Pairwise-comparison estimation with nonparametric controls

Koen Jochmans

SciencesPo. | Department of Economics

Discussion paper 2013-04

Pairwise-comparison estimation with nonparametric controls

Koen Jochmans

Sciences Po Economics Discussion Papers

Pairwise-comparison estimation with nonparametric controls

Koen Jochmans[†]

[†] *Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France.*
E-mail: koen.jochmans@sciences-po.org

[This version: December 7, 2012]

Summary The purpose of this paper is the presentation of distribution theory for generic estimators based on the pairwise comparison of observations in problems where identification is achieved through the use of control functions. The controls can be specified semi- or non-parametrically. The criterion function may be non-smooth. The theory is applied to the estimation of the coefficients in a monotone linear-index model and to inference on the link function in a partially-linear transformation model. A number of simulation exercises serve to assess the small-sample performance of these techniques.

Keywords: *control function, discontinuous criterion function, distribution theory, empirical process, Euclidean class, pairwise comparisons, two-step estimation.*

1. INTRODUCTION

There is a variety of econometric problems where conditioning on control variables is required to obtain moment conditions that identify the parameters of interest. In many such cases, a prior estimation step is required to construct a feasible criterion function based on these identifying restrictions. One prime example is the instrumental-variable estimation of a linear simultaneous-equation model by two-stage least squares. Another familiar illustration is accounting for non-random sample selection through the two-step procedure of Heckman (1979).

In this paper, I consider incorporating controls into estimators that are based on the pairwise comparison of observations. This class of estimators is broad. For example, it includes pairwise-likelihood and pairwise-differencing estimators, as well as estimators based on ranks. Conditions are provided under which such estimators are $\sqrt{n}$-consistent and asymptotically normal. The theory is applied to the estimation of the coefficients in a monotone linear-index model and to inference on the link function in a partially-linear transformation model.

Of course, several other papers have investigated inference in semiparametric models with generated regressors. Extending the approach of Newey (1994), Hahn and Ridder (2012) provided distribution theory for a general class of three-step control-function estimators defined as the solution to empirical moment conditions. Three features of pairwise-comparison estimators make them distinct from their setup, however. First, they maximize U-processes of order two. Second, many pairwise-comparison estimators, and rank estimators in particular, build on moment inequalities. This generally implies the resulting criterion function to be non-smooth. Third, a comparison between observations typically avoids the need to estimate nuisance functions. In practice, this means that the intermediate estimation step in the setup of Hahn and Ridder (2012) can be avoided.

In two recent contributions, Honoré and Powell (2005) and Aradillas-López et al. (2007) extended Ahn and Powell (1993) by deriving distribution theory for a class of

nonlinear pairwise-differencing estimators. While their conditions allow for some lack of differentiability, their approach requires the objective function to be Hölder continuous. This rules out rank estimators and, more generally, criterion functions that behave like step functions. Here, this type of estimation procedures is explicitely allowed for. The extension to severe non-smoothness entertained here is not merely technical, as the requirement of Hölder continuity can substantially restrict the scope of the approach if one is unwilling to impose substantial parametric structure; see, for example, the discussion in Honoré and Powell (2005) [pp. 526 and pp. 528] or the illustration in Aradillas-López et al. (2007). The desire to remain agnostic about the distribution of unobservables is the natural paradigm in most control problems.

The theory in Honoré and Powell (2005) and Aradillas-López et al. (2007) builds on concave combinations of left and right derivatives of the criterion function, upon which a Euclidean condition is subsequently imposed. Here, we directly start from the class of Euclidean criterion functions. Expansions of the large-sample objective function are used to make short work of the non-smooth nature of its small-sample counterpart, as in Sherman (1993). In this sense, our results complement the recent work by Ichimura and Lee (2010) on two-step M-estimators, that is, estimators that maximize U-processes of order one. In addition to the gain in generality compared to Honoré and Powell (2005) and Aradillas-López et al. (2007), constructing the distribution theory in this way leads to more direct and much shortened proofs, as well as to more elementary conditions that are easier to verify.

2. SETUP AND EXAMPLES

Let $V_1, \ldots, V_n$ be independent realizations of a random variable V whose distribution is supported on a set $\mathscr{V}$. Let γ_0 and θ_0 denote unknown elements of two finite-dimensional parameter spaces Γ and Θ. Throughout, θ_0 will indicate the main parameter of interest and γ_0 will denote an auxiliary parameter. Partitioning the parameter vector in this way will enable us to analyze estimators of θ_0 that depend on plug-in estimates of γ_0.[1] Introduce the class of real-valued functions $\mathcal{S} \equiv \{s(\cdot,\cdot;\gamma,\theta) : \gamma \in \Gamma, \theta \in \Theta\}$ on the product set $\mathscr{V} \otimes \mathscr{V}$. Refer to $s(\cdot,\cdot;\gamma,\theta)$ as the score and assume without loss of generality that it is symmetric in its arguments. For each (γ,θ) in $\Gamma \times \Theta$,

$$\binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} s(V_i, V_j; \gamma, \theta) \tag{2.1}$$

is a U-statistic (of order two); the collection over $\Gamma \times \Theta$ is a U-process. Pairwise-comparison estimators are defined as maximizers of U-processes of this form. Distribution theory for this type of estimators can be found in Sherman (1993) and Honoré and Powell (1994), for example.

Incorporating controls in (2.1) will lead to inference from what may be called local U-processes. To describe a local U-process, suppose that, in addition to observations on V, we also observe realizations of a random variable W for the same units; V and W need not be disjoint. Let $D \equiv (V,W)$ and let $D_1, \ldots, D_n$ denote a sequence of n random draws from a distribution P that is supported on $\mathscr{D} \equiv \mathscr{V} \times \mathscr{W}$. For a function $\Xi(\cdot) : \mathscr{W} \mapsto \mathscr{R}^{\mathrm{d}(\Xi)}$, define the (vector-valued) control variable $\Xi = \Xi(W)$. Here and later, $\mathrm{d}(A)$ will denote

[1] This allows for the inclusion of generated regressors and covers multistep estimators, for example.

the dimension of the vector A. For each (γ, θ) in $\Gamma \times \Theta$, define a local U-statistic as

$$\widehat{Q}_n(\theta;\gamma) \equiv \sigma_k^{-\mathrm{d}(\Xi)} \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} s(V_i, V_j; \gamma, \theta)\ k\Big(\frac{\widehat{\Xi}_i - \widehat{\Xi}_j}{\sigma_k}\Big)\ t_i t_j, \tag{2.2}$$

where $\widehat{\Xi}_i \equiv \widehat{\Xi}(W_i)$ is some first-step estimator of $\Xi_i \equiv \Xi(W_i)$, $k(\cdot) : \mathscr{R}^{\mathrm{d}(\Xi)} \mapsto \mathscr{R}$ is a (symmetric) kernel function, $\sigma_k = \sigma_k(n)$ is a bandwidth that shrinks to zero as $n \to \infty$, and t_i $(i = 1, 2, \ldots, n)$ is a trimming term. The trimming will serve to keep (2.2) well defined when the control is estimated nonparametrically; see below.

The estimator of θ_0 that is of interest here may then be defined as

$$\widehat{\theta} \equiv \arg\max_{\theta \in \Theta} \widehat{Q}_n(\theta; \widehat{\gamma}),$$

where $\widehat{\gamma}$ is an auxiliary estimator obtained in a prior stage. Under conditions given below, the introduction of kernel weights leads to asymptotically retaining only those pairs of observations for which $\Xi_i - \Xi_j$ lies in a shrinking neighborhood of zero. The contribution to (2.2) of pairs for which this is not the case fades out as $n \to \infty$. Inference from local U-processes of this kind can be traced back to the seminal contributions of Powell (1987) and Ahn and Powell (1993), who proposed kernel-weighted least-squares estimators of the coefficients in a linear regression model when the data is subject to endogenous sample selection.

To flavor the discussion, and to further motivate the class of estimators that is of focus here, it is useful to sketch some situations of practical interest that can be cast into it. In each of the examples, $V = (Y, X)$ and $\mathscr{V} = \mathscr{Y} \times \mathscr{X}$; Y is a univariate outcome variable and X is a vector of covariates. Throughout, $\mathcal{E}$ denotes the expectations operator with respect to the measure P. Examples 2.2 and 2.3 will be revisited in more detail in Section 4. The first example introduces sample selection into a censored-regression model and modifies Honoré and Powell's (1994) estimator based on artificial censoring; see also Honoré and Powell (2005). Honoré (1992) presents a fixed-effect panel-data version of this model.

Example 2.1. Suppose that $Y = \max\{0, X'\theta_0 - \eta\}$, where η is a latent disturbance. Under independence of η and X,

$$\mathcal{E}[Y|X = x, Y > 0] = x'\theta_0 - \int_{-\infty}^{x'\theta_0} \frac{\varepsilon}{F(x'\theta_0)}\ \mathrm{d}F(\varepsilon),$$

where F denotes the marginal distribution of η. Let (x_1, x_2) be an element of $\mathscr{X} \otimes \mathscr{X}$ and define

$$g(x_1'\theta, x_2'\theta) \equiv \int_{-\infty}^{\min\{x_1'\theta, x_2'\theta\}} \frac{\varepsilon}{F(\min\{x_1'\theta, x_2'\theta\})}\ \mathrm{d}F(\varepsilon).$$

Then, for $i = 1, 2$, $\mathcal{E}[Y|X = x_i, \eta < \min\{x_1'\theta_0, x_2'\theta_0\}] = x_i'\theta_0 - g(x_1'\theta_0, x_2'\theta_0)$ and so it follows that

$$\mathcal{E}[Y|X = x_1, \eta < \min\{x_1'\theta_0, x_2'\theta_0\}] - \mathcal{E}[Y|X = x_2, \eta < \min\{x_1'\theta_0, x_2'\theta_0\}] = (x_1 - x_2)'\theta_0.$$

Let $\eta_i \equiv x_i'\theta_0 - y_i$. Let $c(x_1, x_2; \theta) \equiv \max\{0, (x_1 - x_2)'\theta\}$. Note that the events $\{\eta_i < \min\{x_1'\theta_0, x_2'\theta_0\}\}$ and $\{y_i > c(x_1, x_2; \theta_0)\}$ are equivalent. A least absolute-deviation (LAD) estimator of θ_0 based on these moment restrictions maximizes (2.1) using non-

censored observations for

$$s(V_i, V_i; \gamma, \theta) = -\|(Y_i - Y_j) - (X_i - X_j)'\theta\| \; 1\{Y_i > c(X_i, X_j; \theta) \vee Y_j > c(X_j, X_i; \theta)\}$$

and any γ; cfr. Honoré and Powell (1994). Now let $W = (E, Z)$, $\mathscr{W} = \{0, 1\} \times \mathscr{Z}$, and suppose that

$$Y = E \times \max\{0, X'\theta_0 - \eta\}, \qquad E = 1\{\varrho(Z) \geq \epsilon\},$$

for some function $\varrho(\cdot) : \mathscr{Z} \mapsto \mathscr{R}$ and unobservables (η, ϵ). Suppose that $(\eta, \epsilon) \perp (X, Z)$. Write F_u for the distribution of η given $\epsilon = u$ and G for the marginal distribution of ϵ. Then

$$\mathcal{E}[Y|X = x, Z = z, Y > 0, E = 1] = x'\theta_0 - \int_{-\infty}^{\varrho(z)} \int_{-\infty}^{x'\theta_0} \frac{\varepsilon_2}{G(\varrho(z)) F_{\varepsilon_1}(x'\theta_0)} \; \mathrm{d}F_{\varepsilon_1}(\varepsilon_2) \; \mathrm{d}G(\varepsilon_1),$$

which depends on z through $\varrho(z)$ only. Hence, the artificial-censoring argument will hold only if $(z_1, z_2) \in \mathscr{Z} \otimes \mathscr{Z}$ are so that $\varrho(z_1) = \varrho(z_2)$. If G is strictly increasing, $\Pr[E = 1|Z = z] = G(\varrho(z))$ and $\varrho(z)$ are one-to-one. Therefore, a local LAD estimator of θ_0 maximizes (2.2) using a nonparametric estimator of the propensity score as $\widehat{\Xi}_i$.

The second illustration concerns endogeneity bias as induced by simultaneity or by measurement error in covariates, for example, in the context of a rank estimator for nonlinear index models.

Example 2.2. Let $W = (E, Z)$, $\mathscr{W} = \mathscr{E} \times \mathscr{Z}$, where E is a subset of X. Consider the triangular simultaneous-equation model

$$Y = \phi(X'\beta(\theta_0), \eta), \qquad E = \mu_E(Z) + \epsilon,$$

for a coefficient vector $\beta(\theta) \equiv (1, \theta')'$ and latent disturbances (η, ϵ). Here, $\phi \equiv \phi_2 \circ \phi_1$ is a composite link function, with $\phi_1(\cdot)$ strictly increasing and $\phi_2(\cdot, \cdot)$ weakly increasing. On assuming that $\eta \perp X$, the model becomes a single-index model and monotonicity of $\phi(\cdot)$ implies the sign restriction

$$\mathcal{E}[Y|X = x_1] > \mathcal{E}[Y|X = x_2] \Rightarrow x_1'\beta(\theta_0) > x_2'\beta(\theta_0)$$

for any (x_1, x_2) in $\mathscr{X} \otimes \mathscr{X}$. Cavanagh and Sherman (1998) proposed estimating θ_0 by the maximizer of (2.1) with

$$s(V_i, V_j; \gamma, \theta) = \frac{1}{2} m(Y_i) \; 1\{X_i'\beta(\theta) > X_j'\beta(\theta)\} + \frac{1}{2} m(Y_j) \; 1\{X_i'\beta(\theta) < X_j'\beta(\theta)\}$$

for some increasing function $m(\cdot) : \mathscr{Y} \mapsto \mathscr{R}$ and any γ. Now relax the independence between the unobservables by demanding only that the distribution of η given $X = x, Z = z$ varies with x, z only through the value they induce on ϵ; write F_u for the distribution of η given $\epsilon = u$. Then

$$\mathcal{E}[m(Y)|X = x, Z = z] = \int m\{\phi(x'\beta(\theta_0), \varepsilon)\} \; \mathrm{d}F_u(\varepsilon) = \mu_{m(Y)}(x'\beta(\theta_0), u) \quad \text{(say)},$$

for $u = e - \mu_E(z)$, and so

$$\mu_{m(Y)}(x_1'\beta(\theta_0), u) > \mu_{m(Y)}(x_2'\beta(\theta_0, u)) \Rightarrow x_1'\beta(\theta_0) > x_2'\beta(\theta_0).$$

The realizations of ϵ can be estimated by the residuals from a non- or semi-parametric regression. This leads to a local version of the rank estimator of Cavanagh and Sherman (1998) with $\Xi_i = E_i - \mu_E(Z_i)$.

The last example introduces a three-step estimator of the link function in a partially-linear transformation model under conditional-independence restrictions.

EXAMPLE 2.3. With $\phi(\cdot) : \mathscr{Y} \mapsto \mathscr{R}$ strictly increasing, a generic formulation of the linear tranformation model is $\phi(Y) = X'\beta(\gamma_0) + \eta$, where $\eta \sim F$, independent of X, and $\beta(\gamma) = (1, \gamma')'$. Note that $\Pr[Y \leq y|X = x] = F(\phi(y) - x'\beta(\gamma_0))$. Fix two values of Y in $\mathscr{Y}$, (y, y_0). Then, for any (x_1, x_2) in $\mathscr{X} \otimes \mathscr{X}$,

$$\Pr[Y \geq y|X = x_1] \geq \Pr[Y \geq y_0|X = x_2] \Rightarrow (x_1 - x_2)'\beta(\gamma_0) \geq \phi(y) - \phi(y_0).$$

Following Chen (2002), an estimator of $\theta_0 = \phi(y) - \phi(y_0)$ maximizes (2.1) for

$$s(V_i, V_j; \widehat{\gamma}, \theta) = \frac{1}{2} r(Y_i, Y_j)\ 1\{(X_i - X_j)'\beta(\widehat{\gamma}) \geq \theta\} + \frac{1}{2} r(Y_j, Y_i)\ 1\{(X_j - X_i)'\beta(\widehat{\gamma}) \geq \theta\},$$

where $r(Y_i, Y_j) \equiv 1\{Y_i \geq y\} - 1\{Y_j \geq y_0\}$ and $\widehat{\gamma}$ is a first-step estimator of γ_0. Now relax the linear-index specification by considering the augmented transformation model

$$\phi(Y) = X'\beta(\gamma_0) + \varrho(\Xi) + \eta$$

for an unknown but smooth function $\varrho(\cdot) : \mathscr{R}^{\mathrm{d}(\Xi)} \mapsto \mathscr{R}$. Replace the full independence assumption by demanding that $X \perp \eta | \Xi = \xi$, a.e. ξ. From smoothness of $\varrho(\cdot)$, it follows that $\varrho(\xi_1) - \varrho(\xi_2) \to 0$ as $\xi_1 - \xi_2 \to 0$. Consequently,

$$\Pr[Y \geq y|X = x_1, \Xi = \xi] \geq \Pr[Y \geq y_0|X = x_2, \Xi = \xi] \Rightarrow (x_1 - x_2)'\beta(\gamma_0) \geq \theta_0$$

An estimator that fits into (2.2) readily follows. Note that it does not require estimating the function $\varrho(\cdot)$. Observe also that the objective function requires a consistent plug-in estimate of γ_0. The estimator discussed in Example 2.2 will do for this purpose; see also Abrevaya and Shin (2011) for a maximum rank-correlation approach to estimate γ_0 when Ξ_i is univariate and observable. An alternative estimator of γ_0 is presented in Blundell and Powell (2004).

3. DISTRIBUTION THEORY

In the sequel, I will assume that all elements of Ξ are continuous. This is strictly a matter of convenience, as discrete conditioning variables can be tackled by including indicators for the equality of two of its realizations. This would cause no additional theoretical complications. The notation $\overset{\frown}{}$ will refer to convergence in probability of a random variable, $\rightsquigarrow$ will mean convergence in distribution, and $\|A\|$ will indicate the Euclidean norm when A is a vector and the matrix norm when A is a matrix.

By virtue of symmetry, $\widehat{Q}_n(\theta; \gamma) = (n(n-1))^{-1} \sum_{i \neq j} \widehat{q}_n(D_i, D_j; \gamma, \theta)$ for

$$\widehat{q}_n(D_i, D_j; \gamma, \theta) \equiv \sigma_k^{-\mathrm{d}(\Xi)} s(V_i, V_j; \gamma, \theta)\ k\Big(\frac{\widehat{\Xi}_i - \widehat{\Xi}_j}{\sigma_k}\Big)\ t_i t_j.$$

Because bias induced by kernel weighting can be dealt with under familiar regularity- and smoothness conditions on the kernels and conditional expectations involved, the largest chunk of our endeavors will be devoted to establishing the impact of estimation error in the control on the asymptotic variance of $\widehat{\theta}$. In doing so, it will be useful to interpret $\widehat{Q}_n(\theta; \gamma)$ as an approximation to $Q_n(\theta; \gamma) \equiv (n(n-1))^{-1} \sum_{i \neq j} q_n(D_i, D_j; \gamma, \theta)$,

$$q_n(D_i, D_j; \gamma, \theta) \equiv \sigma_k^{-\mathrm{d}(\Xi)} s(V_i, V_j; \gamma, \theta)\ k\Big(\frac{\Xi_i - \Xi_j}{\sigma_k}\Big)\ t_i t_j,$$

which, apart from the presence of trimming, would be the objective function of choice if the control were directly observable.

Under the assumptions postulated below, the expectation of $q_n(D_i, D_j; \gamma, \theta)$ is well defined. Let

$$\tau_n(d; \gamma, \theta) \equiv \mathcal{E}[q_n(d, D; \gamma, \theta)], \qquad \tau(d; \gamma, \theta) \equiv \lim_{n\to\infty} \tau_n(d; \gamma, \theta),$$

and note that $\mathcal{E}[q_n(d, D; \gamma, \theta)] = \mathcal{E}[q_n(D, d; \gamma, \theta)]$ by symmetry. While both functions are non-stochastic, $\tau_n(d; \gamma, \theta)$ depends on the sample size through σ_k, a consequence of the presence of kernel weights in $q_n(\cdot, \cdot; \gamma, \theta)$. The limit objective function for our problem then is $Q(\theta; \gamma) \equiv \mathcal{E}[\tau(D; \gamma, \theta)]$.

3.1. Maintained assumptions

The following high-level conditions will be imposed throughout.

ASSUMPTION 3.1. *The function $Q(\theta; \gamma_0)$ is continuous on Θ and uniquely maximized at $\theta = \theta_0$.*

This assumption essentially imposes that θ_0 can be identified as a functional of $Q(\theta; \gamma_0)$. Precise conditions for Assumption 3.1 to hold will depend on the problem at hand. Section 4 contains an example.

The following requirement is standard in the analysis of nonlinear problems with non-concave objective functions.

ASSUMPTION 3.2. *The space Θ is compact and θ_0 is interior to it.*

The class of estimators to which the results in this section will apply are those that satisfy a Euclidean condition.

ASSUMPTION 3.3. *The class $\mathcal{S}$ on $\mathscr{D} \otimes \mathscr{D}$ is Euclidean for an envelope* S *whose second moment is finite.*

Pakes and Pollard (1989) provide a definition and many illustrations of functions that satisfy this condition. The classes introduced in the examples above, for instance, are all Euclidean. Assumption 3.3 is a basic building block in the analysis of estimators that solve non-smooth optimization problems. It can roughly be thought of as replacing a Hölder-continuity condition on the objective function in conventional estimation procedures; compare with Assumption 9 in Aradillas-López et al. (2007), for example. As a notable example, indicator functions are not Hölder continuous.

The next assumption demands $k(\cdot)$ to be a symmetric higher-order kernel.

ASSUMPTION 3.4. *For some positive integer κ, $k(\cdot)$ is a symmetric kernel function of order κ. Furthermore, $k(\cdot)$ is bounded and is twice differentiable with bounded derivatives $k'(\cdot)$ and $k''(\cdot)$.*

The use of a higher-order kernel is motivated by the desire to control for bias induced by kernel weighting. As Assumption 3.3 allows for non-concave $s(\cdot, \cdot; \gamma, \theta)$, working with

a higher-order kernel does not introduce any additional complexity.[2] Although it can be relaxed, imposing symmetry on $k(\cdot)$ is natural given that the weight that is assigned to the score contribution of a pair of observations should not depend on the order in which they enter $s(\cdot,\cdot;\gamma,\theta)$. It also facilitates the construction of a higher-order kernel.

To ensure trimming does not have an adverse effect on the behavior of the objective function, maintain Assumption 3.5.

ASSUMPTION 3.5. *The trimming term is bounded by* T*, which is square-integrable.*

Define the class $\mathcal{K} \equiv \{k((\cdot - \cdot)/\sigma) : \sigma > 0)\}$ on $\mathscr{R}^{\mathrm{d}(\Xi)} \otimes \mathscr{R}^{\mathrm{d}(\Xi)}$. By Lemma 22(ii) in Nolan and Pollard (1987), this class is Euclidean for the constant envelope $\mathrm{K} \equiv \sup_{\varepsilon \in \mathscr{R}^{\mathrm{d}(\Xi)}} \|k(\varepsilon)\|$. On combining this with Assumptions 3.3 and 3.5, the class

$$\mathcal{Q} \equiv \{\sigma^{\mathrm{d}(\Xi)} q_n(\cdot,\cdot;\gamma,\theta) : \gamma \in \Gamma, \theta \in \Theta, \sigma > 0\}$$

is Euclidean for the envelope $\mathrm{Q} \equiv \mathrm{SKT}^2$ by Lemma 2.14 in Pakes and Pollard (1989). Thus, for appropriate choices of the kernel function and trimming scheme, our objective function inherits the Euclidean properties from its unweighted counterpart.

The analysis will be confined to plug-in estimators of γ_0 that are asymptotically linear.

ASSUMPTION 3.6. *The estimator* $\widehat{\gamma}$ *satisfies*

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nu(D_i;\gamma_0) + o_p(1)$$

for a measurable function $\nu(\cdot;\gamma)$ *so that* $\mathcal{E}[\nu(D;\gamma_0)] = 0$ *and* $\mathcal{E}[\nu(D;\gamma_0)\nu(D;\gamma_0)'] < \infty$.

Most commonly-used estimators satisfy Assumption 3.6, so this requirement is fairly mild. It would be of interest to extend the results given here to situations where γ_0 is infinite dimensional and $\widehat{\gamma}$ is a nonparametric estimator. In Khan (2001), for example, $\widehat{\gamma}$ is a conditional-quantile estimator. However, with the score being allowed to be non-smooth in γ here, such a generalization would require another approach as the one taken here.

3.2. Nonparametric first step

I first consider nonparametric first-step estimation of controls that can be written as deviations from a conditional mean, i.e., with W partitioned as $(E', Z')'$ and $\mathscr{W} = \mathscr{E} \times \mathscr{Z}$,

$$\Xi(w) = g(e) - \mu_{h(E)}(z)$$

for chosen functions $g(\cdot) : \mathscr{E} \mapsto \mathscr{R}^{\mathrm{d}(\Xi)}$, $h(\cdot) : \mathscr{E} \mapsto \mathscr{R}^{\mathrm{d}(\Xi)}$, and associated conditional-mean function $\mu_{h(E)}(\cdot) : \mathscr{Z} \mapsto \mathscr{R}^{\mathrm{d}(\Xi)}$. Verify that the controls in Examples 2.1–2.3 are of this form. Here, $\Xi(\cdot)$ is unknown because $\mu_{h(E)}(\cdot)$ is not specified. While, in principle, any type of nonparametric estimator could be used, I work with a kernel estimator. It is

[2]This is in contrast to Honoré and Powell (2005) and Aradillas-López et al. (2007), where bias is eliminated by a jackknife in order to retain concavity. The jackknife technique could impose a substantial computational cost as it requires estimating θ_0 for a range of different bandwidth values.

given by

$$\widehat{\Xi}(w) = g(e) - \widehat{\mu}_{h(E)}(z) = g(e) - \frac{\sum_{i=1}^n h(E_i)\ \ell\Big(\frac{z-Z_i}{\sigma_\ell}\Big)}{\sum_{i=1}^n \ell\Big(\frac{z-Z_i}{\sigma_\ell}\Big)},$$

for a kernel function $\ell(\cdot) : \mathscr{R}^{\mathrm{d}(Z)} \mapsto \mathscr{R}$ and a smoothing parameter $\sigma_\ell = \sigma_\ell(n)$. The Nadaraya-Watson estimator will prove a convenient choice for our purposes. However, the limit distribution of $\widehat{\theta}$ will not depend on the particular choice for the first-step estimator, so long as it satisfies certain conditions.

The set of admissible controls can be extended to cover other location parameters, such as conditional quantiles for example. The key requirement for our subsequent claims is a uniform rate of convergence and a linear representation result for the corresponding nonparametric estimator. Assumptions 3.7–3.9, in tandem, provide conditions for such a result for $\widehat{\Xi}(w)$, which is stated in Lemma 3.1.

Assumption 3.7. *For some positive integer ℓ, $\ell(\cdot)$ is a symmetric ℓth-order kernel. In addition, $\ell(\cdot)$ is bounded and α-Hölder for some $\alpha > 0$.*

Assumption 3.8. *The bandwidth σ_ℓ is nonnegative and proportional to $n^{-\lambda}$, where $\lambda \in (1/2\ell, (1-\varkappa)/2\mathrm{d}(Z))$ for some $\varkappa > 0$.*

Assumption 3.9. *Let Z have Lebesgue density p_Z and let $\mathscr{Z}_c$ be a compact subset of $\mathscr{Z}$ so that $\inf_{z\in\mathscr{Z}_c} p_Z(z) > 0$ and $\sup_{z\in\mathscr{Z}_c} p_Z(z) < \infty$. Then, for each z in $\mathscr{Z}_c$, $p_Z(z)$ and $\mu_{h(E)}(z)$ are ℓ-times continuously differentiable with bounded derivatives. In addition, $h(E)$ has an envelope whose fourth moment exists and whose conditional variance given $Z = z$ is continuous in z.*

The first of these assumptions again postulates the use of a bias-reducing kernel. As usual, the required kernel order is increasing in the number of regressors.[3] The dimension of Z also affects the speed at which the associated bandwidth is allowed to shrink to zero and the degree of differentiability that is required from its density. In addition to imposing smoothness conditions, Assumption 3.9 introduces a subset of the support of the regressors on which $p_Z(\cdot)$ is known to be bounded away from zero. This is a technical requirement that prevents the denominator of the first-step estimator from getting arbitrarily close to zero. It also prevents $\widehat{\mu}_{h(E)}(z)$ from converging too slowly due to boundary effects.

The first auxiliary result follows.

Lemma 3.1. *Let Assumptions 3.7–3.9 hold. Then*

$$(i)\quad \sup_{z\in\mathscr{Z}_c} \Big\|\widehat{\mu}_{h(E)}(z) - \mu_{h(E)}(z)\Big\| = \mathcal{O}_p\Big(\sqrt{\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}}\Big),\quad \textit{and}$$

$$(ii)\quad \widehat{\mu}_{h(E)}(z) - \mu_{h(E)}(z) = \frac{1}{n\sigma_\ell^{\mathrm{d}(Z)}} \sum_{i=1}^n \frac{[h(E_i) - \mu_{h(E)}(z)]}{p_Z(z)} \ell\Big(\frac{z-Z_i}{\sigma_\ell}\Big) + \mathcal{O}_p\Big(\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}\Big)$$

[3] The presence of discrete regressors would require a rewriting of Assumption 3.9 in terms of conditional densities and a corresponding adjustment to the kernel function, but would not complicate the analysis beyond notational inconvenience.

uniformly over $\mathscr{Z}_c$.

To accompany Lemma 3.1, strenghten Assumption 3.5 so that only observations for which Z_i lies in $\mathscr{Z}_c$ are retained.

ASSUMPTION 3.10. *The trimming term takes the form* $t_i = t(Z_i) = 1\{Z_i \in \mathscr{Z}_c\}\ \bar{t}(Z_i)$ *where* $\bar{t}(\cdot) : \mathscr{R}^{\mathrm{d}(Z)} \mapsto \mathscr{R}_0^+$ *is bounded and* ℓ*-times differentiable with bounded derivatives.*

The fixed trimming prescribed here comes at a cost in terms of asymptotic efficiency as it implies that a fraction of the data is ignored asymptotically. It is, however, convenient for proving consistency and asymptotic normality and has been applied elsewhere. Arguably, the analysis below could be adjusted to allow for this fraction to converge to zero slowly with the sample size.

Accomodate the slow convergence rate of the first-step estimator by imposing the following shrinkage rate on the second-step bandwidth.

ASSUMPTION 3.11. *The bandwidth* σ_k *is nonnegative and proportional to* $n^{-\kappa}$, $\kappa \in (1/2\mathsf{k}, (1 - \varkappa - 2\mathrm{d}(Z)\lambda)/2(\mathrm{d}(\Xi) + 2))$.

It is apparent from Assumption 3.11 that the shrinkage speed of the bandwidths σ_ℓ and σ_k are interrelated.

Let p_Ξ be the marginal density of Ξ. Introduce

$$\pi(v, \xi; \gamma, \theta) \equiv \mathcal{E}[s(v, V; \gamma, \theta)t(Z)|\Xi = \xi]\ p_\Xi(\xi).$$

Assume $\pi(v, \xi; \gamma, \theta)$ to be smooth, in the following sense.

ASSUMPTION 3.12. *For each* (γ, θ) *in* $\Gamma \times \Theta$, v *in* $\mathscr{V}$, *and* ξ *in* $\mathscr{R}^{\mathrm{d}(\Xi)}$, *the function* $\pi(v, \xi; \gamma, \theta)$ *is continuous in* γ *and* $(\mathsf{k} + 1)$*-times differentiable in its second argument with uniformly bounded derivatives. Furthermore, the first derivative,* $\nabla_\Xi \pi(v, \Xi(w); \gamma, \theta)$, *is* ℓ*-times differentiable in* z, *and the derivatives are uniformly bounded.*

This differentiability condition, in combination with the previous assumptions, implies that

$$\tau_n(d; \gamma, \theta) = \tau(d; \gamma, \theta) + o(1/\sqrt{n}), \quad \tau(d; \gamma, \theta) = t(z)\pi(v, \Xi(w); \gamma, \theta),$$

uniformly over $\Gamma \times \Theta$.

The above conditions suffice for $\widehat{\theta} = \arg\max_{\theta\in\Theta} \widehat{Q}_n(\theta; \widehat{\gamma}) \rightsquigarrow \arg\max_{\theta\in\Theta} Q(\theta; \gamma_0) = \theta_0$.

THEOREM 3.1. *Let Assumptions 3.2–3.12 hold. Then* $\|\widehat{\theta} - \theta_0\| = o_p(1)$.

Theorem 3.1 does not utilize the higher-order smoothness requirements from Assumption 3.12, nor does it require the use of a bias-reducing kernel to hold. However, these latter conditions will be necessary to ensure that $\sqrt{n}(\widehat{\theta} - \theta_0)$ converges to a zero-mean random variable.

Moving on to the asymptotic distribution of $\widehat{\theta}$ requires establishing the impact of the first-step estimation error up to $o_p(1/\sqrt{n})$ and calls for a somewhat more elaborate argument.

For each w in $\mathscr{W}$ and (γ,θ) in $\Gamma\times\Theta$, let

$$\zeta(w;\gamma,\theta)\equiv -t(z)[h(e)-\mu_{h(E)}(z)]'\,\psi(z;\gamma,\theta),\quad \psi(z;\gamma,\theta)\equiv \mathcal{E}[\nabla_{\Xi}\pi(V,\Xi(W);\gamma,\theta)|Z=z].$$

The following lemma shows that $\widehat{Q}_n(\theta;\gamma)$ asymptotically behaves like the sum of two U-statistics and is key in deriving the limit distribution of $\widehat{\theta}$.

Lemma 3.2. *Let Assumptions 3.3–3.4 and 3.7–3.12 hold. Then*

$$\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)=\frac{2}{n}\sum_{i=1}^{n}\zeta(W_i;\gamma,\theta)+\mathcal{o}_p(1/\sqrt{n})$$

uniformly over $\Gamma\times\Theta$.

Recall that $Q_n(\theta;\widehat{\gamma})$ is the infeasible criterion function that uses Ξ as a control. The function $\zeta(\cdot;\gamma,\theta)$ captures the nonnegligible effect of replacing Ξ by a nonparametric estimator. Therefore, Lemma 3.2 shows that we can handle the variation that is induced through the first estimation step separately from the analysis of an infeasible estimator that assumes the control variables to be observable.

The proof of asymptotic normality builds on expansions of each of the U-processes in Lemma 3.2. Two additional assumptions are needed. The first of these relates to the large-n limit of $Q_n(\theta;\gamma)$, the second concerns $\zeta(\cdot;\gamma,\theta)$.

Assumption 3.13. *Let $\mathscr{N}$ denote a convex neighborhood around (γ_0,θ_0). For each d in $\mathscr{D}$ and (γ,θ) in $\mathscr{N}$, all mixed third partial derivatives of $\tau(d;\gamma,\theta)$ exist and there exists an integrable function $\mathcal{M}_\tau(\cdot)$ so that $\|\nabla_{\theta\theta'}\tau(d;\gamma_0,\theta)-\nabla_{\theta\theta'}\tau(d;\gamma_0,\theta_0)\|\leq\mathcal{M}_\tau(d)\;\|\theta-\theta_0\|$. Also, the moments $\mathcal{E}[\|\nabla_\theta\tau(D;\gamma_0,\theta_0)\|^2]$, $\mathcal{E}[\|\nabla_{\theta\theta'}\tau(D;\gamma_0,\theta_0)\|]$, and $\mathcal{E}[\|\nabla_{\theta\gamma'}\tau(D;\gamma_0,\theta_0)\|]$ are finite, and the matrix $\mathcal{E}[\nabla_{\theta\theta'}\tau(D;\gamma_0,\theta_0)]$ is negative definite.*

Assumption 3.14. *For each w in $\mathscr{W}$ and (γ,θ) in $\mathscr{N}$, all mixed third partial derivatives of $\zeta(w;\gamma,\theta)$ exist and $\|\nabla_{\theta\theta'}\zeta(w;\gamma_0,\theta)-\nabla_{\theta\theta'}\zeta(w;\gamma_0,\theta_0)\|\leq\mathcal{M}_\zeta(w)\;\|\theta-\theta_0\|$ for an integrable function $\mathcal{M}_\zeta(\cdot)$. Furthermore, $\mathcal{E}[\|\nabla_\theta\zeta(W;\gamma_0,\theta_0)\|^2]$, $\mathcal{E}[\|\nabla_{\theta\gamma'}\zeta(W;\gamma_0,\theta_0)\|]$, and $\mathcal{E}[\|\nabla_{\theta\theta'}\zeta(W;\gamma_0,\theta_0)\|]$ are all finite.*

Assumptions 3.13 and 3.14 postulate conditions that allow for expansions of $\tau(\cdot;\gamma,\theta)$ and $\zeta(\cdot;\gamma,\theta)$ in a neighborhood of (γ_0,θ_0). Also imposed is the existence of certain moments of the derivatives of $\tau(\cdot;\gamma,\theta)$ and $\zeta(\cdot;\gamma,\theta)$ under P. These assumptions permit the application of a standard law of large numbers and a central limit theorem.

All ingredients are now available to validate the asymptotically-linear representation

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega(D_i;\gamma_0,\theta_0)+\mathcal{o}_p(1)$$

with influence function $\omega(d;\gamma_0,\theta_0)\equiv-\Sigma_{\theta\theta}^{-1}[\nabla_\theta\tau(d;\gamma_0,\theta_0)+\Sigma_{\theta\gamma}\nu(d;\gamma_0)+\nabla_\theta\zeta(d;\gamma_0,\theta_0)]$ for the matrices $\Sigma_{\theta\theta}\equiv\mathcal{E}[\nabla_{\theta\theta'}\tau(D;\gamma_0,\theta_0)]/2$ and $\Sigma_{\theta\gamma}\equiv\mathcal{E}[\nabla_{\theta\gamma'}\tau(D;\gamma_0,\theta_0)]/2$; the factor of one half appears because the objective function is a U-process of order two. On noting that the influence function has zero mean and finite variance under P, we arrive at Theorem 3.2.

THEOREM 3.2. *Let Assumption 3.2–3.14 hold. Then* $\widehat{\theta}$ *satisfies* $\sqrt{n}\|\widehat{\theta}-\theta_0\| = \mathcal{O}_p(1)$ *and*

$$\sqrt{n}(\widehat{\theta}-\theta_0) \rightsquigarrow \mathcal{N}(0,\Omega),$$

for covariance matrix $\Omega \equiv \mathcal{E}[\omega(D;\gamma_0,\theta_0)\omega(D;\gamma_0,\theta_0)']$.

That is, $\widehat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal. Its asymptotic covariance matrix has the usual 'sandwich' form. The influence-function representation is convenient for evaluating the impact of the estimation noise in $\widehat{\gamma}$ and the controls, which are given by $-\Sigma_{\theta\theta}^{-1}\Sigma_{\theta\gamma}\nu(\cdot;\gamma_0)$ and $-\Sigma_{\theta\theta}^{-1}\nabla_\theta\zeta(\cdot;\gamma_0,\theta_0)$, respectively.

On letting $\widetilde{\theta} \equiv \arg\max_{\theta\in\Theta} Q_n(\theta;\widehat{\gamma})$,

$$\sqrt{n}(\widehat{\theta}-\widetilde{\theta}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma_{\theta\theta}^{-1}\nabla_\theta\zeta(D_i;\gamma_0,\theta_0) + o_p(1).$$

The estimator $\widetilde{\theta}$ is applicable to a variety of nonlinear models with observable controls. Robinson (1988) argued that remaining agnostic about how control variables affect the outcome of interest can be useful in preventing misspecification bias. He established $\sqrt{n}$-consistency of a least-squares estimator of θ_0 in a model of the form $Y = X'\theta_0 + \varrho(\Xi)+\eta$, where $\varrho(\xi)$ is replaced by a nonparametric estimator. The results presented here generalize Robinson (1988) to a nonlinear setting, cfr. Examples 2.2 and 2.3. Estimators that maximize an objective function of the form in (2.2) further prevent the need to estimate the nuisance function $\varrho(\cdot)$.

3.3. Semiparametric first step

Suppose that the control function is known up to a finite-dimensional parameter δ_0, i.e., $\Xi(w) = \Xi(w;\delta_0)$, and is estimated as $\widehat{\Xi}_i = \Xi(W_i;\widehat{\delta})$, where $\widehat{\delta}$ satisfies an asymptotic-linearity condition.

ASSUMPTION 3.15. *The estimator* $\widehat{\delta}$ *satisfies*

$$\sqrt{n}(\widehat{\delta}-\delta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\upsilon(W_i;\delta_0) + o_p(1),$$

for a measurable function $\upsilon(\cdot;\delta_0)$ *so that* $\mathcal{E}[\upsilon(W;\delta_0)] = 0$ *and* $\mathcal{E}[\upsilon(W;\delta_0)\upsilon(W;\delta_0)'] < \infty$.

Let the control obey the following smoothness requirement.

ASSUMPTION 3.16. *For each* w *in* $\mathscr{W}$ *and* δ *in an* $\mathcal{O}_p(1/\sqrt{n})$ *neighborhood of* δ_0, $\Xi(w;\delta)$ *is twice differentiable in* δ, *with bounded derivatives.*

With a semiparametrically-specified control, trimming becomes unnecessary.

ASSUMPTION 3.17. *The trimming term* t_i *is equal to the identity for all* i.

The requirements on the second-step bandwidth also become less stringent as now the first step satisfies $\sup_{w\in\mathscr{W}}\|\widehat{\Xi}(w) - \Xi(w)\| = \mathcal{O}_p(1/\sqrt{n})$.

ASSUMPTION 3.18. *The bandwidth σ_k is nonnegative and proportional to $n^{-\kappa}$, $\kappa \in (1/2\mathcal{k}, 1/2(\mathrm{d}(\Xi)+2))$.*

The smoothness requirement that was the topic of Assumption 3.12, finally, can be reduced to the following assumption; the second part of Assumption 3.12 is replaced by Assumption 3.16.

ASSUMPTION 3.19. *For each (γ,θ) in $\Gamma\times\Theta$, v in $\mathscr{V}$, and ξ in $\mathscr{R}^{\mathrm{d}(\Xi)}$, the function $\pi(v,\xi;\gamma,\theta)$ is continuous in γ and $(\mathcal{k}+1)$-times differentiable in its second argument, with uniformly bounded derivatives.*

Given the efforts made so far, it is a small task to show that, with a semiparametric control,

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega(D_i;\gamma_0,\theta_0)+\mathcal{o}_p(1),$$

with influence function $\omega(d;\gamma_0,\theta_0)\equiv -\Sigma_{\theta\theta}^{-1}[\nabla_\theta\tau(d;\gamma_0,\theta_0)+\Sigma_{\theta\gamma}\nu(d;\gamma_0)+\Sigma_{\theta\delta}\upsilon(w;\delta_0)]$, for $\Sigma_{\theta\delta}\equiv\mathcal{E}[\nabla_{\theta\delta'}\tau(D;\gamma_0,\theta_0)]$.

THEOREM 3.3. *Let Assumptions 3.2-3.1, 3.13–3.14, and 3.15–3.19 hold. Then $\widehat{\theta}$ satisfies $\sqrt{n}\|\widehat{\theta}-\theta_0\|=\mathcal{O}_p(1)$ and*

$$\sqrt{n}(\widehat{\theta}-\theta_0)\rightsquigarrow\mathcal{N}(0,\Omega)$$

for covariance matrix $\Omega\equiv\mathcal{E}[\omega(D;\gamma_0,\theta_0)\omega(D;\gamma_0,\theta_0)']$.

4. EXAMPLES

This section presents details on a two-step estimator of monotone linear-index models (cfr. Example 2.2) and on a three-step estimator of the link function in a partially-linear transformation model (cfr. Example 2.3).

4.1. Monotone index models

For brevity I will work under the presumption that γ_0 is absent (or known). Consequently, throughout, I will write $\tau(d;\theta)$ rather than $\tau(d;\gamma,\theta)$, etc. The model is characterized by the sign restriction

$$\mu_{m(Y)}(x_1'\beta(\theta_0),\xi) > \mu_{m(Y)}(x_2'\beta(\theta_0),\xi) \Rightarrow (x_1-x_2)'\beta(\theta_0)>0,\ \text{a.e. } \xi, \tag{4.1}$$

where, recall, $\beta(\theta)=(1,\theta')'$ and $m(\cdot):\mathscr{Y}\mapsto\mathscr{R}$ is increasing. This setup extends the semiparametric control-function approach of Ahn and Powell (1993) and Blundell and Powell (2004) to a large class of nonlinear models.

Recall from Example 2.2 that the score for the estimator by Cavanagh and Sherman (1998) is

$$s(V_i,V_j;\theta)=\frac{1}{2}m(Y_i)\ 1\{(X_i-X_j)'\beta(\theta)>0\}+\frac{1}{2}m(Y_j)\ 1\{(X_j-X_i)'\beta(\theta)>0\}.$$

For each (v_1,v_2) in $\mathscr{V}\otimes\mathscr{V}$, the collection $\{s(v_1,v_2;\theta):\theta\in\Theta\}$ is Euclidean for the envelope $m(\max\{y_1,y_2\})$, and so Assumption 3.3 is satisfied provided that $m(Y)$ has finite variance under P.

The results that follow are readily modified to fit Han's (1987) pioneering and closely-related maximum rank-correlation estimator. This estimator maximizes Kendall's tau between Y and $X'\beta(\theta)$ as a function of θ. Its score is

$$s(V_i, V_j; \theta) = \frac{1}{2} 1\{Y_i > Y_j\}\ 1\{(X_i - X_j)'\beta(\theta) > 0\} + \frac{1}{2} 1\{Y_j > Y_i\}\ 1\{(X_j - X_i)'\beta(\theta) > 0\}$$

and the corresponding class $\mathcal{S}$ on Θ is Euclidean for the constant envelope of unity. I will mention the necessary modifications where needed. Localized versions of the maximum rank-correlation estimator with a semiparametrically-specified or an observable univariate control have also been analyzed in Abrevaya et al. (2010) and Abrevaya and Shin (2011). Proposition 4.2 below contains the asymptotic covariance matrix of these estimators as special cases.

Equation (4.1) bears resemblance to the moment condition behind the estimator of Blundell and Powell (2004) which, on assuming invertibility of $\mu_Y(\cdot, \xi)$, is

$$\mu_Y(x_1'\beta(\theta_0), \xi) = \mu_Y(x_2'\beta(\theta_0), \xi) \Rightarrow (x_1 - x_2)'\beta(\theta_0) = 0, \text{ a.e. } \xi. \tag{4.2}$$

Although (4.1) is weaker than (4.2), the relative asymptotic efficiency of the localized versions of the estimators of Han (1987) and Cavanagh and Sherman (1998) and the estimator of Blundell and Powell (2004) depends on the model at hand.

4.1.1. Identification Point identification of θ_0 can be established under an obvious modification to well known conditions. Throughout this subsection and the next, X is a $\mathrm{d}(X)$-dimensional vector whose first component is continuous; let $\mathcal{X}$ denote the remaining $\mathrm{d}(X) - 1$ components.

PROPOSITION 4.1. *Let the first component of X have an everywhere-positive Lebesgue density given $(\mathcal{X}, \Xi) = (\chi, \xi)$ a.e. and assume that the support of the conditional density of X given $\Xi = \xi$ is not contained in a proper linear subspace of $\mathscr{R}^{\mathrm{d}(X)}$. Then θ_0 is identified and Assumption 3.1 is satisfied.*

It is easy to see that the large-support requirement is sufficient but not necessary for identification. Write $\mathscr{X}_\xi$ for the support of the distribution of X given $\Xi = \xi$. The conditions in Proposition 4.1 imply that the set

$$\Big\{(x_1, x_2) \in \mathscr{X}_\xi \otimes \mathscr{X}_\xi : \operatorname{sign}\{(x_1 - x_2)'\theta\} \neq \operatorname{sign}\{(x_1 - x_2)'\theta_0\}\Big\},$$

has non-zero measure for almost all ξ in $\mathscr{R}^{\mathrm{d}(\Xi)}$ and each θ in Θ except for $\theta = \theta_0$. This leads to point identification. Some form of exclusion restriction will typically be warranted to credibly support the conditions in Proposition 4.1, as is common in control-function problems. Proposition 4.1 extends to Han's (1987) estimator without modification.

4.1.2. Asymptotic variance Let $I \equiv X'\beta(\theta_0)$, $I_i \equiv X_i'\beta(\theta_0)$, and write $p(\iota, \xi)$ for the density of (I, Ξ) at (ι, ξ). Let

$$\mathcal{X}(\chi, z, \iota, \xi) \equiv t(z)\ \mu_{t(Z)}(\iota, \xi)\Big[\chi - \frac{\mu_{t(Z)X}(\iota, \xi)}{\mu_{t(Z)}(\iota, \xi)}\Big], \quad S(y, \iota, \xi) \equiv m(y) - \mu_{m(Y)}(\iota, \xi),$$

where $\mu_{t(Z)}(\cdot)$ and $\mu_{t(Z)X}(\cdot)$ are conditional-mean functions, in obvious notation. The following proposition provides an expression for the components of the influence function $\omega(D_i; \theta_0)$ when Ξ_i is estimated nonparametrically.

PROPOSITION 4.2. *Let $\mu_{m(Y)}(\cdot,\cdot)$ and $p(\cdot,\cdot)$ be differentiable. Then*

$$\begin{aligned}\nabla_\theta \tau(D_i;\theta_0) &= \mathcal{X}(\chi_i, Z_i, I_i, \Xi_i)\ S(Y_i, I_i, \Xi_i)\ p(I_i, \Xi_i),\\ \nabla_\theta \zeta(W_i;\theta_0) &= \mathcal{E}[\mathcal{X}(\chi, Z_i, I, \Xi)\ \nabla_{\Xi'}\mu_{m(Y)}(I,\Xi)\ p(I,\Xi)|Z=Z_i]\ [h(E_i)-\mu_{h(E)}(Z_i)],\\ -\Sigma_{\theta\theta} &= \mathcal{E}[\mathcal{X}(\chi, Z, I, \Xi)\ \mathcal{X}(\chi, Z, I, \Xi)'\ \nabla_I \mu_{m(Y)}(I,\Xi)\ p(I,\Xi)],\end{aligned}$$

provided that the first two moments of $t(Z)$, χ, and $t(Z)\chi$ exist.

On redefining

$$S(y,\iota,\xi) = \mathcal{E}[1\{y > Y\} - 1\{y < Y\}|I=\iota, \Xi=\xi]$$

Proposition 4.2 covers the components of the covariance matrix of a local maximum rank-correlation estimator.

Proposition 4.2 can be used to construct a kernel-based estimator of the asymptotic variance. Consider estimating $p(\iota,\xi)$, $\mathcal{X}(\chi,z,\iota,\xi)$, and $S(y,\iota,\xi)$ and its derivatives by means of nonparametric density and regression estimators. For example, a Rosenblatt-Parzen estimator of $p(I_i,\Xi_i)$ is given by

$$\widehat{p}(\widehat{I}_i, \widehat{\Xi}_i) \equiv \frac{1}{(n-1)}\frac{1}{\sigma_\jmath \sigma_k^{\mathrm{d}(\Xi)}} \sum_{j\neq i} \jmath\Big(\frac{\widehat{I}_i - \widehat{I}_j}{\sigma_\jmath}\Big) k\Big(\frac{\widehat{\Xi}_i - \widehat{\Xi}_j}{\sigma_k}\Big),$$

where $\widehat{I}_i \equiv X_i'\beta(\widehat{\theta})$, $\jmath(\cdot) : \mathscr{R} \mapsto \mathscr{R}$ is a kernel function, and $\sigma_\jmath$ is a bandwidth. These estimators can be combined to arrive at plug-in estimates $\widehat{\omega}(D_i;\widehat{\theta})$, say, of the influence function $\omega(D_i;\theta_0)$ $(i=1,\ldots,n)$.

An alternative covariance-matrix estimator is obtained by working with derivatives of a smoothed objective function. Let $J(\epsilon) \equiv \int_{-\infty}^{\epsilon} \jmath(\varepsilon)\,\mathrm{d}\varepsilon$. The function

$$\widehat{q}(D_i;\widehat{\theta}) \equiv \frac{1}{n-1}\sum_{j\neq i} \frac{m(Y_i)-m(Y_j)}{\sigma_J \sigma_k^{\mathrm{d}(\Xi)}} J\Big(\frac{\widehat{I}_i - \widehat{I}_j}{\sigma_J}\Big) k\Big(\frac{\widehat{\Xi}_i - \widehat{\Xi}_j}{\sigma_k}\Big)\ t_i t_j$$

is a smoothed version of $(n-1)^{-1}\sum_{j\neq i} q(D_i, D_j;\theta)$, up to an inessential constant. An estimator of $\omega(D_i;\theta_0)$ can readily be constructed from the derivatives of $\widehat{q}(D_i;\widehat{\theta})$. Under regularity conditions, $\nabla_\theta \widehat{q}(D_i;\widehat{\theta}) \rightharpoonup \nabla_\theta \tau(D_i;\theta_0)$, $n^{-1}\sum_{i=1}^n \nabla_{\theta\theta'}\widehat{q}(D_i;\widehat{\theta}) \rightharpoonup 2\Sigma_{\theta\theta}$, etc.

4.1.3. Numerical illustrations A simple parametrization of Example 2.2 is the bivariate linear simultaneous-equation model

$$Y = X_1 + E\theta_0 + \eta, \qquad E = (X_1, X_2)\delta_0 + \epsilon,$$

for disturbances (η,ϵ) and regressors (X_1, X_2), drawn as

$$\begin{pmatrix}\eta\\ \epsilon\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_{\eta\epsilon}\\ \rho_{\eta\epsilon} & 1\end{pmatrix}\right], \quad \begin{pmatrix}X_1\\X_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_X\\ \rho_X & 1\end{pmatrix}\right],$$

respectively. I report results for the kernel-weighted pairwise-differenced least-squares estimator of Ahn and Powell (1993), $\widehat{\theta}_{\mathrm{AP1}}$, and its least-absolute deviation counterpart, $\widehat{\theta}_{\mathrm{AP2}}$, the matching estimator of Blundell and Powell (2004), $\widehat{\theta}_{\mathrm{BP}}$, and the local version of the rank estimator of Cavanagh and Sherman (1998) for $m(y) = y$, $\widehat{\theta}_{\mathrm{CS}}$. The first two (correctly) utilize the linearity of the model. The third estimator (correctly) invokes invertibility of the conditional-mean function. Tables 1–2 report on the performance

of the various estimators on the basis of their mean, median, standard deviation, and interquartile range as obtained over 1000 Monte Carlo runs. Throughout, the sample size was fixed at $n = 100$. The first-stage coefficient vector δ_0 was initially set to either $(-1, 1)'$ or $(1, -1)'$, and subsequently rescaled in each simulation run to maintain a value for the concentration parameter of 100, avoiding problems of weak instrumental variables.

To ensure that a proper comparison between the different approaches can be made, all were computed using the same kernel $k(\cdot)$ and bandwidth σ_k, and the same estimates of the first-stage disturbance. Because bias-reducing kernels only start giving worthwhile improvements over kernels of order two for reasonably large samples, $k(\cdot)$ was taken to be the standard-normal density, with its argument scaled down by its empirical standard deviation. The bandwidth was determined via a least-squares cross-validation procedure on a nonparametric regression of Y_i on X_{1i}, E_i, and $\widehat{\Xi}_i$. No trimming was performed. The control was obtained both as the residual from a least-squares regression and as the deviation of E from a nonparametric estimate of its conditional mean. In the latter case, the first-step estimator used a standard-normal kernel together with a cross-validated bandwidth. The kernel arguments were again scaled down by their respective standard deviations.

Table 1 presents the results for the data generating process with $\delta_0 \propto (-1, 1)'$. All the estimators considered perform well. None of them dominates in any category uniformly across the designs. While, not surprisingly, point estimates obtained through $\widehat{\theta}_{\mathrm{BP}}$ and $\widehat{\theta}_{\mathrm{CS}}$ tend to be somewhat more volatile, they often yield small biases coupled to only small increases in standard deviation and interquartile range compared to the estimators that exploit linearity of the model. As expected, using a nonparametric estimator of the first-stage error results in larger standard deviations and interquartile ranges for all estimators.

Table 2 provides the simulation results for the design where $\delta_0 \propto (1, -1)'$. This second parametrization treats all estimators less favorable, except for $\widehat{\theta}_{\mathrm{AP2}}$, whose performance is virtually unaffected. The standard deviation of all other estimators increased compared to Table 1. In terms of root mean-squared error and mean absolute error, as to evaluate estimator risk (not reported), the least absolute-deviation approach is clearly the superior choice. As before, the local-rank estimator tends to keep up with $\widehat{\theta}_{\mathrm{AP1}}$. The estimator that suffers most from the change in δ_0 is $\widehat{\theta}_{\mathrm{BP}}$. Now, it shows substantial bias combined with very high volatility.

The Monte Carlo experiment was repeated for the left-censored simultaneous-equation model

$$Y = \max\{0, X_1 + E\theta_0 + \eta\}, \qquad E = (X_1, X_2)\delta_0 + \epsilon,$$

with covariates and disturbances drawn as before. Because of the nonlinearity in the model, the squared-loss and absolute-deviation estimators, $\widehat{\theta}_{\mathrm{AP1}}$ and $\widehat{\theta}_{\mathrm{AP2}}$, were replaced by their respective censoring-corrected counterparts, $\widehat{\theta}_{\mathrm{HP1}}$ and $\widehat{\theta}_{\mathrm{HP2}}$, based on Honoré and Powell (1994); see Example 2.1. Kernels and bandwidths were chosen as before.

Table 3 contains the results for $\delta_0 \propto (-1, 1)'$. Here, $\widehat{\theta}_{\mathrm{HP1}}$ and $\widehat{\theta}_{\mathrm{HP2}}$ tend to be more biased than both $\widehat{\theta}_{\mathrm{BP}}$ and $\widehat{\theta}_{\mathrm{CS}}$, and are less variable. The largest standard deviation and interquartile range are often found for the local-rank estimator, but not always. The performance of the first three estimators appears sensitive to the specific constellation of the dependence among the errors and the covariates. The bias of the rank estimator is often smallest and much more stable across the variations in $\rho_{\eta\epsilon}$ and ρ_X. It may again

Table 1. Results for the linear triangular model with $\delta_0 \propto (-1, 1)'$.

Semiparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	−.015	−.046	−.104	−.012	−.028	−.046	−.122	−.030
−.5	.5	−.015	−.042	−.098	−.014	−.029	−.047	−.108	−.030
.5	−.5	.017	.041	.002	.021	.014	.049	−.000	.012
.5	.5	.033	.044	.028	.039	.030	.049	.025	.034
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.127	.102	.138	.156	.166	.134	.178	.205
−.5	.5	.130	.104	.142	.161	.171	.136	.179	.207
.5	−.5	.099	.097	.090	.121	.125	.135	.122	.146
.5	.5	.115	.093	.121	.153	.153	.117	.161	.205
Nonparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.049	−.007	.024	.048	.031	−.013	.003	.021
−.5	.5	.045	−.006	.025	.043	.028	−.017	.001	.022
.5	−.5	−.021	−.005	.086	−.015	−.020	.004	.081	−.019
.5	.5	−.026	.005	.126	−.016	−.024	.013	.111	−.024
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.158	.129	.195	.199	.200	.167	.232	.242
−.5	.5	.162	.129	.192	.197	.200	.164	.246	.243
.5	−.5	.101	.119	.109	.123	.128	.159	.148	.157
.5	.5	.126	.119	.157	.171	.161	.151	.199	.220

Note: $n = 100$, 1000 replications.

be observed that using a nonparametric estimator of the control inflates the disperion measures for all approaches.

On turning to the second design for the censored model, Table 4 reveals a similar pattern as did Table 2 for the linear model. Both bias and volatility measures increase for all estimators. The results are again most striking for $\widehat{\theta}_{\text{BP}}$, which tends to be centered around −1 rather than 1 and is very imprecise.

4.1.4. Empirical illustration I used PSID-1975 data to estimate a canonical model of married women's labor supply; see, e.g., Mroz (1987).[4] The outcome variable of interest

[4]PSID: Panel Study of Income Dynamics public use dataset. Produced and distributed by the University of Michigan with primary funding from the National Science Foundation, the National Institute of Aging, and the National Institute of Child Health and Human Development. Ann Arbor, MI.

Table 2. Results for the linear triangular model with $\delta_0 \propto (1, -1)'$.

Semiparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	−.057	−.048	−.434	−.043	−.088	−.051	−.443	−.099
−.5	.5	−.060	−.038	−.418	−.054	−.086	−.039	−.430	−.099
.5	−.5	.373	.045	1.202	.481	.330	.048	−1.931	.353
.5	.5	.121	.048	.127	.151	.097	.051	.027	.104
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.201	.096	.380	.319	.246	.129	.239	.322
−.5	.5	.198	.100	.225	.274	.236	.134	.227	.319
.5	−.5	.452	.095	52.790	.679	.511	.120	1.318	.819
.5	.5	.228	.097	.798	.322	.282	.133	.465	.390
Nonparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	−.081	−.007	−.159	−.074	−.110	−.015	−.354	−.114
−.5	.5	−.080	.002	−.363	−.061	−.105	−.002	−.334	−.115
.5	−.5	.520	.004	−7.889	.608	.477	.005	−3.101	.487
.5	.5	.159	.015	.887	.194	.137	.016	.477	.136
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{AP1}}$	$\widehat{\theta}_{\text{AP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.197	.122	2.499	.280	.249	.159	.331	.322
−.5	.5	.197	.123	1.551	.288	.247	.163	.295	.309
.5	−.5	.487	.110	138.010	.706	.550	.138	3.134	.843
.5	.5	.252	.116	2.351	.379	.314	.154	.808	.444

Note: $n = 100$, 1000 replications.

is the number of hours worked (throughout the year) and the covariates are the log of the wage rate, other household income, the number of kids in the houshold not older than six and older than six, the age of the woman, and her level of education (in years of schooling). The classic concerns here are that women will self-select into the sample by only supplying labor if their (expected) utility from doing so exceeds utility from outside options, and the endogeneity of the wage rate. The sample consists of 753 women; 428 of whom were employed at some point during the year.

I obtained estimates of the (normalized) coefficient vector in the hours-worked equation via least squares, $\widehat{\theta}_{\text{LS}}$, two-stage least squares, $\widehat{\theta}_{\text{2SLS}}$, as well as $\widehat{\theta}_{\text{AP1}}$, $\widehat{\theta}_{\text{BP}}$, and $\widehat{\theta}_{\text{CS}}$ from above. These latter estimators control for both sample selection and endogeneity. All estimates were scaled to lie on the unit hypersphere, which is the most natural

Table 3. Results for the censored triangular model with $\delta_0 \propto (-1, 1)'$.

Semiparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$
−.5	−.5	−.164	−.165	−.062	−.008	−.169	−.166	−.075	−.030
−.5	.5	−.193	−.194	−.104	−.022	−.194	−.197	−.122	−.042
.5	−.5	−.026	−.032	−.019	.037	−.025	−.033	−.026	.024
.5	.5	−.080	−.083	−.009	.050	−.085	−.087	−.017	.031
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$
−.5	−.5	.118	.119	.138	.195	.159	.159	.171	.211
−.5	.5	.124	.128	.169	.181	.159	.171	.215	.227
.5	−.5	.112	.113	.108	.156	.139	.144	.141	.194
.5	.5	.123	.127	.168	.196	.157	.162	.209	.224
Nonparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$
−.5	−.5	−.127	−.127	.051	.033	−.133	−.135	.029	−.001
−.5	.5	−.162	−.160	.083	.036	−.173	−.170	.038	.005
.5	−.5	−.036	−.039	.105	−.007	−.032	−.039	.094	−.016
.5	.5	−.107	−.110	.225	−.006	−.113	−.110	.195	−.018
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$	$\widehat{\theta}_{\mathrm{HP1}}$	$\widehat{\theta}_{\mathrm{HP2}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$
−.5	−.5	.139	.142	.185	.243	.184	.176	.215	.256
−.5	.5	.154	.160	.297	.231	.204	.207	.338	.286
.5	−.5	.123	.126	.142	.160	.152	.156	.188	.189
.5	.5	.140	.145	.271	.214	.180	.181	.328	.261

Note: $n = 100$, 1000 replications.

normalization for applied work.[5] For the choice of instrumental variables I followed Ahn and Powell (1993) and included the region's unemployment rate, a dummy for whether or not the woman lives in a metropolitan area, her experience in the labor market (in years), and the education levels of her mother and father, together with all covariates except for the wage rate. The controls were estimated nonparametrically using the same ingredients as in the simulations above.[6] Indicator functions were used to handle discrete conditioning variables in the Nadaraya-Watson estimators.

[5]The least squares and two-stage least squares estimators included a constant term, but these are not reported.

[6]Note that the implementation of $\widehat{\theta}_{\mathrm{AP1}}$ to deal with the potential endogeneity of the wage rate is different from Ahn and Powell (1993). They used an instrumental-variable procedure while, here, a control-function approach is retained.

Table 4. Results for the censored triangular model with $\delta_0 \propto (1, -1)'$.

Semiparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	−.411	−.418	−1.817	−.077	−.412	−.414	−1.816	−.231
−.5	.5	−.334	−.340	−1.680	−.024	−.333	−.336	−1.064	−.088
.5	−.5	−.251	−.261	−1.985	.578	−.239	−.253	−1.960	.401
.5	.5	−.196	−.200	−1.438	.157	−.198	−.203	−1.605	.099
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.158	.164	.218	.607	.207	.222	.279	.554
−.5	.5	.145	.150	21.527	.343	.199	.2010	.785	.385
.5	−.5	.171	.180	.231	.874	.230	.239	.279	1.042
.5	.5	.154	.159	14.758	.406	.201	.212	1.264	.472
Nonparametric first step									
		Mean bias				Median bias			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	−.410	−.419	−2.069	−.169	−.413	−.426	−2.002	−.275
−.5	.5	−.328	−.334	−1.410	−.049	−.334	−.341	−1.186	−.096
.5	−.5	−.315	−.327	−2.300	.678	−.314	−.328	−2.225	.472
.5	.5	−.246	−.250	−13.718	.182	−.243	−.243	−2.412	.133
		Standard deviation				Interquartile range			
$\rho_{\eta\epsilon}$	ρ_X	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$	$\widehat{\theta}_{\text{HP1}}$	$\widehat{\theta}_{\text{HP2}}$	$\widehat{\theta}_{\text{BP}}$	$\widehat{\theta}_{\text{CS}}$
−.5	−.5	.177	.186	.433	.506	.235	.245	.407	.514
−.5	.5	.176	.181	14.673	.326	.232	.234	1.949	.391
.5	−.5	.196	.206	.408	.898	.262	.280	.456	1.136
.5	.5	.176	.184	329.544	.433	.234	.238	3.753	.528

Note: $n = 100$, 1000 replications.

Table 5 presents the point estimates along with their standard errors as obtained via a nonparametric bootstrap. The least-squares coefficient on log wage is negative, although insignificant. Two-stage least squares finds a strong positive effect of an increase in the wage rate on hours worked. It also reports no significant effects of the number of offspring on labor intensity and asserts that, ceteris paribus, women with higher education are expected to decrease their number of hours worked. This latter effect is significant at the 5% level. Correcting for endogenous sample selection through kernel weighting on the propensity score shows that the number of children does have a negative and statistically significant expected influence on hours. It also results in significantly larger standard errors. Ahn and Powell (1993) justified the large uncertainty through the sensitivity of the results to the selection-equation specification in more parametrized versions of the model (see Mroz, 1987). The estimator by Blundell and Powell (2004) gives a picture similar

Table 5. Labor-force equation results

	Point estimate (Standard error)				
Regressor	$\widehat{\theta}_{\mathrm{LS}}$	$\widehat{\theta}_{\mathrm{2SLS}}$	$\widehat{\theta}_{\mathrm{AP1}}$	$\widehat{\theta}_{\mathrm{BP}}$	$\widehat{\theta}_{\mathrm{CS}}$
log wage	−.048	.979	.540	.995	.577
	(.229)	(.067)	(.320)	(.385)	(.304)
other income	−.012	−.006	−.003	−.010	−.020
	(.012)	(.004)	(.017)	(.006)	(.126)
# kids≤ 6	−.946	−.162	−.791	.091	−.769
	(.144)	(.167)	(.269)	(.578)	(.309)
# kids> 6	−.318	−.040	−.268	.013	−.270
	(.129)	(.051)	(.104)	(.019)	(.097)
age	−.021	−.008	−.014	.000	−.014
	(.019)	(.008)	(.016)	(.009)	(.079)
education	−.040	−.114	−.105	−.047	−.065
	(.068)	(.022)	(.066)	(.020)	(.127)

Note: All estimates rescaled to lie on unit hypersphere; 199 bootstrap replications.

to two-stage least-squares—log wage has a positive and significant effect, the size of the household has no significant impact, and years of schooling has a negative influence—again with more uncertainty around the point estimates. The local-rank estimator, in contrast, yields results that are much in line with $\widehat{\theta}_{\mathrm{AP1}}$, although it does not maintain additive-separability of the latent disturbances in the model. The only covariate that is found to significantly affect the expected number of hours worked at the 5% significance level is the number of young children.

4.2. Transformation models

Transformations $\phi(\cdot) : \mathscr{Y} \mapsto \mathscr{R}$ of an outcome variable have a long history. One popular motivation for their use is the hope that standard regression machinery is more applicable to $\phi(Y)$ than it is directly to Y; a textbook example is a log-linear model. Alternatively, duration data is often analyzed with particular forms of the generic linear transformation model $\phi(Y) = X'\beta(\gamma_0) + \eta$ with $\eta \perp X$; examples include the proportional-hazard model and the GAFT model.

Here, I focus on inferring the transformation $\phi(\cdot)$ (up to location) in the partially-linear model from Example 2.3, restated here for convenience as

$$\phi(Y) = X'\beta(\gamma_0) + \varrho[\Xi] + \eta, \quad \eta \perp X | \Xi = \xi \text{ a.e.}, \quad \beta(\gamma) = (1, \gamma')',$$

for a smooth but unknown function $\varrho(\cdot) : \mathscr{R}^{\mathrm{d}(\Xi)} \mapsto \mathscr{R}$ and $\phi(\cdot)$ invertible, normalized increasing. The additional sources for heterogeneity in this model are substantial. The function $\varrho(\cdot)$ allows for an unspecified impact of the control on Y, and the unobservable η can be correlated with the covariates through Ξ. Nevertheless, I am not aware of any work that deals with nonparametric inference on $\phi(\cdot)$ in this setting.

Example 2.3 showed that, for a fixed (y, y_0) in $\mathscr{Y}$, $\theta_0 = \phi(y) - \phi(y_0)$ can be estimated

via a local version of Chen's (2002) rank estimator. The score for this estimator is

$$s(V_i, V_j; \widehat{\gamma}, \theta) = \frac{1}{2} r(Y_i, Y_j)\ 1\{(X_i - X_j)'\beta(\widehat{\gamma}) \geq \theta\} + \frac{1}{2} r(Y_j, Y_i)\ 1\{(X_j - X_i)'\beta(\widehat{\gamma}) \geq \theta\},$$

with $r(Y_i, Y_j) = 1\{Y_i \geq y\} - 1\{Y_j \geq y_0\}$ and $\widehat{\gamma}$ an asymptotically-linear estimator of γ_0, e.g., the local-rank estimator from the previous subsection. The class $\mathcal{S}$ on $\Gamma \times \Theta$ so formed is Euclidean for the constant envelope of unity. Assumption 3.3 is thus trivially satisfied and the technique falls into our general framework.

4.2.1. Identification Fix (y, y_0) throughout and normalize $\phi(y_0) = 0$. Then $\theta_0 = \phi(y)$ is the unknown parameter value of interest. From Chen (2002) [Assumption 2] it is immediate that the conditions stated in Proposition 4.1 are also sufficient to identify θ_0 in the augmented transformation model.

4.2.2. Asymptotic variance As before, let $I \equiv X'\beta(\gamma_0)$ and write $p(\iota, \xi)$ for the density of (I, Ξ) at (ι, ξ). For any (y_1, y_2), let

$$S_{y_1,y_2}(y, \iota, \xi) \equiv 1\{y < y_1\} - F_\xi(\theta(y_2) - \iota - \varrho[\xi]),$$

where F_ξ is the conditional distribution of η given $\Xi = \xi$. Notice that

$$\mathcal{E}[S_{y,y}(Y, I, \Xi) | I = \iota, \Xi = \xi] = 0$$

for any y. Introduce

$$\mathcal{X}(\iota_1, \iota_2, \xi) \equiv \mathcal{T}(\iota_1, \xi)\ \mathcal{E}[t(Z)\mathcal{X} | I = \iota_2, \Xi = \xi], \qquad \mathcal{T}(\iota, \xi) \equiv \mathcal{E}[t(Z) | I = \iota, \Xi = \xi],$$

and, lastly, let

$$R(\iota, \xi) \equiv f_\xi(-\iota - \varrho[\xi])\ \varrho'[\xi] + \nabla_\Xi F_\xi(-\iota - \varrho[\xi]),$$

where f_ξ is the conditional density of η given $\Xi = \xi$. The asymptotic variance of $\widehat{\theta}$ when the control is estimated nonparametrically is presented in Proposition 4.3.

PROPOSITION 4.3. *Let $p(\cdot, \cdot)$ and $F_\xi(\cdot)$ be differentiable with respect to their arguments. Let $\varrho'(\cdot)$ denote the derivative of $\varrho(\cdot)$. Assume the first moments of $t(Z)$ and $\mathcal{X}$ exist. Then $\nabla_\theta \tau(D_i; \gamma_0, \theta_0)$ is given by*

$$t(Z_i)\big[\mathcal{T}(I_i - \theta_0, \Xi_i) S_{y,y}(Y_i, I_i, \Xi_i) p(I_i - \theta_0, \Xi_i) - \mathcal{T}(I_i + \theta_0, \Xi_i) S_{y_0,y_0}(Y_i, I_i, \Xi_i) p(I_i + \theta_0, \Xi_i)\big].$$

Further,

$$-\Sigma_{\theta\theta} = \int f_\xi(-\iota - \varrho[\xi])\ p(\iota + \theta_0, \xi)\ p(\iota, \xi)\ \mathcal{T}(\iota, \xi)\mathcal{T}(\iota + \theta_0, \xi)\ \mathrm{d}(\iota, \xi),$$

$$\Sigma_{\theta\gamma} = \int f_\xi(-\iota - \varrho[\xi])\ p(\iota + \theta_0, \xi)\ p(\iota, \xi)\big[\mathcal{X}(\iota + \theta_0, \iota, \xi) - \mathcal{X}(\iota, \iota + \theta_0, \xi)\big]'\ \mathrm{d}(\iota, \xi),$$

and $\nabla_\theta \zeta(W_i; \gamma_0, \theta_0) = -t(Z_i)\ \nabla_\theta \psi(Z_i; \gamma_0, \theta_0)\ [h(E_i) - \mu_{h(E)}(Z_i)]$ where $\nabla_\theta \psi(Z_i; \gamma_0, \theta_0)$ takes the form

$$\int \big[\mathcal{T}(\iota - \theta_0, \xi) R(\theta_0 - \iota, \xi)' p(\iota - \theta_0) - \mathcal{T}(\iota + \theta_0, \xi) R(-\iota, \xi)' p(\iota + \theta_0)\big]\ p(\iota, \xi | Z_i)\ \mathrm{d}(\iota, \xi)$$

and $p(\cdot, \cdot | z)$ is the conditional density of the indices given $Z = z$.

Note that the asymptotic variance depends on the estimator of the index coefficients. This is different from the corresponding result in Chen (2002) [Theorem 1]; see Jochmans (2012) for details.

The large-n variance can be estimated as in the previous subsection.

4.2.3. Uniform consistency In the transformation model the object of interest is a function rather than a finite-dimensional parameter; write $\theta_0(y) \equiv \phi(y) - \phi(y_0)$ to make its dependence on y explicit. In light of this, it is of interest to provide a uniform-consistency result. To this end, let $\mathcal{Y} \equiv [\underline{y}, \overline{y}] \subset \mathscr{Y}$ so that $[\theta_0(\underline{y}) - \epsilon, \theta_0(\overline{y}) + \epsilon] \subset \Theta$ for some $\epsilon > 0$ and consider inferring $\{\theta_0(y) : y \in \mathcal{Y}\}$.

Contrary to Chen (2002) [Theorem 1], uniform consistency is not immediate as, here, it is not clear whether the estimator of $\theta_0(\cdot)$ can be guaranteed to be monotonic on $\mathcal{Y}$ for any finite n. The reason is the necessity for $k(\cdot)$ to be a kernel of a higher order than two.[7,8] The following result shows that uniform consistency is attainable if $\theta_0(\cdot)$ is smooth.

PROPOSITION 4.4. *Let $\theta_0(\cdot)$ be continuous on $\mathcal{Y}$. Then* $\sup_{y\in\mathcal{Y}}\|\widehat{\theta}(y) - \theta_0(y)\| = o_p(1)$.

From a practical point of view the possible non-monotonicity of $\widehat{\theta}(\cdot)$ does not bring forth any genuine complication, as the original point estimates may always be rearranged to a monotonic sequence; see Chernozhukov et al. (2009) for this proposal in a generic setting and results on its improvement over the original (non-monotonic) estimator in terms of general L^p norms.

4.2.4. Numerical illustrations To assess the finite-sample performance of $\widehat{\theta}(\cdot)$ it was applied to the model

$$\phi(Y) = I + \Xi + \eta, \qquad \begin{pmatrix} I \\ \Xi \\ \eta \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{I\Xi} & 0 \\ \rho_{I\Xi} & 1 & \rho_{\Xi\eta} \\ 0 & \rho_{\Xi\eta} & 1 \end{pmatrix}\right],$$

with $\phi(y)$ set either to (i) y (linear model; $y_0 = 0$), (ii) $\log(y)$ (log-linear model; $y_0 = 1$), (iii) $\sinh(2y)/13$ (hyperbolic-sine model; $y_0 = 0$), or (iv) $2(\sqrt{y} - 1)$ (a Box-Cox power transform; $y_0 = 1$). The sample size was set to 250 observations. Like before, the estimation was conducted using a cross-validated bandwidth for the conditional mean of Y given the indices. A fourth-order Epanechnikov kernel was used for the weighting but none of the estimated functions were found to be non-monotonic.

Figure 1 shows fifty realizations of $\widehat{\theta}(\cdot)$ (dashed lines), together with $\theta_0(\cdot)$ (solid line),

[7]Suppose that $\theta_0(y_1) < \theta_0(y_2)$. A calculation shows that the estimates, $\widehat{\theta}(y_1)$ and $\widehat{\theta}(y_2)$, must satisfy $\sum_{i=1}^n 1\{y_1 \leq Y_i < y_2\}\big[\mathcal{R}_i(\widehat{\theta}(y_2)) - \mathcal{R}_i(\widehat{\theta}(y_1))\big] \leq 0$, where

$$\mathcal{R}_i(\theta) \equiv \sum_{j\neq i} \frac{1\{X_j'\beta(\widehat{\gamma}) \leq X_i'\beta(\widehat{\gamma}) - \theta\}}{\sigma_k^{\mathrm{d}(\Xi)}} k\Big(\frac{\widehat{\Xi}_i - \widehat{\Xi}_j}{\sigma_k}\Big).$$

Provided that $k(\varepsilon) \geq 0$ for all ε, $\operatorname{sign}\{\mathcal{R}_i(\widehat{\theta}(y_2)) - \mathcal{R}_i(\widehat{\theta}(y_1))\} = \operatorname{sign}\{\widehat{\theta}(y_1) - \widehat{\theta}(y_2)\}$ for all i, leading to monotonicity.

[8]The same problem would arise when using a jackknife approach to bias reduction because it will typically require recomputing the estimator for more than two alternative bandwidth choices. This implies some of the jackknife weights to be negative.

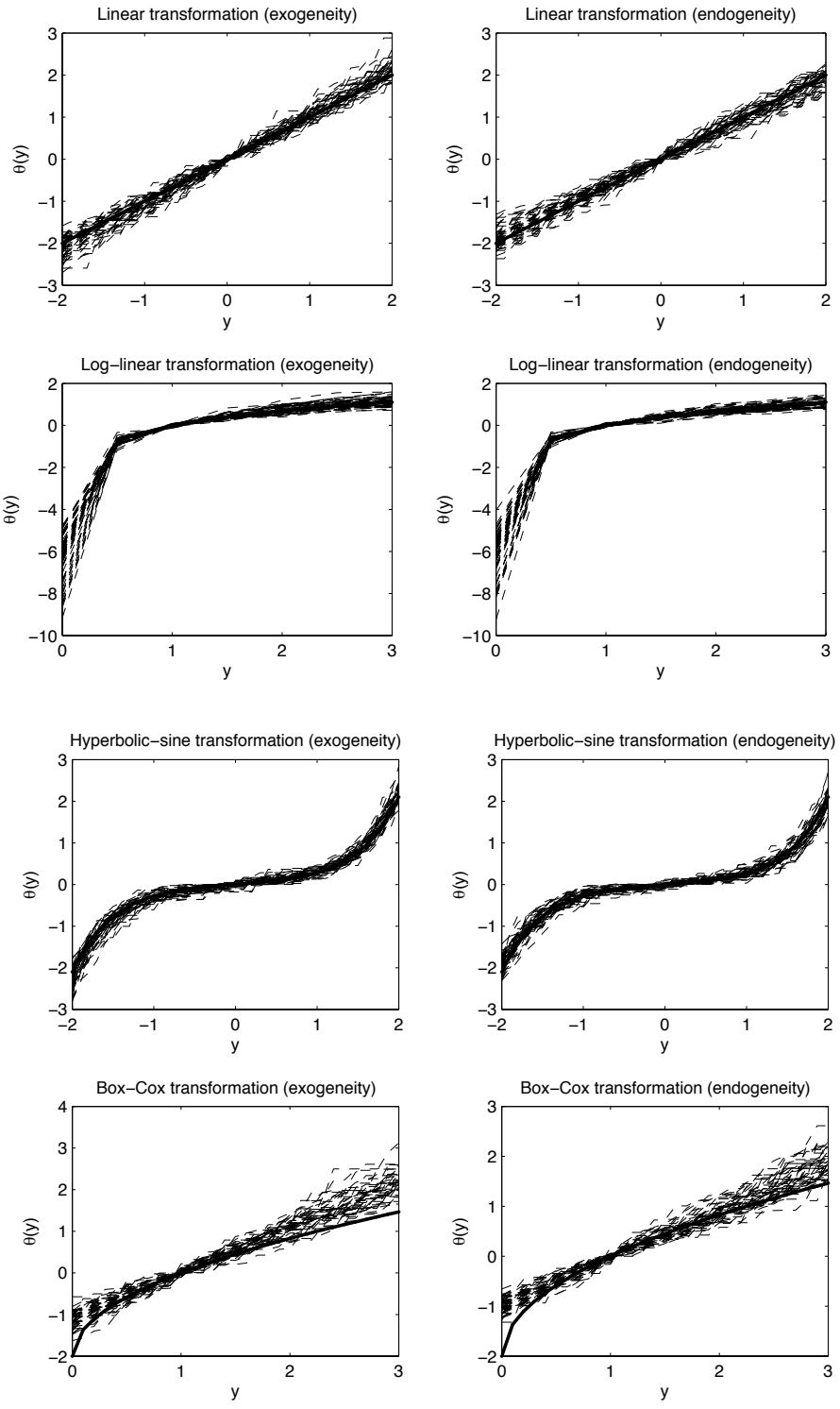

Figure 1. Fifty realizations of $\widehat{\theta}(\cdot)$.

for transformations (i)–(iv). The left panels were constructed from data generated with $\rho_{I\Xi} = -.50$ and $\rho_{\Xi\eta} = 0$, which corresponds to Ξ being an exogenous regressor that is correlated with I. The right panels were obtained from data where $\rho_{I\Xi} = -.50$ and $\rho_{\Xi\eta} = .50$. Here, I and η are dependent unconditional on realizations of Ξ. Overall the estimated link functions closely capture the shape of $\theta_0(\cdot)$. Only with the Box-Cox transformation is there a noticable (upward) bias. These general patterns were confirmed in a larger set of Monte Carlo results, which are omitted here for the sake of brevity.

ACKNOWLEDGEMENTS

I would like to thank Jaap Abbring and two referees for very constructive comments. I am grateful to Manuel Arellano, Richard Blundell, Stéphane Bonhomme, Songnian Chen, Victor Chernozhukov, Aureo de Paula, Geert Dhaene, Iván Fernández-Val, Bryan Graham, Frank Kleibergen, Dennis Kristensen, Arthur Lewbel, Blaise Melly, Jean-Marc Robin, Bernard Salanié, James Stock, and Ke Yang for discussions on this and related work.

REFERENCES

Abrevaya, J., J. A. Hausman, and S. Khan (2010). Testing for causal effects in a generalized regression model with endogenous regressors. *Econometrica* 78, 2043–2061.

Abrevaya, J. and Y. Shin (2011). Rank estimation of partially linear index models. *Econometrics Journal* 14, 409–437.

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–29.

Aradillas-López, A., B. E. Honoré, and J. L. Powell (2007). Pairwise difference estimation with nonparametric control variables. *International Economic Review* 48, 1119–1158.

Blundell, R. W. and J. L. Powell (2004). Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71, 655–679.

Carroll. R. P., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* 92, 477–489.

Cavanagh, C. and R. P. Sherman (1998). Rank estimators for monotonic index models. *Journal of Econometrics* 84, 351–381.

Chen, S. (2002). Rank estimation of transformation models. *Econometrica* 70, 1683–1697.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* 96, 559–575.

Hahn, J. and G. Ridder (2012). The asymptotic variance of semi-parametric estimators with generated regressors. Forthcoming in *Econometrica*.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics* 35, 303–316.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153–161.

Honoré, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 553–565.

Honoré, B. E. and J. L. Powell (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics* 64, 241–278.

Honoré, B. E. and J. L. Powell (2005). Pairwise difference estimators for nonlinear models. In D. W. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models*, 520–553. Cambridge University Press.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics* 159, 252–266.

Jochmans, K. (2012). The variance of a rank estimator of transformation models. *Economics Letters* 117, 168–169.

Khan, S. (2001). Two-stage rank estimation of quantile index models. *Journal of Econometrics* 100, 319–355.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55, 765–799.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–1382.

Nolan, D. and D. Pollard (1987). U-processes: Rates of convergence. *Annals of Statistics* 15, 780–799.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Powell, J. L. (1987). Semiparametric estimation of bivariate latent variable models. Unpublished manuscript.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* 56, 931–954.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123–137.

Sherman, R. P. (1994a). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory* 10, 372–395.

Sherman, R. P. (1994b). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* 22, 439–459.

APPENDIX: PROOFS OF RESULTS

Throughout the Appendix, denote the product measure $P\otimes P$ on the product space $\mathscr{D}\otimes\mathscr{D}$ by $\mathbb{P}$. Write P_n for the empirical measure generated on sampling at random from P and, similarly, let $\mathbb{P}_n$ be the random probability measure placing mass $1/n(n-1)$ on each ordered pair of observations (D_i, D_j). For a measurable function $f(\cdot,\cdot)$ on $\mathscr{D}\otimes\mathscr{D}$, write $\mathcal{E}[f(d,\cdot)] = f(d,P)$ for the expectation of $f(\cdot,\cdot)$ given its first argument and let $\mathbb{E}[f(\cdot,\cdot)] = f(P,P)$. The empirical counterparts are $\mathcal{E}_n[f(d,\cdot)] = f(d,P_n) = n^{-1}\sum_{i=1}^n f(d,D_i)$ and $\mathbb{E}_n[f(\cdot,\cdot)] = (n(n-1))^{-1}\sum_{i\neq j} f(D_i,D_j)$, respectively.

Proof of Lemma 3.1: The result follows from standard kernel-smoothing arguments; see e.g., Aradillas-López et al. (2007). □

Proof of Theorem 3.1: Given random sampling, the construction of Θ, identification of θ_0, and continuity of $Q(\theta;\gamma_0)$, the proof amounts to verifying that

$$\sup_{\theta\in\Theta}\left\|\widehat{Q}_n(\theta;\widehat{\gamma}) - Q(\theta;\gamma_0)\right\| = \mathcal{o}_p(1).$$

The triangle inequality provides the bound

$$\begin{aligned}\left\|\widehat{Q}_n(\theta;\widehat{\gamma}) - Q(\theta;\gamma_0)\right\| \leq &\left\|\widehat{Q}_n(\theta;\widehat{\gamma}) - Q_n(\theta;\widehat{\gamma})\right\| + \left\|Q_n(\theta;\widehat{\gamma}) - \tau_n(P;\widehat{\gamma},\theta)\right\| \\ &+\left\|\tau_n(P;\widehat{\gamma},\theta) - \tau_n(P;\gamma_0,\theta)\right\| + \left\|\tau_n(P;\gamma_0,\theta) - Q(\theta;\gamma_0)\right\|,\end{aligned} \tag{A.1}$$

which holds uniformly over Θ. Hence, it suffices to show that each of these terms converges to zero in probability.

The first right-hand side term in (A.1) captures the noise in the control. Recall that $\widehat{Q}_n(\theta;\widehat{\gamma}) - Q_n(\theta;\widehat{\gamma})$ equals

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{s(V_i,V_j;\widehat{\gamma},\theta)t(Z_i)t(Z_j)}{\sigma_k^{\mathrm{d}(\Xi)}}\Big[k\Big(\frac{\widehat{\Xi}_i-\widehat{\Xi}_j}{\sigma_k}\Big)-k\Big(\frac{\Xi_i-\Xi_j}{\sigma_k}\Big)\Big].$$

Use a mean-value expansion, the symmetry of $\widehat{Q}_n(\theta;\gamma)$ and $Q_n(\theta;\gamma)$, Assumptions 3.3, 3.4, 3.8, and 3.11, and Lemma 3.1(i) to see that, uniformly over Θ,

$$\begin{aligned}\Big\|\widehat{Q}_n(\theta;\widehat{\gamma}) - Q_n(\theta;\widehat{\gamma})\Big\| &= \Big\|\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\frac{s(V_i,V_j;\widehat{\gamma},\theta)t(Z_i)t(Z_j)}{\sigma_k^{\mathrm{d}(\Xi)+1}}\;[\widehat{\Xi}_i-\Xi_i]'\;k'(*)\Big\| \\ &\leq 2\;\mathbb{E}_n[\mathrm{Q}']\;\frac{\sup_{z\in\mathscr{Z}_c}\|\widehat{\mu}_{h(E)}(z)-\mu_{h(E)}(z)\|}{\sigma_k^{\mathrm{d}(\Xi)+1}} \\ &= \frac{1}{\sigma_k^{\mathrm{d}(\Xi)+1}}\mathcal{O}_p\Big(\sqrt{\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}}\Big) = o_p(1),\end{aligned}$$

where $k'(*)$ is $k'(\cdot)$ evaluated in a vector that, elementwise, lies inbetween ${\sigma_k}^{-1}[\widehat{\Xi}_i-\widehat{\Xi}_j]$ and ${\sigma_k}^{-1}[\Xi_i-\Xi_j]$, and $\mathrm{Q}'\equiv\mathrm{SK}'\mathrm{T}^2$ for $\mathrm{K}'\equiv\sup_{\varepsilon\in\mathscr{R}^{\mathrm{d}(\Xi)}}\|k'(\varepsilon)\|$.

The second right-hand side term in (A.1) involves a zero-mean U-process. Because the class $\mathcal{Q}$ is Euclidean for an envelope whose second moment is finite, Corollary 7 of Sherman (1994b) can be applied to get

$$\begin{aligned}\sup_{\theta\in\Theta}\Big\|Q_n(\theta;\widehat{\gamma})-\tau_n(P;\widehat{\gamma},\theta)\Big\| &= \sup_{\theta\in\Theta}\Big\|\mathbb{E}_n\big[q_n(\cdot,\cdot;\widehat{\gamma},\theta)\big]-\mathbb{E}\big[q_n(\cdot,\cdot;\widehat{\gamma},\theta)\big]\Big\| \\ &= \frac{1}{\sigma_k^{\mathrm{d}(\Xi)}}\mathcal{O}_p\Big(\frac{1}{\sqrt{n}}\Big) = o_p(1),\end{aligned}$$

with the last transition following by Assumption 3.11.

For the third right-hand side term in (A.1), recall that $\|\widehat{\gamma}-\gamma_0\| = \mathcal{O}_p(1/\sqrt{n})$ and $\tau_n(d;\gamma,\theta)$ is continuous in γ for all d in $\mathscr{D}$ by Assumptions 3.6 and 3.12, respectively. Therefore, it follows that

$$\sup_{\theta\in\Theta}\Big\|\tau_n(P;\widehat{\gamma},\theta)-\tau_n(P;\gamma_0,\theta)\Big\| = \frac{1}{\sigma_k^{\mathrm{d}(\Xi)}}\mathcal{O}_p\Big(\frac{1}{\sqrt{n}}\Big) = o_p(1),$$

again employing Assumption 3.11.

For the non-stochastic right-hand side term in (A.1), finally, standard kernel-smoothing arguments, as validated by Assumptions 3.4, 3.11, and 3.12 can be used. For a fixed $d=(v,w)$ in $\mathscr{D}$, rewrite $\tau_n(d;\gamma_0,\theta)$ as a kernel-weighted average of $\pi(v,\xi;\gamma_0,\theta)$ under P, i.e.,

$$\tau_n(d;\gamma_0,\theta) = t(z)\int\frac{\pi(v,\xi;\gamma_0,\theta)}{\sigma_k^{\mathrm{d}(\Xi)}}\;k\Big(\frac{\Xi(w)-\xi}{\sigma_k}\Big)\;\mathrm{d}\xi.$$

By a mean-value expansion of $\pi(v,\xi;\gamma_0,\theta)$ around $\Xi(w)$, followed by a change of variable

from ξ to $\varepsilon = \frac{\Xi(w)-\xi}{\sigma_k}$,

$$\sup_{\theta\in\Theta}\Big\|\tau_n(d;\gamma_0,\theta)-\tau(d;\gamma_0,\theta)\Big\| = \sigma_k \sup_{\theta\in\Theta}\Big\|t(z)\int \nabla_\Xi\pi(v,*;\gamma_0,\theta)\ \varepsilon\ k(\varepsilon)\ \mathrm{d}\varepsilon\Big\|$$
$$\leq \sigma_k t(z)\int \sup_{\theta\in\Theta}\big\|\nabla_\Xi\pi(v,*;\gamma_0,\theta)\big\|\ \big\|\varepsilon\big\|\ \big\|k(\varepsilon)\big\|\ \mathrm{d}\varepsilon = \mathcal{O}(\sigma_k).$$

Dominated convergence and Assumption 3.11 then yield

$$\sup_{\theta\in\Theta}\Big\|\tau_n(P;\gamma_0,\theta)-\tau(P;\gamma_0,\theta)\Big\| = \mathcal{O}(\sigma_k) = o(1),$$

which completes the proof. □

Proof of Lemma 3.2: The point of departure is a second-order expansion of each $\widehat{q}_n(D_i,D_j;\gamma,\theta)$ around the scaled difference ${\sigma_k}^{-1}[\Xi_i-\Xi_j]$. Average across all pairs and exploit symmetry to write

$$\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma) = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\frac{s(V_i,V_j,\theta)t(Z_i)t(Z_j)}{\sigma_k^{\mathrm{d}(\Xi)+1}}\ [\widehat{\Xi}_i-\Xi_i]'\ k'\Big(\frac{\Xi_i-\Xi_j}{\sigma_k}\Big)+R_n$$

for a remainder term R_n that captures the contribution of

$$\Big[\frac{\widehat{\Xi}_i-\widehat{\Xi}_j}{\sigma_k}-\frac{\Xi_i-\Xi_j}{\sigma_k}\Big]'k''(*)\Big[\frac{\widehat{\Xi}_i-\widehat{\Xi}_j}{\sigma_k}-\frac{\Xi_i-\Xi_j}{\sigma_k}\Big].$$

The remainder term can be ignored for our purposes because, uniformly over $\Gamma\times\Theta$,

$$R_n \leq 2\ \mathbb{E}_n[\mathrm{Q}'']\ \frac{\sup_{z\in\mathscr{Z}_c}\|\widehat{\mu}_{h(E)}(z)-\mu_{h(E)}(z)\|^2}{\sigma_k^{\mathrm{d}(\Xi)+2}} = o_p\Big(\frac{1}{\sqrt{n}}\Big)$$

for $\mathrm{Q}''\equiv \mathrm{ST}^2\mathrm{K}''$, $\mathrm{K}''\equiv\sup_{\varepsilon\in\mathscr{R}^{\mathrm{d}(\Xi)}}\|k''(\varepsilon)\|$. The inequality follows from the Euclidean properties of the class $\mathcal{S}$ together with Assumptions 3.4 and 3.5. The rate of convergence can be seen to hold on combining Lemma 3.1(i) with Assumption 3.11.

Lemma 3.1(ii) implies that

$$\widehat{\Xi}_i-\Xi_i = -\frac{1}{n\sigma_\ell^{\mathrm{d}(Z)}}\sum_{k=1}^{n}\frac{[h(E_k)-\mu_{h(E)}(Z_i)]}{p_Z(Z_i)}\ell\Big(\frac{Z_i-Z_k}{\sigma_\ell}\Big)+\mathcal{O}_p\Big(\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}\Big)$$

for all Z_i in $\mathscr{Z}_c$. On plugging this expression into $\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)$. and rearranging terms,

$$\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma) = \frac{1}{3}\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j}r_n(D_i,D_j,D_k;\gamma,\theta)+o_p\Big(\frac{1}{\sqrt{n}}\Big),$$

where $r_n(D_i,D_j,D_k;\gamma,\theta)$ is defined as

$$-\frac{s(V_i,V_j;\gamma,\theta)t(Z_i)t(Z_j)}{\sigma_k^{\mathrm{d}(\Xi)+1}\sigma_\ell^{\mathrm{d}(Z)}}\ \frac{[h(E_k)-\mu_{h(E)}(Z_i)]'}{p_Z(Z_i)}k'\Big(\frac{\Xi_i-\Xi_j}{\sigma_k}\Big)\ell\Big(\frac{Z_i-Z_k}{\sigma_\ell}\Big).$$

The influence of the remainder term in the linear representation of $\widehat{\Xi}(w)-\Xi(w)$ on $\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)$ is

$$\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\frac{s(V_i,V_j;\gamma,\theta)t(Z_i)t(Z_j)}{\sigma_k^{\mathrm{d}(\Xi)+1}}k'\Big(\frac{\Xi_i-\Xi_j}{\sigma_k}\Big)\mathcal{O}_p\Big(\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}\Big)$$

and is asymptotically negligible; argue as in the proof of Theorem 3.1 to see that it is bounded by

$$\frac{2\ \mathbb{E}_n[\mathrm{Q}']}{\sigma_k^{\mathrm{d}(\Xi)+1}}\mathcal{O}_p\Big(\frac{n^{\varkappa/2}}{n\sigma_\ell^{\mathrm{d}(Z)}}\Big)=o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Likewise, the contributions of the terms with $k=i$ or $k=j$ to $\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)$ are uniformly bounded by

$$\begin{aligned}\Big\|\binom{n}{2}^{-1}\sum_{i=1}^n\sum_{j\neq i}\frac{r_n(D_i,D_j,D_i;\gamma,\theta)}{n}\Big\| &\leq \frac{\sum_{i=1}^n\sum_{j\neq i}\mathrm{Q}'(D_i,D_j)\ \|h(E_i)-\mu_{h(E)}(Z_i)\|}{(4\ \ell(0))^{-1}n^2(n-1)\sigma_k^{\mathrm{d}(\Xi)+1}\sigma_\ell^{\mathrm{d}(Z)}}\\ &=\mathcal{O}_p\Big(\frac{n^{[\mathrm{d}(\Xi)+1]\kappa+\mathrm{d}(Z)\lambda}}{n}\Big)=o_p\Big(\frac{1}{\sqrt{n}}\Big)\end{aligned}$$

and

$$\begin{aligned}\Big\|\binom{n}{2}^{-1}\sum_{i=1}^n\sum_{j\neq i}\frac{r_n(D_i,D_j,D_j;\gamma,\theta)}{n}\Big\| &\leq \frac{\sum_{i=1}^n\sum_{j\neq i}\mathrm{Q}'(D_i,D_j)\ \|h(E_j)-\mu_{h(E)}(Z_j)\|}{(4\ \ell(0))^{-1}n^2(n-1)\sigma_k^{\mathrm{d}(\Xi)+1}\sigma_\ell^{\mathrm{d}(Z)}}\\ &=\mathcal{O}_p\Big(\frac{n^{[\mathrm{d}(\Xi)+1]\kappa+\mathrm{d}(Z)\lambda}}{n}\Big)=o_p\Big(\frac{1}{\sqrt{n}}\Big),\end{aligned}$$

respectively. A symmetrization argument then yields

$$\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)=\frac{1}{3}\binom{n}{3}^{-1}\sum_{\iota_3}r(D_i,D_j,D_k;\gamma,\theta)+o_p\Big(\frac{1}{\sqrt{n}}\Big),$$

where

$$\begin{aligned}r(D_i,D_j,D_k,\theta)\equiv{}& r_n(D_i,D_j,D_k;\gamma,\theta)+r_n(D_i,D_k,D_j;\gamma,\theta)+r_n(D_j,D_i,D_k;\gamma,\theta)\\ &+r_n(D_k,D_i,D_j;\gamma,\theta)+r_n(D_j,D_k,D_i;\gamma,\theta)+r_n(D_k,D_j,D_i;\gamma,\theta).\end{aligned}$$

and $\iota_3=(i,j,k)$ ranges over the $n(n-1)(n-2)$ ordered triplets of distinct integers from the set $\{1,2,\ldots,n\}$. It is readily verified that the expectation of $\|r(\cdot,\cdot,\cdot;\gamma,\theta)\|^2/n$ is $\mathcal{O}(1)$ so that, up to $o_p(1/\sqrt{n})$,

$$\widehat{Q}_n(\theta;\gamma)-Q_n(\theta;\gamma)=\frac{1}{3}r(P,P,P;\gamma,\theta)+\frac{1}{n}\sum_{i=1}^n\big[r(D_i,P,P;\gamma,\theta)-r(P,P,P;\gamma,\theta)\big]$$

by Lemma A.3 in Ahn and Powell (1993), and it remains only to work out the expectations involved.

Write $r_n^{i,j,k}(\cdot,\cdot,\cdot;\gamma,\theta)$, $r_n^{i,k,j}(\cdot,\cdot,\cdot;\gamma,\theta)$, etc. for the six components of $r(D_i,D_j,D_k;\gamma,\theta)$. The contribution of the first four terms to $r(D_i,P,P;\gamma,\theta)$ can be neglected. Consider the term $r_n^{i,j,k}(D_i,P,P;\gamma,\theta)$, for example. For a random variable A, write P_a for the distribution of D given $A=a$. Notice that, uniformly over $\Gamma\times\Theta$,

$$\begin{aligned}\Big\|r_n^{i,j,k}(D_i,P,P;\gamma,\theta)\Big\|\leq t(Z_i)\ &\Big\|\iint\frac{[h(e)-\mu_{h(E)}(Z_i)]}{p_Z(Z_i)\sigma_\ell^{\mathrm{d}(Z)}}\ \mathrm{d}P_z\ \ \ell\Big(\frac{Z_i-z}{\sigma_\ell}\Big)\ p_Z(z)\ \mathrm{d}z\Big\|\\ \times\quad &\Big\|\iint\frac{-s(V_i,v;\gamma,\theta)t(z)}{\sigma_k^{\mathrm{d}(\Xi)+1}}\ \mathrm{d}P_\xi\ k'\Big(\frac{\Xi_i-\xi}{\sigma_k}\Big)\ p_\Xi(\xi)\ \ \mathrm{d}\xi\Big\|.\end{aligned}$$

By iterated expectations and Assumptions 3.7, 3.8, and 3.9, the first right-hand side term

can be shown to equal

$$\mathrm{t}(Z_i)\Big\| \int \frac{[\mu_{h(E)}(z) - \mu_{h(E)}(Z_i)]}{p_Z(Z_i)\sigma_\ell^{\mathrm{d}(Z)}}\; p_Z(z)\; \ell\Big(\frac{Z_i - z}{\sigma_\ell}\Big)\; \mathrm{d}z\Big\| = \mathcal{O}(\sigma_\ell^l) = o\Big(\frac{1}{\sqrt{n}}\Big)$$

using standard arguments. Next, iterate expectations on the second right-hand side term to see that

$$\sup_{(\gamma,\theta)\in\Gamma\times\Theta} \Big\| r_n^{i,j,k}(D_i, P, P; \gamma, \theta)\Big\| \leq o\Big(\frac{1}{\sqrt{n}}\Big) \sup_{(\gamma,\theta)\in\Gamma\times\Theta} \Big\| \int \frac{\pi(V_i, \xi; \gamma, \theta)}{\sigma_k^{\mathrm{d}(\Xi)+1}}\; k'\Big(\frac{\Xi_i - \xi}{\sigma_k}\Big)\; \mathrm{d}\xi\Big\|.$$

Change variable from ξ to $\varepsilon = \frac{\Xi_i - \xi}{\sigma_k}$ and integrate by parts to find that, uniformly over $\Gamma \times \Theta$,

$$\begin{aligned}\Big\| \int \frac{\pi(V_i, \xi; \gamma, \theta)}{\sigma_k^{\mathrm{d}(\Xi)+1}}\; k'\Big(\frac{\Xi_i - \xi}{\sigma_k}\Big)\; \mathrm{d}\xi\Big\| &= \Big\| \int \nabla_\Xi \pi(V_i, \Xi_i - \varepsilon\sigma_k; \gamma, \theta)\; k(\varepsilon)\; \mathrm{d}\varepsilon\Big\| \\ &= \Big\| \nabla_\Xi \pi(V_i, \Xi_i; \gamma, \theta)\Big\| + \mathcal{O}(\sigma_k^\kappa),\end{aligned}$$

which is bounded. Here, the last transition follows again by standard kernel-smoothing arguments and Assumption 3.12. A similar exercise can be done for $r_n^{i,k,j}(D_i, P, P; \gamma, \theta)$, $r_n^{j,i,k}(D_i, P, P; \gamma, \theta)$, and $r_n^{k,i,j}(D_i, P, P; \gamma, \theta)$. It then follows that

$$\sup_{(\gamma,\theta)\in\Gamma\times\Theta} \|r_n^{i,j,k}(D_i, P, P; \gamma, \theta)\| = o\Big(\frac{1}{\sqrt{n}}\Big), \quad \sup_{(\gamma,\theta)\in\Gamma\times\Theta} \|r_n^{j,i,k}(P, D_i, P; \gamma, \theta)\| = o\Big(\frac{1}{\sqrt{n}}\Big),$$

$$\sup_{(\gamma,\theta)\in\Gamma\times\Theta} \|r_n^{i,k,j}(D_i, P, P; \gamma, \theta)\| = o\Big(\frac{1}{\sqrt{n}}\Big), \quad \sup_{(\gamma,\theta)\in\Gamma\times\Theta} \|r_n^{k,i,j}(P, D_i, P; \gamma, \theta)\| = o\Big(\frac{1}{\sqrt{n}}\Big),$$

and are, thus, all asymptotically negligible.

Now turn to the two remaining terms. Iterate expectations and argue as in the previous paragraph to write $r_n^{j,k,i}(P, P, D_i; \gamma, \theta)$ as

$$-\iint t(z) \frac{[h(E_i) - \mu_{h(E)}(z)]'}{\sigma_\ell^{\mathrm{d}(Z)} p_Z(z)} \Big[\nabla_\Xi \pi(v, \Xi(w); \gamma, \theta) + \mathcal{O}(\sigma_k^\kappa)\Big]\; \mathrm{d}P_z\; \ell\Big(\frac{z - Z_i}{\sigma_\ell}\Big)\; p_Z(z)\; \mathrm{d}z.$$

Assumptions 3.7, 3.9, and 3.12 imply that the bias induced by the kernel-weighting is $\mathcal{O}(\sigma_\ell^l) = o(1/\sqrt{n})$, and so

$$\begin{aligned} r_n^{j,k,i}(P, P, D_i; \gamma, \theta) &= -t(Z_i)\big[h(E_i) - \mu_{h(E)}(Z_i)\big]'\; \psi(Z_i; \gamma, \theta) + o_p\Big(\frac{1}{\sqrt{n}}\Big) \\ &= \zeta(W_i; \gamma, \theta) + o_p\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

Because $r_n^{k,j,i}(P, P, D_i; \gamma, \theta)$ has an identical structure,

$$\begin{aligned} r_n^{k,j,i}(P, P, D_i; \gamma, \theta) &= -t(Z_i)\big[h(E_i) - \mu_{h(E)}(Z_i)\big]'\; \psi(Z_i; \gamma, \theta) + o_p\Big(\frac{1}{\sqrt{n}}\Big) \\ &= \zeta(W_i; \gamma, \theta) + o_p\Big(\frac{1}{\sqrt{n}}\Big),\end{aligned}$$

follows immediately from the same reasoning.

On combining all results,

$$\widehat{Q}_n(\theta; \gamma) - Q_n(\theta; \gamma) = -\frac{2}{3} r(P, P, P, \theta) + \frac{2}{n} \sum_{i=1}^n \zeta(W_i; \gamma, \theta) + o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

The proof is complete as $r(P,P,P;\gamma,\theta)=\mathcal{o}(1/\sqrt{n})$; use the fact that $\mathcal{E}[h(E)|Z=z]=\mu_{h(E)}(z)$ to deduce that the dominant term in $\zeta(\cdot;\gamma,\theta)$ has mean zero conditional on $Z=z$. □

Proof of Theorem 3.2: The proof proceeds by approximating $\widehat{Q}_n(\theta;\widehat{\gamma})$ by a smooth function and then applying results of Sherman (1994a) to this approximation. In light of Lemma 3.2, the construction of this expansion is split up in two parts. The first deals with $Q_n(\theta;\widehat{\gamma})$, the second with $\zeta(P_n;\widehat{\gamma},\theta)$.

Define the functions $p_n(\theta;\gamma)\equiv Q_n(\theta;\gamma)-Q_n(\theta_0;\gamma)$ and $p(\theta;\gamma)\equiv\mathbb{E}[p_n(\theta;\gamma)]$ on $\mathscr{N}$. Applying a Hoeffding decomposition to $p_n(\theta;\gamma)$ gives

$$p_n(\theta;\gamma)=p(\theta;\gamma)+2\big[\tau_n(P_n;\gamma,\theta)-\tau_n(P_n;\gamma,\theta_0)-p(\theta;\gamma)\big]+\mathbb{E}_n\big[r_n(\cdot,\cdot;\gamma,\theta)\big].$$

The first two right-hand side terms constitute the projection of $Q_n(\theta;\gamma)$. The remainder term takes the form

$$r_n(D_i,D_j;\gamma,\theta)=b_n(D_i,D_j;\gamma,\theta)-b_n(D_i,P;\gamma,\theta)-b_n(P,D_j;\gamma,\theta)+p(\theta;\gamma)$$

for $b_n(\cdot,\cdot;\gamma,\theta)\equiv q_n(\cdot,\cdot;\gamma,\theta)-q_n(\cdot,\cdot;\gamma,\theta_0)$. Consider the projection term. For each d in $\mathscr{D}$,

$$\tau_n(d;\gamma,\theta)=\tau(d;\gamma,\theta)+\mathcal{o}(1/\sqrt{n})$$

by the arguments in the proof of Lemma 3.2. By Assumption 3.13 and the $\sqrt{n}$-consistency of $\widehat{\gamma}$,

$$\tau_n(d;\widehat{\gamma},\theta)=\tau(d;\gamma_0,\theta)+\nabla_\gamma\tau(d;\gamma_0,\theta)'(\widehat{\gamma}-\gamma_0)+\mathcal{o}_p(1/\sqrt{n})$$

as $\widehat{\gamma}$ lies in $\mathscr{N}$ for n sufficiently large. So,

$$\begin{aligned}\tau_n(d;\widehat{\gamma},\theta)-\tau_n(d;\widehat{\gamma},\theta_0)&=[\tau(d;\gamma_0,\theta)-\tau(d;\gamma_0,\theta_0)]\\&\quad+[\nabla_{\gamma'}\tau(d;\gamma_0,\theta)-\nabla_{\gamma'}\tau(d;\gamma_0,\theta_0)]'(\widehat{\gamma}-\gamma_0)+\mathcal{o}_p(1/\sqrt{n}).\end{aligned}$$

A second-order expansion of $\tau(d;\gamma_0,\theta)$ around θ_0 gives

$$\begin{aligned}\tau(d;\gamma_0,\theta)-\tau(d;\gamma_0,\theta_0)=\;&(\theta-\theta_0)'\nabla_\theta\tau(d;\gamma_0,\theta_0)+\frac{1}{2}(\theta-\theta_0)'\nabla_{\theta\theta'}\tau(d;\gamma_0,\theta_0)(\theta-\theta_0)\\&+\frac{1}{2}(\theta-\theta_0)'\big[\nabla_{\theta\theta'}\tau(d;\gamma_0,*)-\nabla_{\theta\theta'}\tau(d;\gamma_0,*)\big](\theta-\theta_0)\end{aligned}$$

for some $*$ inbetween θ and θ_0; clearly, $*$ lies in $\mathscr{N}$. Similarly,

$$\nabla_{\gamma'}\tau(d;\gamma_0,\theta)-\nabla_{\gamma'}\tau(d;\gamma_0,\theta_0)=(\theta-\theta_0)'\nabla_{\theta\gamma'}\tau(d;\gamma_0,\theta_0)+\mathcal{O}(\|\theta-\theta_0\|^2)$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0. Next, invoke the Lipschitz condition under Assumption 3.13, take expectations, use the asymptotic linearity of $\widehat{\gamma}$, and observe that $\nabla_\theta\tau(P;\gamma_0,\theta_0)=0$ by the first-order conditions to the limiting maximization problem. Then,

$$\begin{aligned}p(\theta;\widehat{\gamma})&=\tau_n(P;\widehat{\gamma},\theta)-\tau_n(P;\widehat{\gamma},\theta_0)\\&=(\theta-\theta_0)'\nabla_{\theta\gamma'}\tau(P;\gamma_0,\theta_0)\,\nu(P_n;\gamma_0)+\frac{1}{2}(\theta-\theta_0)'\nabla_{\theta\theta'}\tau(P;\gamma_0,\theta_0)(\theta-\theta_0)\\&\quad+\mathcal{o}(\|\theta-\theta_0\|^2)+\mathcal{o}_p\Big(\frac{1}{\sqrt{n}}\Big)\end{aligned}$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0. For the second term in the projection, likewise, average $\tau_n(D_i;\widehat{\gamma},\theta)-\tau_n(D_i;\widehat{\gamma},\theta_0)$ across observations, subtract $p(\theta;\widehat{\gamma})$, and invoke

Assumptions 3.6 and 3.13 to see that

$$\tau_n(P_n;\widehat{\gamma},\theta)-\tau_n(P_n;\widehat{\gamma},\theta_0)-p(\theta;\widehat{\gamma})=(\theta-\theta_0)'\nabla_\theta\tau(P_n;\gamma_0,\theta_0)+\mathcal{o}_p(\|\theta-\theta_0\|^2)+\mathcal{o}_p\Big(\frac{1}{\sqrt{n}}\Big)$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0. On collecting terms,

$$\begin{aligned}p_n(\theta;\widehat{\gamma})&=(\theta-\theta_0)'\Sigma_{\theta\theta}(\theta-\theta_0)\\&\quad+\frac{2}{\sqrt{n}}(\theta-\theta_0)'\sqrt{n}\big[\nabla_\theta\tau(P_n;\gamma_0,\theta_0)+\Sigma_{\theta\gamma}\;\nu(P_n;\gamma_0)+\mathcal{o}_p(1)\big]\\&\quad+\mathcal{o}_p(\|\theta-\theta_0\|^2)+\mathbb{E}_n[r_n(\cdot,\cdot;\widehat{\gamma},\theta)]\end{aligned}$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0.

Now turn to $\zeta(P_n;\gamma,\theta)$. Taylor-expand around θ_0 and γ_0, in turn. Then appeal to the Lipschitz condition and the finiteness of the population moments in Assumption 3.14 to dispense with $\nabla_{\theta\theta'}\zeta(P_n,\gamma_0,\theta_0)$. Recall that $\|\widehat{\gamma}-\gamma_0\|=\mathcal{O}_p(1/\sqrt{n})$ and observe that $\nabla_{\theta\gamma'}\zeta(P_n;\gamma_0,\theta_0)\rightsquigarrow\nabla_{\theta\gamma'}\zeta(P;\gamma_0,\theta_0)=0$ by the law of large numbers. Therefore,

$$\zeta(P_n;\widehat{\gamma},\theta)-\zeta(P_n;\widehat{\gamma},\theta_0)=(\theta-\theta_0)'\big[\nabla_\theta\zeta(P_n;\gamma_0,\theta_0)+\mathcal{o}_p(1/\sqrt{n})\big]+\mathcal{o}_p(\|\theta-\theta_0)^2\|)$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0.

Combine the expansions of $p_n(\theta;\widehat{\gamma})$ and $\zeta(P_n;\widehat{\gamma},\theta)-\zeta(P_n;\widehat{\gamma},\theta_0)$ with Lemma 3.2. Then, on rearranging terms and letting

$$\varsigma(d;\gamma_0,\theta_0)\equiv\nabla_\theta\tau(d;\gamma_0,\theta_0)+\Sigma_{\theta\gamma}\nu(d;\gamma_0)+\nabla_\theta\zeta(d;\gamma_0,\theta_0),$$

one obtains

$$\begin{aligned}\widehat{Q}_n(\theta;\widehat{\gamma})-\widehat{Q}_n(\theta_0;\widehat{\gamma})&=(\theta-\theta_0)'\Sigma_{\theta\theta}(\theta-\theta_0)+(\theta-\theta_0)'\frac{2[\sqrt{n}\varsigma(P_n;\gamma_0,\theta_0)+\mathcal{o}_p(1)]}{\sqrt{n}}\\&\quad+\mathcal{o}_p(\|\theta-\theta_0\|^2)+\mathbb{E}_n[r_n(\cdot,\cdot;\widehat{\gamma},\theta)]\end{aligned}$$

uniformly over $\mathcal{o}_p(1)$ neighborhoods of θ_0. Establishing Theorem 3.2 then requires inspection of the remainder term in this approximation. Because of the presence of kernel weights in the objective function, showing $\sqrt{n}$-consistency of $\widehat{\theta}$ requires proceeding in two steps. Once $\sqrt{n}$-consistency has been established, asymptotic normality will follow readily.

First, combine the Euclidean properties of the class $\mathcal{Q}$ with Corollary 17 and Corollary 21 in Nolan and Pollard (1987) to see that the class $\{\sigma_k^{\mathrm{d}(\Xi)}\mathrm{r}_n(\cdot,\cdot;\gamma,\theta):\gamma\in\Gamma,\theta\in\Theta\}$ is Euclidean for an envelope whose second moment under $\mathbb{P}$ exists; denote its bound by R. Further observe that $\mathrm{r}_n(\cdot,\cdot;\gamma,\theta)$ is P-degenerate on $\mathscr{D}\otimes\mathscr{D}$ and that $\mathrm{r}_n(\cdot,\cdot;\gamma,\theta_0)=0$. Apply Theorem 3 in Sherman (1994a) with, using Sherman's notation, $\Theta_n=\Theta$, $\gamma_n=1$, and any $\alpha\in(0,1)$ to get

$$\mathbb{E}_n[\mathrm{r}_n(\cdot,\cdot;\widehat{\gamma},\theta)]=\mathcal{O}_p\Big(\frac{1}{\sigma_k^{\mathrm{d}(\Xi)}n}\Big)=\mathcal{o}_p(1)$$

uniformly over Θ_n. By Theorem 1 of Sherman (1994a), this yields a preliminary convergence rate of $\mathcal{O}_P(1/\sqrt{\sigma_k^{\mathrm{d}(\Xi)}n})$ for $\widehat{\theta}$. Next, reset $\gamma_n=\mathbb{E}_n[\mathrm{R}]/\sigma_k^{\mathrm{d}(\Xi)}$ and let $\delta_n=1/\sqrt{\sigma_k^{\mathrm{d}(\Xi)}n}$. Then, on setting α sufficiently close to unity, by another application of The-

orem 3 of Sherman (1994a), in tandem with Assumption 3.11,

$$\mathbb{E}_n[\mathrm{r}(\cdot,\cdot;\widehat{\gamma},\theta)] = \frac{\mathbb{E}_n[\mathrm{Q}]}{\sigma_k^{\mathrm{d}(\Xi)}}\ \mathcal{O}_p\Big(\frac{(\gamma_n\delta_n)^\alpha}{n}\Big) = o_p\Big(\frac{1}{n}\Big)$$

uniformly over $\mathcal{O}_p(1/\sqrt{\sigma_k^{\mathrm{d}(\Xi)}n}) = o_p(1)$ neighborhoods of θ_0. The $\sqrt{n}$-consistency of $\widehat{\theta}$ then follows from another application of Theorem 1 of Sherman (1994a).

Given $\sqrt{n}$-consistency and the order of the remainder term, asymptotic normality follows from Theorem 2 in Sherman (1994a). □

Proof of Theorem 3.3: The proof mimics the proof of Theorem 3.2. □

Proof of Proposition 4.1: The proof follows from straightforward modifications to the proof of Theorem 1(i) and 1(iii) in Cavanagh and Sherman (1998). □

Proof of Proposition 4.2: The proof follows similar steps as Sherman (1993) [proof of Theorem 4]. Fix θ in Θ. Manipulate the inequalities in the score and iterate expectations to verify that

$$\begin{aligned}\tau(D_i;\theta) &= t(Z_i)\int \big[m(Y_i)-m(y)\big]\ t(z)\ 1\{(X_i-x)'\beta(\theta)>0\}\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)\\ &= t(Z_i)\int\Big[\int_{-\infty}^{\varepsilon(\chi,\theta)}\mu_{t(Z)}(\chi,\iota,\Xi_i)\ S(Y_i,\iota,\Xi_i)\ p_I(\iota|\chi,\Xi_i)\ \mathrm{d}\iota\Big]p_X(\chi|\Xi_i)\ \mathrm{d}\chi\ p_\Xi(\Xi_i)\end{aligned}$$

up to the constant $\mathrm{c}\equiv t(Z_i)\int m(y)\ t(z)\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)$, where $\varepsilon(\chi,\theta)=I_i+(\mathcal{X}_i-\chi)'(\theta-\theta_0)$, and $\mu_{t(Z)}(\cdot)$ is the conditional mean of the trimming function.

The inner integral in the displayed equation above can be differentiated with respect to θ. On doing so, one finds that $\nabla_\theta\tau(D_i;\theta)$ equals

$$t(Z_i)\int \mu_{t(Z)}(\chi,\varepsilon(\chi,\theta),\Xi_i)\ (\mathcal{X}_i-\chi)\ S(Y_i,\varepsilon(\chi,\theta),\Xi_i)\ p_I(\varepsilon(\chi,\theta)|\chi,\Xi_i)\ p_X(\chi|\Xi_i)\ \mathrm{d}\chi\ p_\Xi(\Xi_i).$$

On evaluating at $\theta=\theta_0$ and computing the integral, this yields the expression for $\tau(D_i;\theta_0)$ stated in the proposition.

The form of $\Sigma_{\theta\theta}$ can be derived immediately from the above equation, on differentiating with respect to θ, integrating, and observing that $\mathcal{E}[S(Y,I,\Xi)|I=\iota,\Xi=\xi]=0$.

Recall that $\tau(D_i;\theta)=t(Z_i)\pi(V_i,\Xi_i;\theta)$ and that

$$\zeta(W_i;\theta) = -t(Z_i)[h(E_i)-\mu_{h(E)}(Z_i)]'\mathcal{E}[\nabla_\Xi\pi(V,\Xi;\theta)|Z=Z_i]$$

when the functional form of $\Xi(\cdot)$ is unknown. From above,

$$t(Z_i)\nabla_{\theta\Xi'}\pi(V_i,\Xi(W_i);\theta) = -\mathcal{X}(\mathcal{X}_i,Z_i,I_i,\Xi_i)\ \nabla_{\Xi'}\mu_{m(Y)}(I_i,\Xi_i)\ p(I_i,\Xi_i).$$

The expression for $\nabla_\theta\zeta(W_i;\theta)$ then follows on integrating and rearranging the resulting equation. □

Proof of Proposition 4.3: Observe that

$$\begin{aligned}\tau(D_i;\gamma,\theta) &= t(Z_i)\int(1\{y<y_0\}-1\{Y_i<y\})\ 1\{x'\beta(\gamma)\le X_i'\beta(\gamma)-\theta\}\ t(z)\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)\\ &\quad - t(Z_i)\int(1\{Y_i<y_0\}-1\{y<y\})\ 1\{x'\beta(\gamma)<X_i'\beta(\gamma)+\theta\}\ t(z)\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)\end{aligned}$$

up to c $\equiv \int(1\{Y_i < y_0\} - 1\{y < y\})\ t(Z_i)t(z)\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)$, which does not depend on unknown parameters. Iterate expectations and reverse the order of both integrals to find that, up to c,

$$\tau(D_i;\gamma,\theta) = t(Z_i)\int\Big[\int_{-\infty}^{\varepsilon(\chi,\gamma)+\theta} \mu_{t(Z)}(\chi,\iota,\Xi_i)\ \ S_{y_0,y}(Y_i,\iota,\Xi_i)\ p_I(\iota|\chi,\Xi_i)\ \ \mathrm{d}\iota\Big]\mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)$$
$$- t(Z_i)\int\Big[\int_{-\infty}^{\varepsilon(\chi,\gamma)-\theta} \mu_{t(Z)}(\chi,\iota,\Xi_i)\ \ S_{y,y_0}(Y_i,\iota,\Xi_i)\ p_I(\iota|\chi,\Xi_i)\ \mathrm{d}\iota\Big]\ \mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i)$$

for $\varepsilon(\chi,\gamma) = I_i + (\mathcal{X}_i - \chi)'\gamma$. Here, $\mu_{t(Z)}(\cdot)$ denotes the conditional mean of $t(\cdot)$ and $p_I(\cdot|\cdot,\cdot)$ is the conditional density of I, with the conditioning arguments being obvious.

Using Leibniz's rule to differentiate the integrals between brackets in the displayed equation above with respect to θ, $\nabla_\theta\tau(D_i;\gamma,\theta)$ is found to equal

$$t(Z_i)\int\Big[\mu_{t(Z)}(\chi,\varepsilon(\chi,\gamma)-\theta,\Xi_i)\ S_{y,y_0}(Y_i,\varepsilon(\chi,\gamma)-\theta,\Xi_i)\ p_I(\varepsilon(\chi,\gamma)-\theta|\chi,\Xi_i)$$
$$- \mu_{t(Z)}(\chi,\varepsilon(\chi,\gamma)+\theta,\Xi_i)\ S_{y_0,y}(Y_i,\varepsilon(\chi,\gamma)+\theta,\Xi_i)\ p_I(\varepsilon(\chi,\gamma)+\theta|\chi,\Xi_i)\Big]\mathrm{d}P_{\Xi_i}\ p_\Xi(\Xi_i),$$

again reversing the order of the terms. Evaluate at (γ_0,θ_0) and integrate out $\mathcal{X}$ to find the expression for $\nabla_\theta\tau(D_i;\gamma_0,\theta_0)$ stated in the proposition. Verify that $\nabla_\theta\tau(P;\gamma_0,\theta_0) = 0$ because both

$$\mathcal{E}[S_{y,y}(Y,I,\Xi)|I=\iota,\Xi=\xi] = 0 \quad \text{and} \quad \mathcal{E}[S_{y_0,y_0}(Y,I,\Xi)|I=\iota,\Xi=\xi] = 0$$

hold.

To find the expectation of $\nabla_{\theta\gamma'}\tau(\cdot;\gamma_0,\theta_0)$ under P, differentiate the expression for $\nabla_\theta\tau(d;\gamma,\theta)$ with respect to γ', rearrange, and integrate out d. On using the moment conditions in the above display this gives

$$2\Sigma_{\theta\gamma} = \int f_\xi(\theta_0-\iota-\varrho[\xi])\ p(\iota-\theta_0,\xi)\ p(\iota,\xi)\big[\mathcal{X}(\iota-\theta_0,\iota,\xi) - \mathcal{X}(\iota,\iota-\theta_0,\xi)\big]'\mathrm{d}(\iota,\xi)$$
$$- \int f_\xi(-\iota-\varrho[\xi])\ p(\iota+\theta_0,\xi)\ p(\iota,\xi)\big[\mathcal{X}(\iota+\theta_0,\iota,\xi) - \mathcal{X}(\iota,\iota+\theta_0,\xi)\big]'\ \mathrm{d}(\iota,\xi).$$

In the same fashion,

$$2\Sigma_{\theta\theta} = -\int \mathcal{T}(\iota,\xi)\ \mathcal{T}(\iota-\theta_0,\xi)\ f_\xi(\theta_0-\iota-\varrho[\xi])\ p(\iota-\theta_0,\xi)\ p(\iota,\xi)\ \ \mathrm{d}(\iota,\xi)$$
$$- \int \mathcal{T}(\iota,\xi)\ \mathcal{T}(\iota+\theta_0,\xi)\ f_\xi(-\iota-\varrho[\xi])\ p(\iota+\theta_0,\xi)\ p(\iota,\xi)\ \ \mathrm{d}(\iota,\xi).$$

The expressions of $\Sigma_{\theta\gamma}$ and $\Sigma_{\theta\theta}$ in the proposition then follow from a change of variable from ι to $\iota-\theta_0$ under the first integral in both equations.

Differentiate $\nabla_\theta\tau(d;\gamma_0,\theta_0) = t(z)\pi(v,\xi;\gamma_0,\theta_0)$ with respect to ξ' and take expectations conditional on $Z = Z_i$. Then $\nabla_\theta\zeta(W_i;\gamma_0,\theta_0)$ follows readily. Contrary to $\Sigma_{\theta\theta}$ and $\Sigma_{\theta\gamma}$, however, this expression does not symmetrize because of the conditioning on $Z = Z_i$ in $\nabla_\theta\psi(Z_i;\gamma_0,\theta_0)$. □

Proof of Proposition 4.4: Apply Lemma A.1 in Carroll et al. (1997). □