] = \mathrm{Var}\big[W_j(\delta)\big] \\
&= \mathbb{E}\left\{\left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{U_j}{h}\right) - q\right]^2 Z_j Z_j'\right\} - \big[\mathbb{E}W_j(\delta)\big]\big[\mathbb{E}W_j(\delta)\big]' \\
&= \mathbb{E}\left\{\left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{U_j}{h}\right) - q\right]^2 Z_j Z_j'\right\} + O\left(n^{-1} + h^{2r}\right).
\end{aligned}
$$

Now

$$
\begin{aligned}
&\mathbb{E}\left\{\left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{U_j}{h}\right) - q\right]^2 \mid Z_j, X_j\right\} \\
&= \int_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} \left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) - q\right]^2 dF_{U|Z,X}\big(u \mid Z_j, X_j\big) \\
&= \left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) - q\right]^2 F_{U|Z,X}\big(u \mid Z_j, X_j\big)\Big|_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} \\
&\quad + \frac{2}{h}\int_{U_L(Z_j,X_j)}^{U_H(Z_j,X_j)} F_{U|Z,X}\big(u \mid Z_j, X_j\big)\left[G\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) - q\right] G'\left(\frac{X_j'\delta}{\sqrt{n}h} - \frac{u}{h}\right) du \\
&= q^2 + 2\int_{-1}^{1} F_{U|Z,X}\left(hv + \frac{X_j'\delta}{\sqrt{n}} \mid Z_j, X_j\right)[G(-v) - q]G'(-v)dv \\
&= q^2 + 2F_{U|Z,X}\big(0 \mid Z_j, X_j\big)\int_{-1}^{1}[G(-v) - q]G'(-v)dv \\
&\quad + 2hf_{U|Z,X}\big(0 \mid Z_j, X_j\big)\left[\int_{-1}^{1} v[G(-v) - q]G'(-v)dv\right] \\
&\quad + \frac{2}{\sqrt{n}}\left[f_{U|Z,X}\big(0 \mid Z_j, X_j\big)X_j'\delta\int_{-1}^{1}[G(-v) - q]G'(-v)dv\right] + O\left(h^2 + n^{-1}\right) \\
&= q^2 + F_{U|Z,X}\big(0 \mid Z_j, X_j\big)(1-2q) - hf_{U|Z,X}\big(0 \mid Z_j, X_j\big)\left(1 - \int_{-1}^{1} G^2(u)du\right) \\
&\quad + \frac{(1-2q)}{\sqrt{n}}\left[f_{U|Z,X}\big(0 \mid Z_j, X_j\big)X_j'\delta\right] + O\left(h^2 + n^{-1}\right),
\end{aligned}
$$

and so

$$
\begin{aligned}
V_n &= q^2\mathbb{E}\left(Z_j Z_j'\right) + (1-2q)\mathbb{E}\left[F_{U|Z,X}\big(0 \mid Z_j, X_j\big)Z_j Z_j'\right] \\
&\quad + h\left(1 - \int_{-1}^{1} G^2(u)du\right)\mathbb{E}\left[f_{U|Z,X}\big(0 \mid Z_j, X_j\big)Z_j Z_j'\right]
\end{aligned}
$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$$+\frac{(1-2q)}{\sqrt{n}}\mathbb{E}\left\{\left[f_{U|Z,X}(0\mid Z_j,X_j)X_j'\delta\right]Z_jZ_j'\right\}+O\left(n^{-1}+h^2\right)$$
$$=V-hV^{1/2}\mathbb{E}(AA')\left(V^{1/2}\right)'+O\left(n^{-1/2}+h^2\right),$$

where the last line holds because of the above law of iterated expectation extension and

$$\begin{aligned}
&q^2\mathbb{E}\left(Z_jZ_j'\right)+(1-2q)\mathbb{E}\left[F_{U|Z,X}(0\mid Z_j,X_j)Z_jZ_j'\right]\\
&\quad=q^2\mathbb{E}\left(Z_jZ_j'\right)+(1-2q)\mathbb{E}\left\{\mathbb{E}\left[1\{U<0\}\mid Z_j,X_j\right]Z_jZ_j'\right\}\\
&\quad=q^2\mathbb{E}\left(Z_jZ_j'\right)+(1-2q)\mathbb{E}\left\{\mathbb{E}\left[1\{U<0\}Z_jZ_j'\mid Z_j,X_j\right]\right\}\\
&\quad=q^2\mathbb{E}\left(Z_jZ_j'\right)+(1-2q)\mathbb{E}\left(1\{U<0\}Z_jZ_j'\right)=q(1-q)\mathbb{E}\left(Z_jZ_j'\right).
\end{aligned}$$

Let $\Lambda_n=V_n^{1/2}$, then

$$\Lambda_n=V^{1/2}\left[I_d-h\mathbb{E}(AA')+O\left(n^{-1/2}+h^2\right)\right]^{1/2}.$$

Define $W_j^*\equiv W_j^*(\delta)=\Lambda_n^{-1}W_j(\delta)$ and

$$\bar{W}_n^*\equiv\bar{W}_n^*(\delta)=n^{-1}\sum_{j=1}^{n}W_j^*(\delta).$$

Then $\Delta=\sqrt{n}\mathbb{E}\left(W_j^*\right)$ and

$$\begin{aligned}
\|\Delta\|^2&=\left\|V_n^{-1/2}\Sigma_{ZX}\delta+V_n^{-1/2}\sqrt{n}(-h)^rV^{1/2}\mathbb{E}(B)\right\|^2\\
&=\left\|V_n^{-1/2}V^{1/2}\tilde{\delta}+\sqrt{n}(-h)^rV_n^{-1/2}V^{1/2}\mathbb{E}(B)\right\|^2\\
&=\left\|\left[I_d-h\mathbb{E}(AA')\right]^{-1/2}\tilde{\delta}+\sqrt{n}(-h)^r\mathbb{E}(B)\right\|^2[1+o(1)]\\
&=\left\|\left[I_d+\frac{1}{2}h\mathbb{E}(AA')\right]\tilde{\delta}+\sqrt{n}(-h)^r\mathbb{E}(B)\right\|^2[1+o(1)]\\
&=\left\{\left\|\tilde{\delta}\right\|^2+h\tilde{\delta}'\left[\mathbb{E}(AA')\right]\tilde{\delta}+nh^{2r}\mathbb{E}(B)'\mathbb{E}(B)+2\tilde{\delta}'\sqrt{n}(-h)^r\mathbb{E}(B)\right\}[1+o(1)].
\end{aligned}$$

We can now write

$$S_n=m_n(\beta_0)'\hat{V}^{-1}m_n(\beta_0)=S_n^L+e_n$$

where

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

By the same argument as in the proof of Theorem 3, we can show that the presence of e_n generates an approximation error that is not larger than that given in Theorem 5.

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

The characteristic function of S_n^L is

$$\mathbb{E}\left[\exp\left(it S_n^L\right)\right] = C_0(t) - hC_1(t) + O\left(h^2 + n^{-1/2}\right) \text{ where}$$
$$C_0(t) \equiv \mathbb{E}\left\{\exp\left[it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\sqrt{n}\bar{W}_n^*\right)\right]\right\},$$
$$C_1(t) \equiv \mathbb{E}\left\{it\left(\sqrt{n}\bar{W}_n^*\right)'\mathbb{E}\left(AA'\right)\left(\sqrt{n}\bar{W}_n^*\right)\exp\left[it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\sqrt{n}\bar{W}_n^*\right)\right]\right\}.$$

Using the expansion of the PDF of $n^{-1/2}\sum_{j=1}^n\left(W_j^* - \mathbb{E}W_j^*\right)$:

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

where $p(x)$ is an odd polynomial in the elements of x of degree 3, we obtain

$$\begin{aligned} C_0(t) &= \mathbb{E}\left\{\exp\left[it\left(\sqrt{n}\bar{W}_n^*\right)'\left(\sqrt{n}\bar{W}_n^*\right)\right]\right\} \\ &= (2\pi)^{-d/2}\int \exp\left\{it\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right]'\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right]\right\}\exp(-x'x/2)dx \\ &\quad + O\left(n^{-1/2}\right) \\ &= (1-2it)^{-d/2}\exp\left(\frac{it\left\|\sqrt{n}\mathbb{E}\left(W_j^*\right)\right\|^2}{1-2it}\right) + O\left(n^{-1/2}\right). \end{aligned}$$

Similarly,

$$\begin{aligned} C_1(t) = (2\pi)^{-d/2}\int & it\left(x + \sqrt{n}\mathbb{E}W_j^*\right)'\mathbb{E}\left(AA'\right)\left(x + \sqrt{n}\mathbb{E}W_j^*\right) \\ & \times \exp\left\{it\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right]'\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right] - x'x/2\right\}dx + O\left(n^{-1/2}\right). \end{aligned}$$

Since

$$\begin{aligned} & it\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right]'\left[x + \sqrt{n}\mathbb{E}\left(W_j^*\right)\right] - x'x/2 \\ &= -\frac{1}{2}(1-2it)\left[x - \frac{2it}{1-2it}\sqrt{n}\mathbb{E}\left(W_j^*\right)\right]'\left[x - \frac{2it}{1-2it}\sqrt{n}\mathbb{E}\left(W_j^*\right)\right] \\ &\quad + \frac{it}{1-2it}\left(\sqrt{n}\mathbb{E}W_j^*\right)'\left(\sqrt{n}\mathbb{E}W_j^*\right), \end{aligned}$$

we have

$$\begin{aligned} C_1(t) &= (1-2it)^{-d/2}\exp\left[\frac{it}{1-2it}\left(\sqrt{n}\mathbb{E}W_j^*\right)'\left(\sqrt{n}\mathbb{E}W_j^*\right)\right] \\ &\quad \times \mathbb{E}\left[it\left(\mathbb{X} + \sqrt{n}\mathbb{E}W_j^*\right)'\mathbb{E}\left(AA'\right)\left(\mathbb{X} + \sqrt{n}\mathbb{E}W_j^*\right)\right] + O\left(n^{-1/2}\right) \\ &= (1-2it)^{-d/2}\exp\left[\frac{it}{1-2it}\left(\sqrt{n}\mathbb{E}W_j^*\right)'\left(\sqrt{n}\mathbb{E}W_j^*\right)\right] \end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$$\times it\mathrm{tr}\left\{\mathbb{E}(AA')\left[\frac{1}{1-2it}I_d+\left(\frac{2it}{1-2it}+1\right)^2\left(\sqrt{n}\mathbb{E}W_j^*\right)\left(\sqrt{n}\mathbb{E}W_j^*\right)'\right]\right\}$$
$$+O\left(n^{-1/2}\right)$$
$$=(1-2it)^{-d/2}\exp\left[\frac{it}{1-2it}\left(\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2\right)\right]$$
$$\times\frac{it}{1-2it}\mathrm{tr}\left\{\mathbb{E}(AA')\left[I_d+\frac{1}{1-2it}\left(\sqrt{n}\mathbb{E}W_j^*\right)\left(\sqrt{n}\mathbb{E}W_j^*\right)'\right]\right\}+O\left(n^{-1/2}\right),$$

where $\mathbb{X}\sim N\left[\frac{2it}{1-2it}\left(\sqrt{n}\mathbb{E}W_j^*\right),diag(1-2it)^{-1}\right]$.

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

$$\mathbb{E}\left[\exp\left(itS_n^L\right)\right]$$
$$=(1-2it)^{-d/2}\exp\left(\frac{it\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2}{1-2it}\right)$$
$$+(1-2it)^{-d/2-1}\exp\left(\frac{it\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2}{1-2it}\right)h(-it)\mathrm{tr}\left[\mathbb{E}(AA')\right]+O\left(h^2+n^{-1/2}\right)$$
$$+(1-2it)^{-d/2-2}\exp\left(\frac{it\left\|\sqrt{n}\mathbb{E}W_j^*\right\|^2}{1-2it}\right)h(-it)\left(\sqrt{n}\mathbb{E}W_j^*\right)'\mathbb{E}(AA')\left(\sqrt{n}\mathbb{E}W_j^*\right).$$

Let $\mathcal{G}'_d(x;\lambda)$ be the PDF of the noncentral chi-square distribution with noncentrality parameter λ, so

$$\mathcal{G}'_d(x;\lambda)=\frac{1}{2\pi}\int_{\mathbb{R}}(1-2it)^{-d/2}\exp\left(\frac{it\lambda}{1-2it}\right)\exp(-itx)dt, \tag{A.5}$$
$$\mathcal{G}''_d(x;\lambda)=\frac{1}{2\pi}\int_{\mathbb{R}}(-it)(1-2it)^{-d/2}\exp\left(\frac{it\lambda}{1-2it}\right)\exp(-itx)dt.$$

Using the above results and taking a Fourier–Stieltjes inversion,

$$P_{\beta_n}(S_n<x)=\mathcal{G}_d\left(x;\|\Delta\|^2\right)+\mathcal{G}'_{d+2}\left(x;\|\Delta\|^2\right)h\mathrm{tr}\left[\mathbb{E}(AA')\right]$$
$$+\mathcal{G}'_{d+4}\left(x;\|\Delta\|^2\right)h\left[\Delta'\mathbb{E}(AA')\Delta\right]+O\left(h^2+n^{-1/2}\right).$$

Expanding $\mathcal{G}_d\left(x;\|\Delta\|^2\right)$ around $\mathcal{G}_d\left(x;\|\tilde{\delta}\|^2\right)$ yields

$$\mathcal{G}_d\left(x;\|\Delta\|^2\right)=\mathcal{G}_d\left(x;\left\|\tilde{\delta}\right\|^2\right)+\left.\frac{\partial\mathcal{G}_d(x,\lambda)}{\partial\lambda}\right|_{\lambda=\left\|\tilde{\delta}\right\|^2}$$
$$\times\left[h\tilde{\delta}'\mathbb{E}(AA')\tilde{\delta}+nh^{2r}\mathbb{E}(B)'\mathbb{E}(B)+2\tilde{\delta}'\sqrt{n}(-h)^r\mathbb{E}(B)\right][1+o(1)]$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

$$= \mathcal{G}_d\left(x; \left\|\tilde{\delta}\right\|^2\right) - \mathcal{G}'_{d+2}\left(x; \left\|\tilde{\delta}\right\|^2\right) \times \left[h\tilde{\delta}'\mathbb{E}(AA')\tilde{\delta} + nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) + 2\tilde{\delta}'\sqrt{n}(-h)^r\mathbb{E}(B)\right][1+o(1)]$$

using the result that $\frac{\partial}{\partial\lambda}\mathcal{G}_d(x;\lambda) = -\mathcal{G}'_{d+2}(x;\lambda)$, which can be derived from (A.5). Hence

$$P_{\beta_n}(S_n < x) = \mathcal{G}_d\left(x; \left\|\tilde{\delta}\right\|^2\right) - \mathcal{G}'_{d+2}\left(x; \left\|\tilde{\delta}\right\|^2\right)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}(AA')\right]\right\} + \left[\mathcal{G}'_{d+4}\left(x; \left\|\tilde{\delta}\right\|^2\right) - \mathcal{G}'_{d+2}\left(x; \left\|\tilde{\delta}\right\|^2\right)\right]h\left[\tilde{\delta}'\mathbb{E}(AA')\tilde{\delta}\right] - \mathcal{G}'_{d+2}\left(x; \left\|\tilde{\delta}\right\|^2\right)2\tilde{\delta}'\sqrt{n}(-h)^r\mathbb{E}(B) + O\left(h^2 + n^{-1/2}\right).$$

Under the assumption that $\tilde{\delta}$ is uniform on the sphere $\mathcal{S}_d(\tau)$, we can write $\tilde{\delta} = \tau\xi/\|\xi\|$ where $\xi \sim N(0, I_d)$. Then

$$\mathbb{E}_{\tilde{\delta}}\left[P_{\beta_n}(S_n < x)\right] = \mathcal{G}_d\left(x;\tau^2\right) - \mathcal{G}'_{d+2}\left(x;\tau^2\right)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}(AA')\right]\right\} + \left[\mathcal{G}'_{d+4}\left(x;\tau^2\right) - \mathcal{G}'_{d+2}\left(x;\tau^2\right)\right]\tau^2 h\mathrm{tr}\left[\mathbb{E}(AA')\mathbb{E}_\xi\left(\xi\xi'/\|\xi\|^2\right)\right] + O\left(h^2 + n^{-1/2}\right)$$

where $\mathbb{E}_\xi$ is the expectation with respect to ξ. As a consequence,

$$\mathbb{E}_{\tilde{\delta}}\left[P_{\beta_n}(S_n > x)\right] = 1 - \mathcal{G}_d\left(x;\tau^2\right) + \mathcal{G}'_{d+2}\left(x;\tau^2\right)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}(AA')\right]\right\} - \left[\mathcal{G}'_{d+4}\left(x;\tau^2\right) - \mathcal{G}'_{d+2}\left(x;\tau^2\right)\right]\frac{\tau^2}{d}h\mathrm{tr}\left[\mathbb{E}(AA')\right] + O\left(h^2 + n^{-1/2}\right).$$

Letting $x = c_\alpha$ yields the desired result. ■

Proof of Corollary 6

By direct calculations,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\delta}}\left[P_{\beta_n}(S_n > c^*_\alpha)\right] \\
&= 1 - \mathcal{G}_d\left(c^*_\alpha;\tau^2\right) + \mathcal{G}'_{d+2}\left(c^*_\alpha;\tau^2\right)\left\{nh^{2r}\mathbb{E}(B)'\mathbb{E}(B) - h\mathrm{tr}\left[\mathbb{E}(AA')\right]\right\} \\
&\quad - \left[\mathcal{G}'_{d+4}\left(c^*_\alpha;\tau^2\right) - \mathcal{G}'_{d+2}\left(c^*_\alpha;\tau^2\right)\right]\frac{\tau^2}{d}h\mathrm{tr}\left[\mathbb{E}(AA')\right] + O\left(h^2 + n^{-1/2}\right) \\
&= 1 - \mathcal{G}_d\left(c_\alpha;\tau^2\right) + \mathcal{G}'_d\left(c_\alpha;\tau^2\right)\frac{\mathcal{G}'_{d+2}(c_\alpha)}{\mathcal{G}'_d(c_\alpha)}\left(1 - \frac{1}{2r}\right)\mathrm{tr}\left[\mathbb{E}(AA')\right]h^*_{\mathrm{SEE}} \\
&\quad + \mathcal{G}'_{d+2}\left(c_\alpha;\tau^2\right)\left(\frac{1}{2r} - 1\right)\mathrm{tr}\left[\mathbb{E}(AA')\right]h^*_{\mathrm{SEE}} \\
&\quad - \left[\mathcal{G}'_{d+4}\left(c_\alpha;\tau^2\right) - \mathcal{G}'_{d+2}\left(c_\alpha;\tau^2\right)\right]\frac{\tau^2}{d}h\mathrm{tr}\left[\mathbb{E}(AA')\right] + O\left(h^2 + n^{-1/2}\right) \\
&= 1 - \mathcal{G}_d\left(c_\alpha;\tau^2\right) + Q_d\left(c_\alpha,\tau^2,r\right)\mathrm{tr}\left[\mathbb{E}(AA')\right]h^*_{\mathrm{SEE}} + O\left(h^{*2}_{\mathrm{SEE}} + n^{-1/2}\right), \qquad \textbf{(A.6)}
\end{aligned}$$

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

where

$$Q_d\left(c_\alpha, \tau^2, r\right) = \left[\mathcal{G}_d'\left(c_\alpha; \tau^2\right)\frac{\mathcal{G}_{d+2}'(c_\alpha)}{\mathcal{G}_d'(c_\alpha)} - \mathcal{G}_{d+2}'\left(c_\alpha; \tau^2\right)\right]\left(1 - \frac{1}{2r}\right) - \frac{1}{d}\left[\mathcal{G}_{d+4}'\left(c_\alpha; \tau^2\right) - \mathcal{G}_{d+2}'\left(c_\alpha; \tau^2\right)\right]\tau^2$$

as desired. ■

LEMMA A.2. *Let the assumptions in Theorem 7 hold. Then*

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

and

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

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

$$\frac{1}{\sqrt{n}}m_n(\beta) = \frac{1}{\sqrt{n}}\mathbb{E}[m_n(\beta)] + o_p(1)$$

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

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

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

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

uniformly over $\beta \in \mathcal{B}$. Given the uniform convergence and the identification condition in Assumption 6, we can invoke Theorem 5.9 of van der Vaart (1998) to obtain that $\hat{\beta} \xrightarrow{p} \beta_0$.

Next we prove the first result of the lemma. Under Assumption 4(i–ii), we can use the elementwise mean value theorem to obtain

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

where

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

and each $\tilde{\beta}_i$ is a point between $\hat{\beta}$ and β_0. Under Assumptions 1 and 4(i–ii) and that $\mathbb{E}\left[\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta)\right]$ 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

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

Hence, under the additional Assumption 5 and nonsingularity of Σ_{ZX}, we have $\sqrt{n}\left(\hat{\beta}-\beta_0\right)=O_p(1)$. 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:

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

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

$$\left\{\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}\left[m_n\left(\tilde{\beta}\right)-m_n(\beta_0)\right]\right\}=O_p\left(n^{-1/2}\right).$$

Therefore

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right)&=-\left[\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n(\beta_0)+O_p\left(n^{-1/2}\right)\right]^{-1}m_n\\&=-\left(\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\right)^{-1}m_n+O_p\left(n^{-1/2}\right).\end{aligned}$$

Now

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

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press

so

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

As a result,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right) &= -\left[\mathbb{E}\left(\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\right) + O_p\left(\frac{1}{\sqrt{nh}}\right)\right]^{-1} m_n + O_p\left(n^{-1/2}\right)\\ &= -\left[\mathbb{E}\left(\frac{\partial}{\partial\beta'}\frac{1}{\sqrt{n}}m_n\right)\right]^{-1} m_n + O_p\left(\frac{1}{\sqrt{nh}}\right) + O_p\left(n^{-1/2}\right).\end{aligned}$$

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

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

as desired. ■

https://doi.org/10.1017/S0266466615000407 Published online by Cambridge University Press