INTERNATIONAL ECONOMIC REVIEW
Vol. 48, No. 4, November 2007

PAIRWISE DIFFERENCE ESTIMATION WITH NONPARAMETRIC CONTROL VARIABLES*

BY ANDRES ARADILLAS-LOPEZ, BO E. HONORÉ, AND JAMES L. POWELL[1]

Princeton University, U.S.A.; Princeton University, U.S.A.; University of California, Berkeley, U.S.A.

This article extends the pairwise difference estimators for various semilinear limited dependent variable models proposed by Honoré and Powell (*Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* Cambridge: Cambridge University Press, 2005) to permit the regressor appearing in the nonparametric component to itself depend upon a conditional expectation that is nonparametrically estimated. This permits the estimation approach to be applied to nonlinear models with sample selectivity and/or endogeneity, in which a "control variable" for selectivity or endogeneity is nonparametrically estimated. We develop the relevant asymptotic theory for the proposed estimators and we illustrate the theory to derive the asymptotic distribution of the estimator for the partially linear logit model.

1. INTRODUCTION

Using an analogy between the partially linear regression model (Engle et al., 1986; Robinson, 1988) and linear panel data models with fixed effects, Honoré and Powell (2005) showed how partially linear versions of several limited dependent variable models (e.g., logit, censored, and Poisson regression models) could be constructed using the corresponding estimators for panel data versions of these models with individual fixed effects. The resulting estimation method involved "pairwise differences" of observations for which the regressors in the nonparametric component of the regression function are approximately equal, in an analogy of first-differencing to eliminate fixed effects in panel data models. Assuming the regressors in the nonparametric component were either known or linear in parametrically estimated coefficients, the paper derived conditions under which the proposed estimators of the parameters of interest were root-n consistent and asymptotically normal.

Recent work on semiparametric and nonparametric "control function" estimation of nonlinear models with selectivity or endogenous regressors has shown how such models can often be recast as partially linear regression models with

* Manuscript received December 2005; revised February 2007.

[1] The authors are grateful to three anonymous referees for their valuable comments and suggestions. This research is supported by The Gregory C. Chow Econometric Research Program at Princeton University (Aradillas-Lopez and Honoré), the NSF grant No. SES-0417895 and the Danish National Research Foundation, through CAM at The University of Copenhagen (Honoré). Please address correspondence to: Andres Aradillas-Lopez, Department of Economics, Princeton University, Princeton, NJ 08544. E-mail: *aaradill@princeton.edu*.

"control variables" for selection or endogeneity in the nonparametric component of the regression function; these "control variables" are either conditional expectations (e.g., the conditional probability of selection, or "propensity score", as in Ahn and Powell, 1993, to control for selection bias) or residuals from conditional expectations (e.g., first-stage residuals to control for endogenous regressors, as in Newey et al., 1999; Blundell and Powell, 2004) or both (Das et al., 2003). Thus, an obvious approach to estimation of these models would extend the "pairwise difference" estimation method of Honoré and Powell (2005) to accommodate nonparametric estimation methods of the "control variables" appearing in the nonparametric component. This article makes this extension and illustrates how the approach would specialize to a logit model with endogenous regressors.

2. MOTIVATION AND EXAMPLES

Before describing the econometric model of interest, let us denote the vector of observable covariates for the ith observation by $z_i \equiv (v_i, w_i)$, where $v_i \in \mathbb{R}^M$ and $w_i \in \mathbb{R}^L$ may have some elements in common. The rationale behind this partition will become clear below. We will assume throughout that we have an i.i.d. sample $(z_i)_{i=1}^n$ from the population to be described below. The type of econometric models we study have two general features. The first feature is that there is a finite-dimensional parameter of interest $\beta \in \mathbb{R}^K$ and two nuisance parameters: a finite-dimensional vector $\gamma \in \mathbb{R}^D$ and an unknown function $\mu(w_i, \gamma)$, effectively an infinite-dimensional nuisance parameter. The parameter γ is either known or can be estimated by the econometrician. We refer to μ as the "control function," which can be real or vector-valued. We will let β_0 and γ_0 denote the true parameter values. The second feature of the models of interest here is the existence of a function $s(v_i, v_j; \beta)$—which is either known or estimable by the econometrician—such that β_0 is identified as the unique solution to the following problem:

$$\operatorname*{Min}_{\beta} E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0].$$

The specific features of $s(\cdot)$ depend on the model in question. Honoré and Powell (2005) studied the case in which $\mu(w_i, \gamma) = \mu(w_i'\gamma)$. This article relaxes this assumption and studies the general case. We now present a series of examples that fit this general framework. As we shall see, the control function μ may arise naturally in very different contexts including (not necessarily known) transformations of partially linear index models, nonlinear models with endogeneity, as well as some game-theoretic models.

2.1. *Partially Linear Model.* One of the most basic examples of the family of models studied here is a partially linear model, described by $y_i = \beta' x_i + g(w_i) + u_i$, where $g(\cdot)$ is an unknown function and u_i is unobserved but satisfies $E[u_i \mid x_i, w_i] = 0$ almost surely. This yields $y_i - E[y_i \mid w_i] = \beta'(x_i - E[x_i \mid w_i]) + u_i$ a.s. If x_i has full rank and is not deterministic conditional on w_i, it is a well-known result (e.g., Robinson, 1988) that β can be $\sqrt{n}$-consistently estimated by

$$\tilde{\beta} = \left[\sum_{i=1}^{n}(x_i - \hat{E}[x_i \mid w_i])(x_i - \hat{E}[x_i \mid w_i])'\mathbb{I}(w_i)\right]^{-1} \times \sum_{i=1}^{n}(x_i - \hat{E}[x_i \mid w_i])(y_i - \hat{E}[y_i \mid w_i])\mathbb{I}(w_i), \tag{1}$$

where $\hat{E}[x_i \mid w_i]$ is a nonparametric estimator and $\mathbb{I}(w_i)$ is a trimming function, introduced to make the bias of $\hat{E}[x_i \mid w_i]$ disappear uniformly at the same rate. Alternatively, we could reformulate this problem by noting that under the same set of identifying restrictions, with probability one $w_i = w_j$ implies $y_i - y_j = \beta'(x_i - x_j) + u_i - u_j$ and therefore $E[(y_i - y_j - \beta'(x_i - x_j))^2 \mid w_i - w_j = 0]$ is uniquely minimized at $\beta = \beta_0$. This calls for an estimator of the form

$$\hat{\beta} = \operatorname*{argmin}_b \sum_{i<j} K\left(\frac{w_i - w_j}{h}\right)\left[(y_i - y_j) - b'(x_i - x_j)\right]^2, \tag{2}$$

for an appropriately chosen kernel k and bandwidth h. This yields a convenient closed-form estimator,

$$\hat{\beta} = \left[\sum_{i<j} K\left(\frac{w_i - w_j}{h}\right)\Delta x_{ij}\Delta x_{ij}'\right]^{-1} \sum_{i<j} K\left(\frac{w_i - w_j}{h}\right)\Delta x_{ij}\Delta y_{ij}, \tag{3}$$

where $\Delta\xi_{ij} = \xi_i - \xi_j$.

Note that, unlike (1), the estimator proposed requires no trimming function. This is because in this particular example, the control function used (w_i) has a trivial functional form.[2] We will study situations in which the control function used in the procedure analogous to (1) must be estimated in a first stage and w_i is replaced with its estimate $\hat{w}_i$ in a generalized version of (2). For example, the role of w_i could be played by a nonparametric conditional expectation or residual that is not observed but can be uniformly consistently estimated semi or nonparametrically.

2.2. *Other Partially Linear Models.* We now discuss briefly some extensions to the model in Section 2.1. These examples were considered in a more restrictive setting in Honoré and Powell (2005). For illustrative purposes we include here only the partially linear logit and Tobit models. We stress that all the examples mentioned in Section 2 of Honoré and Powell are amenable to the methodology described here. These include the partially linear Poisson regression model and partially linear duration models.

2.2.1. *Partially linear logit model.* The partially linear logit model is given by $y_i = \mathbb{1}\{x_i'\beta + g(w_i) + \varepsilon_i \geq 0\}$, where ε_i is unobserved, independent of $z_i \equiv (x_i, w_i)$ with logistic distribution. We have

[2] If we knew the true functional form for μ, we could reduce the "curse of dimensionality" issues that are present in (3)

$$\Pr(y_i = 1 \mid z_i, z_j, y_i + y_j = 1, w_i - w_j = 0) = \frac{\exp\{(x_j - x_i)'\beta_0\}}{1 + \exp\{(x_j - x_i)'\beta_0\}}, \tag{4}$$

which calls for an estimator of the form

$$\hat{\beta} = \operatorname*{argmin}_b \sum_{\substack{i<j \\ y_i \neq y_j}} K\left(\frac{w_i - w_j}{h}\right) \big[y_i \ln(1 + \exp\{(x_j - x_i)'b\}) + y_j \ln(1 + \exp\{(x_i - x_j)'b\})\big]. \tag{5}$$

As we mentioned above, we shall focus on the case in which w_i is not observed and must be estimated semi or nonparametrically prior to the estimation described in (5).

2.2.2. *Partially linear Tobit model.* The partially linear Tobit model is given by $y_i = \max\{x_i'\beta + g(w_i) + \varepsilon_i\}$. Using the identification insight and results from Honoré (1992) in the context of panel data Tobit models with fixed effects, the framework in Honoré and Powell (2005) suggests an estimator of the form

$$\hat{\beta} = \operatorname*{argmin}_b \sum_{i<j} K\left(\frac{w_i - w_j}{h}\right) q\big(y_i, y_j, (x_i - x_j)'b\big), \tag{6}$$

where $q(\cdot)$ is based on a convex-loss function of the type described in Honoré (1992).

2.3. *Rational Expectations and Interaction-based Models.* Examples of models amenable to our methods also arise in the context of some structural models involving rational expectations with or without strategic interaction across economic agents. We present two examples here.

2.3.1. *Incomplete information games.* Aradillas-Lopez (2006) studies a 2×2 simultaneous game with incomplete information where two players labeled $p = 1, 2$ can choose two actions, labeled $y_p \in \{1, 0\}$. Take $x_1 \in \mathbb{R}^{L_1}, x_2 \in \mathbb{R}^{L_2}$ and denote $x = x_1 \cup x_2$. Players' expected payoffs of choosing $y_p = 1$ are given by $E[U_1 \mid y_1 = 1] = x_1'\delta_1 + \alpha_1 \Pr[Y_2 = 1 \mid x] - \varepsilon_1$ and $E[U_2 \mid y_2 = 1] = x_2'\delta_2 + \alpha_2 \Pr[Y_1 = 1 \mid x] - \varepsilon_2$, where ε_p is independent of (x_1, x_2) with an unknown, but everywhere strictly increasing cdf given by $F_p(\cdot)$. Let $F_p(x) = \Pr(y_p = 1 \mid x)$. Players' beliefs are assumed unobserved to the econometrician. However, if players are expected utility maximizers, Bayesian–Nash equilibrium choice probabilities are[3]

$$\mu_1(x) = F_1(x_1'\delta_1 + \alpha_1\mu_2(x)); \quad \mu_2(x) = F_2(x_2'\delta_2 + \alpha_2\mu_1(x)). \tag{7}$$

Lack of knowledge of $F_p(\cdot)$ requires further parameter normalization. First, we will only be able to estimate (δ_p, α_p) up to a proportionality constant. In addition,

[3] Multiple equilibria concerns are addressed in the paper. We will ignore them here.

an intercept would not be identified. The following normalization will be convenient here. For $p = 1, 2$ we will split $x_p = (w_p, v_p)$, where the coefficient of w_p will be normalized to one. Let $\beta_p = (\delta_p, \alpha_p)$, $z_1 = (v_1', \mu_2(x))$ and $z_2 = (v_2', \mu_1(x))$. We can rewrite (7) as

$$\mu_1(x) = F_1(w_1 + z_1'\beta_1); \quad \mu_2(x) = F_2(w_2 + z_2'\beta_2). \tag{8}$$

Denote $\mu_{p,i} = \mu_p(x_i)$. Using the invertibility properties of the unknown cdf $F_p(\cdot)$, Aradillas-Lopez provides additional conditions on x_p such that $E[((x_{p,i} - x_{p,j})'\delta_p)^2 \mid \mu_{p,i} = \mu_{p,j}]$ is uniquely minimized at $\delta_p = \delta_{p_0} = (1, \beta_{p_0})$. Such conditions involve a simple exclusion restriction and the existence of a continuously distributed covariate with nonzero coefficient. This model has the peculiarity that both the control function $\mu_{p,i}$ and the linear index $x_{p,i}'\delta_p$ involve unknown functions (recall that $x_p'\delta_p = w_p + z_p'\beta_p$, where z_p includes a nonparametric conditional probability). Let $\hat{\mu}_p(x_i) \equiv \hat{\mu}_{1,i}$ denote a nonparametrically estimator of $\mu_{p,i}$ and denote $\hat{\mu}_2(X_i) \equiv \hat{\mu}_{2,i}$ and let $\hat{z}_{1,i} \equiv (v_{1,i}', \hat{\mu}_{2,i})$ and $\hat{z}_{2,i} \equiv (v_{2,i}', \hat{\mu}_{1,i})$. Based on the identification condition stated above, the proposed estimator is

$$\hat{\beta}_p = \operatorname*{argmin}_b \sum_{i<j} K\left(\frac{\hat{\mu}_{p,i} - \hat{\mu}_{p,j}}{h}\right)[(w_{p,i} - w_{p,j}) + (\hat{z}_{p,i} - \hat{z}_{p,j})'b]^2 \phi(x_i)\phi(x_j), \tag{9}$$

where $\phi(\cdot)$ is a trimming function. Note that the control function arises naturally in this model as a consequence of invertibility conditions of $F_p(\cdot)$, the otherwise unknown transformation of the partially linear index $w_p + z_p'\beta_p$. Notice also that the control functions used in (9) are one-dimensional, regardless of the dimension of x. This estimation procedure can be extended to general index models, whose identification conditions are essentially perfectly compatible with the assumptions used in this model.

2.3.2. *Dynamic optimization models.* Consider a model in which agent i solves a dynamic optimization problem of the form

$$\max_{\{q_{it}\}_t} E\left[\sum_{t=0}^{\infty} \delta^t U(x_{it}, s_{it}, q_{it}; \theta) \,\middle|\, \{q_{it}\}_t\right] \text{ subject to } x_{it+1} = \mu(x_{it}, q_{it}) + \xi_{it+1}. \tag{10}$$

Using the terminology of dynamic programming models, q_{it} is agent i's control (not to be confused with the "control function") and x_{it} is the stock. $U(\cdot, \cdot, \cdot; \theta)$ is agent i's per-period utility function and s_{it} is an idiosyncratic shock unobserved by the econometrician, i.i.d. across time and agents, with cdf $F_s(\cdot;\gamma)$ assumed to be known up to the finite-dimensional parameter γ. The agent must choose q_{it} before knowing the realization of s_{it}. The accumulation equation $x_{it+1} = \mu(x_{it}, q_{it}) + \xi_{it+1}$ describes the evolution of the stock variable, with $\mu(\cdot)$ being an unspecified (for the moment) function, increasing in both arguments. ξ_{it+1} is a shock that is i.i.d. across time and individuals and is unobserved at time t. We will assume here that ξ_{it+1} is independent of all other covariates in the model. This model was analyzed in detail by Hong and Shum (2004), who assume a

deterministic accumulation equation of the form $x_{it+1} = x_{it} + q_{it}$. Our assumptions imply

$$x_{it+1} \mid x_{it}, q_{it} \sim x_{it+1} \mid \mu(x_{it}, q_{it}), \tag{11}$$

conditional on $\mu(x_{it}, q_{it})$, x_{it+1} is independent of x_{it}, q_{it}. Note that the evolution shock ξ_{it} does not enter the per-period utility function $U(\cdot, \cdot, \cdot; \theta)$. In a stationary setting, agents' optimal policy functions solve the following Bellman equation for $t = 1, 2, 3, \ldots$:

(12)

$$\max_q U(x_{it}, s_{it}, q; \theta) + \delta \cdot E[V(x_{t+1}, s_{t+1}; \theta, \gamma) \mid x_{it}, s_{it}, \xi_{it}, q], \quad \text{where}$$

$$V(x_{t+1}, s_{t+1}; \theta, \gamma) = \max_{\{q_{i\tau}\}_\tau} E\left[\sum_{\tau=t+1}^{\infty} \delta^{\tau-t-1} U(x_{i,\tau}, s_{i,\tau}, q_\tau; \theta) \,\middle|\, \{q_{i\tau}\}_\tau, x_{t+1}, s_{t+1} \right].$$

Note that this function does not depend on ξ_{t+1} because $E[U(x_{i\tau}, s_{i\tau}, q_\tau; \theta) \mid \{q_{i\tau}\}_\tau, x_{t+1}, s_{t+1}, \xi_{t+1}] = E[U(x_{i\tau}, s_{i\tau}, q_\tau; \theta) \mid \{q_{i\tau}\}_\tau, x_{t+1}, s_{t+1}]$ for all $\tau \geq t+1$. We have

$$\begin{aligned} E[V(x_{t+1}, s_{t+1}; \theta, \gamma) \mid x_t, s_t, \xi_t, q] &= E[V(x_{t+1}, s_{t+1}; \theta, \gamma) \mid x_t, s_t, q] \\ &= E\left[\int V(x_{t+1}, s; \theta, \gamma)\, dF_s(s)\, ds \mid x_t, q \right] \\ &= E\left[\int V(x_{t+1}, s; \theta, \gamma)\, dF_s(s)\, ds \mid \mu(x_t, q) \right] \\ &\equiv \bar{\mathcal{V}}(\mu(x_t, q); \theta, \gamma), \end{aligned} \tag{13}$$

so agent i's optimal policy can be expressed compactly as

$$q(x_{it}, s_{it}; \theta, \gamma) = \operatorname*{argmax}_q U(x_{it}, s_{it}, q; \theta) + \delta \bar{\mathcal{V}}(\mu(x_{it}, q); \theta, \gamma). \tag{14}$$

As in Hong and Shum (2004), the optimal policy function $q(x_t, s_t; \theta, \gamma)$ will be nondecreasing in s_t conditional on x_t if $U(x, s, q; \theta)$ is supermodular in (q, s) given x. This is a useful result because it enables us to recover s_{it} by inverting conditional quantiles of q_{it} given x_{it}. More precisely, for every quantile $\tau \in [0, 1]$ we have $(q \mid x)_\tau = s_\tau$. Therefore, we can estimate s_{it} by $\hat{s}_{it}(\gamma) = F_s^{-1}(\hat{F}(q_{it} \mid x_{it}); \gamma)$, where $\hat{F}(q \mid x)$ is a nonparametric estimator of the conditional cdf of q_t given x_t. Interior solutions to (12) (i.e., those with $q > 0$) satisfy

$$U_{(3)}(x_{it}, s_{it}, q; \theta) + \delta \bar{\mathcal{V}}_{(1)}(\mu(x_{it}, q); \theta, \gamma) \mu_{(2)}(x_{it}, q) = 0, \tag{15}$$

where $f_{(k)}$ denotes the partial derivative of f with respect to its kth argument. Now let us define $g(x, q) = (\mu(x, q), \mu_{(2)}(x, q))$. Notice from (15) that if $q_{it} > 0$

and $q_{jt} > 0$, then $g(x_{it}, q_{it}) = g(x_{jt}, q_{jt})$ implies $U_{(3)}(x_{it}, s_{it}, q_{it}; \theta) - U_{(3)}(x_{jt}, s_{jt}, q_{jt}; \theta) = 0$ when $\theta = \theta_0$. The (vector-valued) control function $g(x, q)$ can be estimated nonparametrically by noting that, given our assumptions,

(16)
$$\mu(x, q) = E[x_{it+1} \mid x_{it} = x, q_{it} = q], \text{ and } \mu_{(2)}(x, q) = \frac{\partial E[x_{it+1} \mid x_{it} = x, q_{it} = q]}{\partial q}.$$

Both objects can be estimated nonparametrically. Following the notation in Hong and Shum, let θ_1 denote the subvector of θ that "survives" after we take the difference $U_{(3)}(x_{it}, s_{it}, q_{it}; \theta) - U_{(3)}(x_{jt}, s_{jt}, q_{jt}; \theta)$ for two observations such that $g(x_{it}, q_{it}) = g(x_{jt}, q_{jt})$. We have a problem now: Both x and q are scalars. If μ is left completely unspecified, then in general x and q could be deterministic conditional on $\mu(x, q)$ and $\mu_{(2)}(x, q)$. Consequently, s would be deterministic too. Whenever $g(x_{it}, q_{it}) = g(x_{jt}, q_{jt})$, we would have $U_{(3)}(x_{it}, s_{it}, q_{it}; \theta) - U_{(3)}(x_{jt}, s_{jt}, q_{jt}; \theta) = 0$ for any θ. For this reason we need to add structure to μ. We will assume that $\mu(x, q) = \alpha_1 x + \alpha_2 q$, with $\hat{\mu}(x, q) = \hat{\alpha}_1 x + \hat{\alpha}_2 q$ being the estimated control function, which can be estimated based on (16) (note that $\mu_{(2)}(x, q)$ is simply a constant now). Letting $\beta = (\theta_1, \gamma)$, the above discussion calls for an estimator of the form[4]

(17)
$$\hat{\beta} = \underset{\gamma, \theta_1}{\operatorname{argmin}} \sum_{t=1}^{T} \sum_{\substack{i<j \\ q_{it}>0 \\ q_{jt}>0}} K\left(\frac{\hat{\mu}(x_{it}, q_{it}) - \hat{\mu}(x_{jt}, q_{jt})}{h}\right) \times [U_{(3)}(x_{it}, \hat{s}_{it}(\gamma), q_{it}; \theta_1) - U_{(3)}(x_{jt}, \hat{s}_{jt}(\gamma), q_{jt}; \theta_1)]^2.$$

Pairwise-differencing allows us to estimate (at least a subset of) θ and γ without having to estimate the value function $\bar{\mathcal{V}}$, which is usually a complex computational task. One would have to undertake this in a second step in order to estimate any parameters that might be in $\theta \setminus \theta_1$ (if there are any such parameters). Hong and Shum describe procedures to do this.

2.4. *The Effect of Pairwise Differencing on Identification.* The discussion at the end of Section 2.3.2 highlights an important issue in pairwise-difference estimation procedures. This is the fact that conditioning on the control variable might effectively "wipe out" some of the parameters of interest from the moment condition used for estimation. This is particularly clear in closed-form estimators like the ones described in Sections 2.1. and 2.3.1. In both cases, simple exclusion restrictions will salvage identification of the entire parameter vector of interest. As we noted above, this issue also arises in nonlinear models and might destroy identification of the entire parameter vector of interest. Conditions that would prevent this from happening are specific to the model at hand, and would go from conditions on the structure of the control function itself (as in Section 2.3.2 to

[4] We may still have $\theta_1 \subset \theta$ if taking the difference $U_{(3)}(x_{it}, s_{it}, q_{it}; \theta) - U_{(3)}(x_{jt}, s_{jt}, q_{jt}; \theta)$ always eliminates a subset of parameters.

exclusion restrictions). In general, the effect of pairwise differencing on identification is analogous to panel data fixed effects models, where differencing for a particular individual across time eliminates the unobserved fixed effect but it also wipes out any covariate that is fixed over time. As in our framework, only a subset of parameters would be identified.

2.5. *On the Presence of Semi or Nonparametric Control Functions in Econometric Models.* The control functions described in Section 2.3 showed up naturally as a result of the primitive assumptions of the underlying economic models. Without further structure, the presence of control functions in Sections 2.1. and 2.2 appears to be artificially introduced into the models described there. We will argue here that these control functions would appear naturally in the econometric model as propensity scores—in models with some selection mechanism—or in cases where endogeneity is modeled in a particular way.

2.5.1. *Control functions in models with selection.* This general notion was studied in a semi or nonparametric context for example in Ahn and Powell (1993) and Honoré and Powell (1994). Suppose $y_i^* = x_i'\beta + \varepsilon_i$, $d_i = \mathbb{1}\{w_i'\gamma + \eta_i > 0\}$ and $y_i = d_i y_i^*$. If (ε_i, ν_i) are independent of (x_i, w_i), Then we can express $y_i = x_i'\beta + g(w_i'\gamma) + \nu_i$, where $g(w_i'\gamma) = E[\varepsilon_i \mid w_i'\gamma + \eta_i > 0]$ and $E[\nu_i \mid x_i, w_i] = 0$. The parameter vector γ could be estimated using index-model methods without having to assume a particular functional form for the distribution of ν_i—as long as it has unbounded support and an everywhere-increasing distribution function. Alternatively, we could assume $d_i = \mathbb{1}\{\phi(w_i) + \eta_i > 0\}$, where $\phi(w_i)$ is an unknown function. This yields $E[d_i \mid w_i] = F_\eta(-\phi(w_i))$, where $\eta_i \sim F_\eta$ is only assumed to be everywhere increasing. Let $g(\phi(w_i)) = E[\varepsilon_i \mid \phi(w_i) + \eta_i > 0]$; invertibility of F_η implies that we can express the model as $y_i = x_i'\beta + \tilde{g}(E[d_i \mid w_i]) + \nu_i$, where $\tilde{g}(z) = g(-F_\eta^{-1}(z))$. An intermediate case is one where $d_i = \mathbb{1}\{\phi(w_i'\gamma) + \eta_i > 0\}$. Under appropriate assumptions about the otherwise unknown function $\phi(\cdot)$, the linear index $\gamma' x_i$ can be consistently estimated. If the econometrician observes only y_i, x_i, and w_i, all these cases yield special cases of the partially linear model described in Section 2.1. A control function also arises naturally in the context of a Tobit model with selection. Let $y_i^* = \max\{0, x_i'\beta + \varepsilon_i\}$, $d_i = \mathbb{1}\{\phi(w_i) + \eta_i > 0\}$ and $y_i = d_i y_i^*$. Suppose we can express[5] $\varepsilon_i = E[\varepsilon_i \mid d_i, w_i] + \nu_i \equiv g_1(\phi(w_i), d_i) + \nu_i$, where $E[\nu_i \mid w_i, x_i, d_i] = 0$, and obtain $y_i^* = \max\{0, x_i'\beta + g(\phi(w_i), d_i) + \nu_i\}$. If d_i is observed by the econometrician, then we can use the condition $E[\Delta y_{ij} - \Delta x_{ij}'\beta_0 \mid z_i, z_j\phi(w_i) - \phi(w_j) = 0] = 0$. The control function $\phi(w_i)$ can be estimated nonparametrically from $E[d_i \mid w_i] = F_\eta(-\phi(w_i))$ if we assume that F_η is invertible everywhere. We can show that if d_i is observed only when $y_i > 0$, we would have to proceed by using the entire vector w_i as the control function.

2.5.2. *Control function and endogeneity.* Consider the model $y_i = \mathbb{1}\{x_i'\beta + \varepsilon_i \geq 0\}$ and partition $x_i = (x_{1i}, x_{2i})$, where x_{1i} is suspicious of endogeneity. Let $w_i = (x_{1i}, w_{2i})$, where x_{2i} is included in w_{2i}. In the spirit of Blundell and Powell

[5] Note that $E[\varepsilon_i \mid d_i, w_i]$ depends on w_i only through $\phi(w_i)$.

(2004), suppose the "reduced form" of the model can be described as follows[6]: $x_{1i} = \phi(w_{2i}'\gamma) + \nu_i$, with $E[\nu_i \mid w_{2i}] = 0$, and $\varepsilon_i = E[\varepsilon_i \mid \nu_i] + \zeta_i \equiv g(\nu_i) + \zeta_i$, where ζ_i is independent[7] of w_i, x_i. The model becomes $y_i = \mathbb{1}\{x_i'\beta + g(\nu_i) + \zeta_i \geq 0\}$. If $\zeta_i \sim$ logistic, we obtain the partially linear logit model of Section 2.2.1. Under appropriate Index-model assumptions, the control function ν_i can be estimated semiparametrically (e.g., Ichimura and Lee, 1991) as a residual. Let $\mu_i \equiv E[y_i \mid x_i, w_i]$. If F_η is only assumed to be strictly increasing with unknown functional form, the model could be approached by noting that $\mu_i = \mu_j$ and $\nu_i = \nu_j$ imply $E[(x_i - x_j)'\delta_0 \mid \mu_i - \mu_j = 0, \nu_i - \nu_j = 0] = 0$, where δ is the vector of identified parameters in this case. Following the discussion in Section 2.4, identification would be completely destroyed in this case if $\mu_i = \mu_j$ and $\nu_i = \nu_j$ implies $x_i = x_j$. This immediately imposes a dimensionality constraint on the vector of covariates x_i. Using these same arguments it is easy to see how the partially linear Tobit model from Section 2.2.2 could arise in the context of endogeneity. Further examples where control functions have been used to control for endogeneity include Newey et al. (1999) and Das et al. (2003), where control functions also appear as propensity scores in the context of selection.

3. LARGE SAMPLE THEORY FOR THE PROPOSED ESTIMATION PROCEDURE

3.1. *Setup.* We will assume throughout that we have an i.i.d. sample $\{s_i\}_{i=1}^n$ of size n on an observable vector z_i; letting $w_i \equiv (w_{1i}, w_{2i})' \in \mathbb{R}^{L_1} \times \mathbb{R}^{L_2}$ be a given subvector of z_i, and $\gamma_0 \in \mathbb{R}^D$ be a vector of nuisance parameters, our "nonparametric control variable" is defined to be a vector-valued function $\mu : \mathbb{R}^{L_2} \times \mathbb{R}^D \longrightarrow \mathbb{R}^L$ of the form

$$\mu(w_i, \gamma_0) = \tau(w_i, \gamma_0) - E[\eta(w_i, \gamma_0) \mid w_{2i}],$$

where the functional forms for $\tau(\cdot)$ and $\eta(\cdot)$ are known, but the exact expression for $\mu(\cdot)$ is unknown due to lack of knowledge about the conditional distribution of w_i given w_{2i}. For example, for censored selection models with a binary indicator variable d_i for the uncensored observations, the control variable $\zeta_i = \zeta(w_i, \gamma_0)$ might be the propensity score $\zeta_i = E[d_i \mid w_{2i}]$, with w_{2i} being a vector of regressors in the selection equation, as in Ahn and Powell (1993); for this application, we would have $w_i \equiv (d_i, w_{2i})$, $\tau(w_i, \gamma_0) \equiv 0$, and $\eta(w_i, \gamma_0) \equiv -d_i$. Alternatively, in applications with endogenous regressors, μ_i might be the difference between the endogenous regressor x_i and its conditional mean given some instrumental variables w_{2i}, as in Blundell and Powell (2004) and the application discussed in Section 4 below (with $\tau(w_i, \gamma_0) \equiv x_i \equiv \eta(w_i, \gamma_0)$). The nuisance parameter γ_0 might appear in applications in which some semiparametric structure (e.g., a single index restriction) is imposed on the control variable.

Now let $v_i = (y_i, x_i)$ be another subvector of z_i; denoting the vector of parameters of interest by β_0, suppose there exists a function $s(v_i, v_j; \beta)$ with the property that the function

[6] As before, a linear index model is not required. We can have in general $x_{1i} = \phi(w_{2i}) + \nu_i$.

[7] More generally, we could simply assume that $\varepsilon_i \mid x_i, \nu_i \sim \varepsilon_i \mid \nu_i$.

$$T(\gamma_0, \beta) \equiv E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$$

is uniquely minimized at $\beta = \beta_0$. In terms of identification, the parameter vector β_0 includes only those that "survive" the pairwise-differencing procedure (see Section 2.4, above). For simplicity, we will focus here on the case where $s(v_i, v_j; \beta)$ is known up to β. We stress however that our methods can be extended to cases where $s(v_i, v_j; \beta)$ includes unknown functions (as in the examples in Section 2.3) or "generated regressors"; for details, see, for example, Aradillas-Lopez (2006). As we illustrated in Section 2 and further examples analyzed in Honoré and Powell (2005), such criterion functions $s(\cdot)$ are available for a number of different nonlinear models, including the partially linear logit, censored regression, and Poisson regression models and the censored regression model with selectivity, in which the control function appears in the nonparametric component of the partially linear regression function. We also showed in Section 2 that further examples have been found in the context of rational expectations and interactions-based models. Following Honoré and Powell (2005), we study the properties of estimators of β_0 defined by minimizing an estimator of $T(\gamma_0, \beta)$ of the form

$$T_n(\hat{\gamma}, b) = \binom{n}{2}^{-1} \sum_{i<j} r_n(z_i, z_j; \hat{\gamma}, b), \quad \text{with} \tag{18}$$

$$r_n(z_i, z_j; \hat{\gamma}, b) = \frac{1}{h_n^L} K\left(\frac{\hat{\mu}_n(w_{2i}, \hat{\gamma}) - \hat{\mu}_n(w_{2j}, \hat{\gamma})}{h_n}\right) s(v_i, v_j; b),$$

where $K(\cdot)$ is a kernel function and $\hat{\mu}_n(\cdot)$ is a nonparametric estimator of $\mu(\cdot)$.

In the following sections, we will provide conditions under which the resulting estimator $\hat{\beta}$ is consistent and asymptotically normal. We assume that the conditioning vector w_{2i} is continuously distributed, and, in our estimation method, realizations near the boundary of its support will be trimmed out in order to avoid the resulting bias on $\hat{\mu}_n(\cdot)$, the nonparametric estimator of $\mu(\cdot)$. We study carefully the implications of trimming on the consistency of our estimators and provide a set of sufficient conditions that ensure the appropriate rate of uniform convergence of the estimator $\hat{\mu}_n(\cdot)$ to achieve $\sqrt{n}$-consistency of $\hat{\beta}$. We also provide conditions under which trimming may disappear asymptotically. Following Honoré and Powell (2005), we propose the use of a simple jackknife procedure as an alternative to the use of a bias-reducing kernel $K(\cdot)$ in (18), which would render the objective function nonconvex and require compactness of the parameter space for uniform consistency. Bias-reducing kernels will be used only in the estimation of $\hat{\mu}_n(\cdot)$, the control function.

3.2. *Some Preliminary Results.* Suppose $w_i \equiv (w_{1i}, w_{2i})' \in \mathbb{R}^{L_1} \times \mathbb{R}^{L_2}$ is a random vector and let $f_{w_2}(w_2)$ denote the marginal density of w_2. Given two constant vectors $\gamma \in \mathbb{R}^D$ and $\omega = (\omega_1, \omega_2) \in \mathbb{R}^{L_1} \times \mathbb{R}^{L_2}$, the function $\eta : \mathbb{R}^{L_1} \times \mathbb{R}^{L_2} \times \mathbb{R}^D \to \mathbb{R}^L$, a kernel $H : \mathbb{R}^{L_2} \to \mathbb{R}$, and a bandwidth sequence $b_n : \mathbb{N} \to \mathbb{R}_{++}$, let $H_{b_n}(t) = H(t/b_n)$ and define

$$R_n(\omega_2, \gamma) = \frac{1}{nh_n^{L_2}} \sum_{i=1}^{n} \eta(w_{1i}, \omega_2, \gamma) H_{b_n}(w_{2i} - \omega_2),$$

$$\hat{f}_{w_{2n}}(\omega_2) = \frac{1}{nh_n^{L_2}} \sum_{i=1}^{n} H_{b_n}(w_{2i} - \omega_2).$$

Let $\mu(\omega, \gamma) = \tau(\omega, \gamma) - E[\eta(w_i, \gamma) \mid w_{2i} = \omega_2]$ and $\hat{\mu}_n(\omega, \gamma) = \tau(\omega, \gamma) - [R_n(\omega_2, \gamma)/\hat{f}_{w_{2n}}(\omega_2)]$. Denote $M = L + L_2 + 3$, and let $\mathbb{S}(\omega)$ denote the support of a random variable ω. Consider the following assumptions:

ASSUMPTION 1.

i. *w_{2i} is absolutely continuous with respect to Lebesgue measure;*
ii. *The density $f_{w_2}(w_2)$ is bounded, M times differentiable with respect to w_2 with bounded Mth derivative everywhere in $\mathbb{S}(w)$.*

ASSUMPTION 2. *There exists $\mathcal{W}_2 \subset \text{interior}\{\mathbb{S}(w_2)\}$ with $\inf_{w_2 \in \mathcal{W}_2} f_{w_2}(w_2) > 0$, and $\Gamma \subset \mathbb{R}^D$ where the following conditions hold:*

i. *$\mu(\omega, \gamma)$ is M times differentiable with respect to ω and γ with bounded Mth derivatives for every $\omega \in \mathbb{S}(\omega)$ and $\gamma \in \Gamma$;*
ii. *There exists a function $\bar{\eta} : \mathbb{R}^{L_1} \to \mathbb{R}_+$ such that $\|\eta(w_{1i}, w_2, \gamma)\| \leq \bar{\eta}(w_{1i})$ w.p.1 for all $w_{1i} \in \mathbb{S}(w_1)$, $w_2 \in \mathcal{W}_2$, and $\gamma \in \Gamma$ that satisfies the following: $E[\bar{\eta}(w_{1i})^2 \mid w_{2i} = w_2]$ exists and is a continuous function of w_2 for all $w_2 \in \mathbb{S}(w_2)$, and $E[\bar{\eta}(w_{1i})^4] < \infty$;*
iii. *There exists a function $\bar{\eta}_1 : \mathbb{R}^{L_1} \to \mathbb{R}_+$ and φ_1 such that $\|\eta(w_{1i}, u, \gamma) - \eta(w_{1i}, u', \gamma)\| \leq \bar{\eta}_1(w_{1i})\|u - u'\|^{\varphi_1}$ w.p.1 for all $w_{1i} \in \mathbb{S}(w_1), u, u' \in \mathcal{W}_2$, $\gamma \in \Gamma$, with $E[\bar{\eta}_1(w_{1i})] < \infty$;*
iv. *There exists a function $\bar{\eta}_2 : \mathbb{R}^{L_1} \to \mathbb{R}_+$ and φ_2 such that $\|\eta(w_1, u, \gamma) - \eta(w_1, u, \gamma')\| \leq \bar{\eta}_2(w_1)\|\gamma - \gamma'\|^{\varphi_2}$ w.p.1 for all $w_1 \in \mathbb{S}(w_1), u \in \mathcal{W}_2, \gamma, \gamma' \in \Gamma$, and $E[\bar{\eta}_2(w_1)] < \infty$.*

Assumption 1 can be relaxed to permit discrete components of w_{2i}, in which case L_2 would be the number of continuously distributed components. Assumptions 2(iii) and 2(iv) can be seen as "in probability" Lipschitz conditions—see, for example, lemma 2.9 in Newey and McFadden (1994). They would be immediately satisfied, for example, if $\eta(w_1, u, \gamma)$ is assumed to be differentiable with respect to u and γ with bounded derivatives in $\mathcal{W}_2$ and Γ.

ASSUMPTION 3.

i. *Define $\mathcal{H} \equiv \{t \in \mathbb{R}^{L_2} : H(t) \neq 0\}$; then $\mathcal{H} \subset \mathbb{R}^{L_2}$ is compact. $H(\cdot)$ is bounded and symmetric about zero, with $\int H(t)\,dt = 1$. Denote $t = (t_1, \ldots, t_{L_2})'$; then $\int \|t\|^M |H(t)|\,dt < \infty$ and $\int (t_1^{q_1} \cdots t_{L_2}^{q_{L_2}}) H(t)\,dt_1 \ldots dt_{L_2} = 0$ for all $0 < q_1 + \cdots + q_{L_2} < M$. There exist $\varphi \in (0, M)$ and $c_H < \infty$ such that $|H(t) - H(t')| \leq c_H \|t - t'\|^{\varphi}\ \forall\, t, t'$.*
ii. *b_n satisfies $b_n = o(1), \ln n b_n^{-2L_2} = o(n), b_n^{2M} = o(n^{-1})$ and $b_n^{-L_2 - 2\varphi} = o(n^{1-\sigma})$ for some $\sigma > 0$.*

iii. The sets $\mathcal{W}_2$ *and* $\mathcal{H}$*, and the bandwidth* b_n *are such that* $b_n t + w_2 \in$ interior $\{\mathbb{S}(w_2)\} \forall\, t \in \mathcal{H}, w_2 \in \mathcal{W}_2, n \in \mathbb{N}$.

The following result will be useful.

THEOREM 1. *If Assumptions (1)–(3) are satisfied, then*

(a) $\sup_{\substack{\omega \in \mathcal{W}_2 \\ \gamma \in \Gamma}} (n^{1-\delta} b_n^{L_2})^{1/2} \|\hat{\mu}_n(\omega, \gamma) - \mu(\omega, \gamma)\| = O_p(1)$ *for any* $\delta > 0$.

(b) $\hat{\mu}_n(\omega, \gamma) - \mu(\omega, \gamma) = \frac{1}{f_{w_2}(\omega)} \frac{1}{n b_n^{L_2}} \sum_{i=1}^n [\tau(\omega, \gamma) - \eta(w_{1i}, \omega_2, \gamma) - \mu(\omega, \gamma)] \times H_{b_n}(w_{2i} - \omega) + \xi_n(\omega, \gamma)$,

where $\sup_{\substack{\omega \in \mathcal{W}_2 \\ \gamma \in \Gamma}} \|\xi_n(\omega, \gamma)\| = O_p(n^{\delta - 1} b_n^{-L_2})$ *for any* $\delta > 0$.

Note that $\tau(\omega, \gamma) - \mu(\omega, \gamma) = E[\eta(w_{1i}, \omega_2, \gamma) \mid w_{2i} = \omega_2]$, so the linear representation in part (b) of Theorem 1 depends only on $\eta(w_i, \gamma) - E[\eta(w_i, \gamma) \mid w_{2i} = \omega_2]$. Theorem 1 is a special case of a more general result shown in Aradillas-Lopez (2005). It will be crucial for the main results presented below. The next result is an immediate consequence:

COROLLARY 1. *Suppose we strengthen the condition* $\ln n b_n^{-L_2} = o(n)$ *to* $n^{1-\delta} b_n^{-2L_2} = o(1)$ *for some* $\delta > 0$. *Let* $\xi_n(\omega, \gamma)$ *be as defined in Theorem 1; then* $\sup_{\substack{\omega \in \mathcal{W}_2 \\ \gamma \in \Gamma}} \|\xi_n(\omega, \gamma)\| = o_p(\mathrm{N}^{-1/2})$.

3.3. *Estimation.* We will examine the case in which the pairwise difference estimator depends on the unknown function $\mu(\omega, \gamma)$, which was defined in Section 3.2. Because this function is unknown, we will use its nonparametric estimator $\hat{\mu}_n(\omega, \gamma)$, which was also defined above. We assume w_{2i} to be continuously distributed. Thus, in order to avoid the influence of points in the boundary of $\mathbb{S}(w_2)$—which would introduce a bias on $\hat{\mu}_n(w_2, \gamma)$, we will analyze a trimmed version of the objective function in Equation (18). We will denote $z_i = (y_i, x_i, w_i)$ and $v_i = (y_i, x_i)$, where $w_i = (w_{1i}, w_{2i})$, and w_i is as described in Section 3.2. We will use $\mathcal{W}_2$ as the trimming set, where $\mathcal{W}_2$ satisfies Assumptions (1) and (2). Let

$$r_n(z_i, z_j; \gamma, b) = \frac{1}{h_n^L} K\left(\frac{\hat{\mu}_n(w_i, \gamma) - \hat{\mu}_n(w_j, \gamma)}{h_n}\right) s(v_i, v_j; b) \cdot a(w_{2i}) a(w_{2j}),$$

where $a(\cdot)$ is the trimming function described as

$$a(w_2) = \begin{cases} \phi(w_2) > 0 & \text{if } w_2 \in \mathcal{W}_2, \\ 0 & \text{otherwise.} \end{cases}$$

The function $\phi(u)$ is bounded, continuous, and strictly positive for all $u \in \mathbb{R}^{L_2}$. We will describe the properties of the function $s(v_i, v_j; b)$ below. Define

$$T_n(\gamma, b) = \binom{n}{2}^{-1} \sum_{i<j} r_n(z_i, z_j; \gamma, b), \tag{19}$$

where $\hat{\mu}_n(\omega, \gamma)$ is as defined in Section 3.2.

ASSUMPTION 4.

i. $E[s(v_i, v_j; b)^2] < \infty$;
ii. $E[\|\mu(w_i, \gamma) - \mu(w_j, \gamma)\|^2] < \infty$;
iii. $\mu(w_i, \gamma_0)$ *is continuously distributed with bounded density,* $f_{\mu(w,\gamma_0)}(\cdot)$, *which is a continuous function;*
iv. Denote $\kappa_s(a_1, a_2, b) = E[s(v_i, v_j; b) \mid z_i = a_1, \mu(w_j, \gamma_0) = a_2]$; *then* $\kappa_s(\cdot)$ *exists and is a continuous function of each of its arguments;*
v. Let $\rho(\mu(w_i, \gamma_0)) = E[a(w_{2i}) \mid \mu(w_i, \gamma_0)]$. *Then* $\rho(\mu(w_i, \gamma_0))$ *is continuous and strictly positive for all* $\mu(w_i, \gamma_0)$ *such that* $w_{2i} \in \mathcal{W}_2$. *Denote* $\ell_s(a_1, a_2, b) = \kappa_s(a_1, a_2, b)\rho(a_2) f_{\mu(w,\gamma_0)}(a_2)$; *then* $|\ell_s(a_1, a_2, b)| \le c_1(a_1, a_2, b)$ *with* $E[c_1(v_i, \mu(w_i, \gamma_0), b)] < \infty$ *for all b;*
vi. $\{z_i, i = 1, \ldots, n\}$ *is an i.i.d. sample.*

The use of trimming in the objective function could have an adverse effect on the consistency of $\hat{\beta}$ if the value of β that maximizes $E[a(w_{2i})a(w_{2j})s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$ differs from the one that corresponds to $E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$. Typically, this issue would be addressed by introducing some form of exclusion restriction. Suppose all we know is that $E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$ is uniquely minimized when $\beta = \beta_0$. The next assumption describes an exclusion restriction that would eliminate any potential trimming bias.

ASSUMPTION 5. $E[s(v_i, v_j; b) \mid v_i, w_{2j}, \mu(w_j, \gamma_0)] = E[s(v_i, v_j; b) \mid v_i, \mu(w_j, \gamma_0)]$.

In essence, what we require is that, conditional on v_i, $s(v_i, v_j; b)$ is mean-independent of w_{2j} conditional on $\mu(w_j, \gamma_0)$. The worst-case scenario would be one in which all we have to work with is the assumption that $E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$ is uniquely minimized when $\beta = \beta_0$, and Assumption 5 does not hold. In Section 3.9, we present a solution to that case. There, we describe an alternative trimming methodology based on a sequence of trimming functions $a_n(\cdot)$ and a sequence of trimming sets $\mathcal{W}_{2_n}$ and we describe conditions under which $E[a_n(w_{2i})a_n(w_{2j})s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) = \mu(w_j, \gamma_0)] \xrightarrow{p} E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) = \mu(w_j, \gamma_0)]$ uniformly in B. If such conditions are satisfied, and the alternative trimming methodology is used, then consistency of the resulting estimator $\hat{\beta}$ would rely exclusively on the assumption that $E[s(v_i, v_j; \beta) \mid \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$ is uniquely minimized when $\beta = \beta_0$ regardless of whether or not Assumption 5 is satisfied. Basically, what we would need is for the properties of $\mathcal{W}_2$ in Assumptions 1 and 2 to be satisfied by any arbitrary compact set in the interior of $\mathbb{S}(w_2)$, the sequence of trimming

functions $a_n(w_2)$ to converge to a positive constant c with probability one, and the tails of $f_{w_2}(\cdot)$ to converge to zero at an appropriate rate relative to the one at which $\mathcal{W}_{2n}$ converges to $\mathbb{S}(w_2)$. We will describe the set of conditions in detail in Section 3.9.

ASSUMPTION 6. *We have $h_n = o(1)$. Let b_n be the bandwidth used in the estimation of $\mu(\cdot)$. We will strengthen Assumption 3(ii) and assume that $n^{1-2\delta} b_n^{L_2} h_n^{2(L+2)} \to \infty$ for some $\delta > 0$ and $n b_n^{2M} h_n^{-2(L+1)} \to 0$.*

ASSUMPTION 7. *K is bounded and symmetric about zero with $\int K(u)\,du = 1$ and $\|u\| \cdot |K(u)| \to 0$ as $\|u\| \to \infty$. $K(\cdot)$ is twice differentiable with bounded derivatives. Denote its gradient by $K^{(1)}(\cdot) \in \mathbb{R}^L$; then $K^{(1)}(t) = -K^{(1)}(-t)$ for all t.*

ASSUMPTION 8. *We assume $\gamma_0 \in \Gamma$, which is described in Theorem 1. We allow the use of an estimator $\hat{\gamma}$ of γ_0 if necessary. Then, either*

i. $\hat{\gamma} = \gamma_0 \in \Gamma$ *or*
ii. $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$, *and* $\hat{\gamma} \in \Gamma$ *for all* n.

ASSUMPTION 9. $|s(v_i, v_j; b_1) - s(v_i, v_j; b_2)| \le B_{ij}\|b_1 - b_2\|^{\alpha}$ *for some* $\alpha > 0$, *where* $E[B_{ij}^2] < \infty$.

Our estimator $\hat{\beta}$ is defined as

$$\hat{\beta} = \operatorname*{argmin}_b T_n(\hat{\gamma}, b). \tag{20}$$

We first analyze the limiting objective function.

3.3.1. *Pointwise convergence to limiting objective function.* Define

$$T(\gamma_0, b) = E[a(w_{2i})\ell_s(v_i, \mu(w_i, \gamma_0), b)],$$

where $\kappa_s(\cdot)$ and $g(\cdot)$ are described in Assumptions 4(iv) and 4(vi)—above. Then, if Assumptions 1–8 hold, $T_n(\hat{\gamma}, b) \to T(\gamma_0, b)$.

PROOF. Define

$$\mathcal{T}_n(\gamma_0, b) = \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0)}{h_n}\right) s(v_i, v_j; b) a(w_{2i}) a(w_{2j}).$$

First we show that $T_n(\gamma_0, b) \to \mathrm{T}(\gamma_0, b)$. Note that if Assumption 7 is satisfied, then

$$
\begin{aligned}
&E[\mathcal{T}_n(\gamma_0, b)] \\
&\quad = E\left[\frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0)}{h_n}\right) s(v_i, v_j; b) a(w_{2i}) a(w_{2j})\right] \\
&\quad = E\left[\frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0)}{h_n}\right) a(w_{2i}) \rho(\mu(w_j, \gamma_0)) \kappa_s(v_i, \mu(w_j, \gamma_0); b)\right] \\
&\quad = E\Bigg[a(w_{2i}) \int K(\psi) \rho(h_n \psi + \mu(w_i, \gamma_0)) \kappa_s(v_i, h_n \psi + \mu(w_i, \gamma_0); b) \\
&\qquad\qquad \times f_\mu(h_n \psi + \mu(w_i, \gamma_0))\, d\psi\Bigg] \\
&\quad \to E[a(w_{2i}) \rho(\mu(w_i, \gamma_0)) \kappa_s(v_i, \mu(w_i, \gamma_0), b)\, f_\mu(\mu(w_i, \gamma_0))] \\
&\quad = E\big[f_\mu(\mu(w_i, \gamma_0)) \rho(\mu(w_i, \gamma_0))^2 E[\kappa_s(v_i, \mu(w_i, \gamma_0), b) \mid \mu(w_i, \gamma_0)]\big],
\end{aligned}
$$

where the next-to-last line follows from Assumptions 4(iv)–4(vi), 5, and 6, and the last line follows once again from Assumption 5. Using Assumptions 4(i), 6, 7, and the properties of $a(\cdot)$, we have

$$
E\left[\left\{\frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0)}{h_n}\right) s(v_i, v_j; b) a(w_{2i}) a(w_{2j})\right\}^2\right] = O(n),
$$

which satisfies lemma A.3 of Ahn and Powell (1993). Consequently, $\mathcal{T}_n(\gamma_0, b) \to E[\mathcal{T}_n(\gamma_0, b)] \to T(\gamma_0, b)$. Next we show that $T_n(\hat{\gamma}, b) \to \mathcal{T}_n(\gamma_0, b)$:

$$
\begin{aligned}
&|T_n(\hat{\gamma}, b) - \mathcal{T}_n(\gamma_0, b)| \\
&\quad \le \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L} \left| K\left(\frac{\hat{\mu}_n(w_i, \hat{\gamma}) - \hat{\mu}_n(w_j, \hat{\gamma})}{h_n}\right) - K\left(\frac{\hat{\mu}_n(w_i, \gamma_0) - \hat{\mu}_n(w_j, \gamma_0)}{h_n}\right)\right| \\
&\qquad \times |s(v_i, v_j; b)| a(w_{2i}) a(w_{2j}).
\end{aligned}
$$

We have

$$
\begin{aligned}
&\left| K\left(\frac{\hat{\mu}_n(w_i, \hat{\gamma}) - \hat{\mu}_n(w_j, \hat{\gamma})}{h_n}\right) - K\left(\frac{\hat{\mu}_n(w_i, \gamma_0) - \hat{\mu}_n(w_j, \gamma_0)}{h_n}\right)\right| \\
&\quad \le \frac{2L}{h_n} \left\| K^{(1)}(d_{ij}^*)\right\| \cdot \Big[\max_i \|\hat{\mu}_n(w_i, \hat{\gamma}) - \mu(w_i, \hat{\gamma})\| + \max_i \|\hat{\mu}_n(w_i, \gamma_0) - \mu(w_i, \gamma_0)\| \\
&\qquad\qquad + \max_i \|\mu(w_i, \hat{\gamma}) - \mu(w_i, \gamma_0)\|\Big] |s(v_i, v_j; b)| a(w_{2i}) a(w_{2j}).
\end{aligned}
$$

By Theorem 1, $a(\cdot)$, and Assumption 8, we have $\max_i \|\hat{\mu}_n(w_i, \hat{\gamma}) - \mu(w_i, \hat{\gamma})\| = O_p(n^{1-\delta} b_n^{L_2})^{-1/2}$ for every $\delta > 0$. The same result holds for $\max_i \|\hat{\mu}_n(w_i, \gamma_0) - \mu(w_i, \gamma_0)\|$. By Assumption 2(i), there exists a $C_1 > 0$ such that $\max_i \|\mu(w_i, \hat{\gamma}) - \mu(w_i, \gamma_0)\| \le C_1 \|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$—using Assumption 8. From Assumption 7,

there exists $C_2 > 0$ such that $\|K^{(1)}(t)\| \le C_2$ for all t. Combining all this and letting $C = 2LC_1C_2$, we have

$$|T_n(\hat{\gamma}, b) - \mathcal{T}_n(\gamma_0, b)| \le \frac{C}{h_n^{L+1}}\Big[O_p\big(n^{1-\delta}b_n^{L_2}\big)^{-1/2} + O_p(n^{-1/2})\Big]\binom{n}{2}^{-1} \times \sum_{i<j} |s(v_i, v_j; b)|a(w_{2i})a(w_{2j})$$

for all $\delta > 0$. Therefore, Assumption 6 yields $h_n^{-L-1}O_p(n^{1-\delta}b_n^{L_2})^{-1/2} = o_p(1)$ and $h_n^{-L-1}O_p(n^{-1/2}) = o_p(1)$. Combining this with Assumptions 4(i), 6, and the properties of $a(\cdot)$, we have $|T_n(\hat{\gamma}, b) - \mathcal{T}_n(\gamma_0, b)| = o_p(1)$, and therefore $T_n(\hat{\gamma}, b) \to T(\gamma_0, b)$, which proves the claim. ■

Suppose that the following modified version of Assumption 5 holds:

ASSUMPTION 5′. $E[s(v_i, v_j; b) \mid v_i, x_j, w_{2j}, \mu(w_j, \gamma_0)] = E[s(v_i, v_j; b) \mid v_i, x_j, \mu(w_j, \gamma_0)]$.

Based on this condition, we can modify Assumption 4 accordingly:

ASSUMPTION 4′. *Maintain 4(i), 4(ii), and 4(vi) but modify 4(iii)–4(v) to assume that*

iii. Let $f_{\mu|x}(\mu; x)$ be the conditional density of $\mu(w_j, \gamma_0)$ given $x_j = x$. Then $f_{\mu|x}(\mu; x)$ is a continuous function of its first argument.

iv. Define $\bar{\kappa}_s(a_1, a_2, a_3) = E[s(v_i, v_j; b) \mid z_i = a_1, x_j = a_2, \mu(w_j, \gamma_0) = a_3]$. Then $\bar{\kappa}_s(a_1, a_2, a_3)$ exists and is a continuous function of each of its arguments;

v. Let $\bar{\rho}(a_2, a_3) = E[a(w_{2j}) \mid x_j = a_2, \mu(w_j, \gamma_0) = a_3]$. Then $\bar{\rho}(a_2, a_3)$ is a continuous function of its second argument and is strictly positive for all $\mu(w_j, \gamma_0)$ such that $w_{2j} \in \mathcal{W}_2$. Define $\bar{\ell}_s(a_1, a_2, a_3, b) = \bar{\rho}(a_2, a_3)\bar{\kappa}_s(a_1, a_2, a_3) f_{\mu|x}(a_3; a_2)$; then $|\ell_s(a_1, a_2, a_3, b)| < \bar{c}_1(a_1, a_2, a_3, b)$, with $E[\bar{c}_1(a_1, a_2, a_3, b)] < \infty$ for all b.

Define

$$\bar{T}(\gamma_0, b) = E[a(w_{2i})\bar{\ell}_s(v_i, x_j, \mu(w_i, \gamma_0))].$$

Then, if Assumptions 1–3, 4′, 5′, and 6–8 hold, $T_n(\hat{\gamma}, b) \to \bar{T}(\gamma_0, b)$. The proof follows the same steps as above, replacing Assumptions 4 and 5 with 4′ and 5′, respectively.

3.4. *Uniform Convergence to Limiting Objective Function.* Let B be the parameter space for β. If B is compact, then $\sup_{b\in B} |T_n(\hat{\gamma}, b) - T(\gamma_0, b)| = o_p(1)$.

PROOF. Using Assumption 9 and following steps parallel to those used to show that $\sup_{b\in B} |\mathcal{T}_n(\gamma_0, b) - T(\gamma_0, b)| = o_p(1)$. Take $b_1, b_2 \in B$. Using the same steps as the ones used in the pointwise convergence proof, we can show that

$$
\begin{aligned}
&|T_n(\hat{\gamma}, b_1) - \mathcal{T}_n(\gamma_0, b_2)| \\
&\quad \le \frac{C}{h_n^{L+1}} \left[O_p\big(n^{1-\delta} b_n^{L_2}\big)^{-1/2} + O_p(n^{-1/2}) \right] \binom{n}{2}^{-1} \\
&\quad \times \sum_{i<j} |s(v_i, v_j; b_1) - s(v_i, v_j; b_1)| a(w_{2i}) a(w_{2j}) \\
&\quad \le \frac{C}{h_n^{L+1}} \left[O_p\big(n^{1-\delta} b_n^{L_2}\big)^{-1/2} + O_p(n^{-1/2}) \right] \binom{n}{2}^{-1} \sum_{i<j} B_{ij} a(w_{2i}) a(w_{2j}) \|b_1 - b_2\|^{\alpha},
\end{aligned}
$$

for all $\delta > 0$, and using Assumptions 6, 9, and the compactness of B, we get $\sup_{b \in B} |T_n(\hat{\gamma}, b) - \mathcal{T}_n(\gamma_0, b)| = o_p(1)$, and consequently $\sup_{b \in B} |T_n(\hat{\gamma}, b) - Q(\gamma_0, b)| = o_p(1)$, which establishes uniform convergence. ■

As with pointwise convergence, a parallel result holds if we replace Assumptions 4 and 5 with 4′ and 5′, respectively. In this case, we have $\sup_{b \in B} |\mathcal{T}_n(\gamma_0, b) - \bar{T}(\gamma_0, b)| = o_p(1)$.

3.5. *Identification.* As we mentioned above, trimming is done to avoid bias of $\hat{\mu}_n(\cdot)$ that would be caused by points on the boundary of $\mathbb{S}(w_2)$. Given this need, we chose $\mathcal{W}_2$ as the trimming set to take advantage of the results of Theorem 1. Without additional assumptions, trimming may cause bias of $\hat{\beta}$ in this setting. The purpose of Assumptions 4(vi), 4(vii), and 5 is to avoid the presence of such bias. As we stated above, if Assumptions 1–3 hold for any compact set in the interior of $\mathbb{S}(w_2)$ and if $f_{w_2}(\cdot)$ satisfies some additional assumptions (see Assumption 18, below), we could make the trimming disappear asymptotically and leave the rate of convergence of our estimator unchanged. This would allow us to relax Assumptions 4(vi), 4(vii), and 7. Below, we will present conditions under which this can be done.

If Assumption 5 is satisfied, the limiting objective function $T(\gamma_0, b)$ is given by

$$
\begin{aligned}
T(\gamma_0, b) &= E[a(w_{2i}) \ell_s(v_i, \mu(w_i, \gamma_0), b)] \\
&= E\big[f_\mu(\mu(w_i, \gamma_0)) \rho(\mu(w_i, \gamma_0))^2 E[s(v_i, v_j, b) \,|\, \mu(w_j, \gamma_0) = \mu(w_i, \gamma_0)] \big],
\end{aligned}
$$

where the last equality comes from Assumption 5. $T(\gamma_0, b)$ is uniquely minimized at β_0 if the following condition holds:

ASSUMPTION 10. *$E[s(v_i, v_j; b) \,|\, \mu(w_j, \gamma_0) = \mu(w_i, \gamma_0)]$ is uniquely minimized at $b = \beta_0$.*

If Assumption 5′ is satisfied, the limiting objective function $\bar{T}(\gamma_0, b)$ is

$$
\begin{aligned}
\bar{T}(\gamma_0, b) &= E[a(w_{2i}) \bar{\ell}_s(v_i, x_j, \mu(w_i, \gamma_0))] \\
&= E[a(w_{2i}) \bar{\rho}(x_j, \mu(w_i, \gamma_0)) f_{\mu|x}(\mu(w_i, \gamma_0); x_j) \\
&\qquad \times E[s(v_i, v_j, b) \,|\, x_i, x_j, \mu(w_j, \gamma_0) = \mu(w_i, \gamma_0)]]
\end{aligned}
$$

The last equality being a consequence of Assumption 5′. $\bar{T}(\gamma_0, b)$ is uniquely minimized at β_0 if the following condition holds:

ASSUMPTION 10′. *$E[s(v_i, v_j; b) \mid x_i, x_j, \mu(w_i, \gamma_0) - \mu(w_j, \gamma_0) = 0]$ is uniquely minimized at $b = \beta_0$.*

As it is the case with extremum estimators, Assumption 10 (or 10′) yields identification and, along with the convergence results of Sections 3.3.1 and 3.4, it will also yield consistency of $\hat{\beta}$. We present the results now.

3.5.1. *Consistency theorem.* Let $\hat{\beta}$ be the minimizer of $T_n(\hat{\gamma}, b)$ over the parameter space of β, denoted by B. We have the following results:

THEOREM 2. *Let $K_{h_n}(t) \equiv K(t/h_n)$. If $K_{h_n}(\mu(w_i, \gamma) - \mu(w_j, \gamma))s(v_i, v_j, b) \times a(w_{2i})a(w_{2j})$ is a continuous and convex function of b and B is a convex set with the true value β_0 in its interior, then $\hat{\beta} \xrightarrow{p} \beta_0$ under Assumptions 1–8 and 10. This result remains true if we replace Assumptions 4, 5, and 10 with 4′, 5′, and 10′, respectively.*

PROOF. Follows from the result in Section 3.3.1 and theorem 2.7 in Newey and McFadden (1994). ■

THEOREM 3. *Let $K_{h_n}(t) \equiv K(t/h_n)$. If $K_{h_n}(\mu(w_i, \gamma) - \mu(w_j, \gamma))s(v_i, v_j, b) \times a(w_{2i})a(w_{2j})$ is a continuous function of b, and B is a compact set that includes the true value β_0 as an interior point; then $\hat{\beta} \xrightarrow{p} \beta_0$ under Assumptions 1–10. This result remains true if we replace Assumptions 4, 5, and 10 with 4′, 5′, and 10′, respectively.*

PROOF. Follows from the result in Section 3.4 and Theorem 2.1 along with lemma 2.9 in Newey and McFadden (1994). ■

3.6. *Asymptotic Normality.* Let β_h be the minimizer of $E[h_n^{-L} K_{h_n}(\mu(w_i, \gamma) - \mu(w_j, \gamma))s(v_i, v_j, b)a(w_{2i})a(w_{2j})]$. The same arguments that lead to the consistency of $\hat{\beta}$ imply that $\beta_h \xrightarrow{p} \beta_0$. Also note that β_h is nonstochastic. In this section we will derive the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta_h)$, where $\hat{\beta}$ is the minimizer of $T_n(\hat{\gamma}, b)$. In all the applications considered in this article, the function $s(v_i, v_j; \beta)$ is left and right differentiable with respect to each component of β. Let

$$v_n(z_i, z_j; \gamma, \beta) = \frac{1}{h_n^L} K\left(\frac{\hat{\mu}_n(w_i, \gamma) - \hat{\mu}_n(w_j, \gamma)}{h_n}\right) t(v_i, v_j, \beta)a(w_{2i})a(w_{2j})$$

$$p_n(z_i, z_j; \gamma, \beta) = \frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma) - \mu(w_j, \gamma)}{h_n}\right) t(v_i, v_j, \beta)a(w_{2i})a(w_{2j})$$

and define

$$\hat{G}_n(\gamma,\beta)=\binom{n}{2}^{-1}\sum_{i<j}v_n(z_i,z_j;\gamma,\beta)\text{ and }G_n(\gamma,\beta)=\binom{n}{2}^{-1}\sum_{i<j}p_n(z_i,z_j;\gamma,\beta),$$

where $t(v_i,v_j,\beta)$ is a convex combination of the left and right derivatives of $s(v_i,v_j;\beta)$ with respect to each component of β. Since $\hat{\beta}$ is the minimizer of $T_n(\hat{\gamma},b)$, the object of interest is $\hat{G}_n(\hat{\gamma},\beta)$. Below, we will characterize the relationship between $\hat{G}_n(\hat{\gamma},\beta)$ and $G_n(\hat{\gamma},\beta)$, which will determine the asymptotic distribution of $\sqrt{n}(\hat{\beta}-\beta_h)$. We will add the following assumptions:

ASSUMPTION 11. *The derivative function* $\{t(\cdot,\cdot;\beta):\beta\in B\}$ *is Euclidean for an envelope* F, *i.e.*,

$$\sup_{n,\beta}|t(z_i,z_j;\beta)|\le F(z_i,z_j),$$

satisfying

$$E\big[F(z_i,z_j)^2\big]<\infty\text{ and }\sup_{\gamma\in\Gamma}E\big[\|F(z_i,z_j)[\tau(w_i,\gamma)-\eta(w_i,\gamma)-\mu(w_j,\gamma)]\|^2\big]<\infty.$$

The set B need not be the whole parameter space, but could be some other set with β_0 in its interior.

ASSUMPTION 12. *The function* $\phi(\cdot)$, *in the definition of the trimming function* $a(\cdot)$, *is* M *times differentiable with bounded derivatives*.

ASSUMPTION 13. *Define*

$$B_n(w_{2i};\gamma,\beta)=E\left[\frac{1}{h_n^{L+1}}K^{(1)}\left(\frac{\mu(w_i,\gamma)-\mu(w_j,\gamma)}{h_n}\right)t(v_i,v_j,\beta)a(w_{2j})\,\middle|\,w_{2i}\right];$$

then $B_n(w_{2i};\gamma,\beta)$ *is* M *times differentiable with respect to* w_{2i} *with bounded derivatives everywhere in* Γ, B.

Note that if Assumption 7 is satisfied, then $\int K^{(1)}(t)\,dt=0$. We will now strengthen Assumption 2(iii), one of the "in probability" Lipschitz conditions for $\eta(\cdot)$:

ASSUMPTION 14. *With probability one in* $\mathbb{S}(w_1)$, *the function* $\eta(w,\gamma)$ *is* M *times differentiable with respect to* w_2 *with bounded* M*th derivatives for all* $w_2\in\mathcal{W}_2$ *and* $\gamma\in\Gamma$.

ASSUMPTION 15. *The true parameter* β_0 *is in the interior of the parameter space*.

Define the projection functions

$$\begin{aligned}p_{1n}(z_i;\gamma,\beta)&=E[p_n(z_i,z_j;\gamma,\beta)\,|\,z_i]-E[p_n(z_i,z_j;\gamma,\beta)],\\ p_{0n}(\gamma,\beta)&=E[p_n(z_i,z_j;\gamma,\beta)],\end{aligned}$$

and let $\tilde{p}_n(z_i,\gamma,\beta)=p_{0n}(\gamma,\beta)+2p_{1n}(z_i;\gamma,\beta)$.

ASSUMPTION 16.

i. $\tilde{p}_n(z_i, \gamma, \beta)$ *satisfies the following:*

(a) $\tilde{p}_n(z_i, \gamma, \beta)$ *is continuously differentiable in* (γ, β) *with a derivative* $\Delta\tilde{p}_n(z_i, \gamma, \beta)$ *with the property that for any sequence* (γ^*, β^*) *that converges in probability to* (γ_0, β_0), $\Delta\tilde{p}_n(z_i, \gamma^*, \beta^*)$ *converges to a matrix* $\Delta\tilde{p}_0(\gamma_0, \beta_0)$. *Let* $\Delta\tilde{p}_0^{\beta}(\gamma_0, \beta_0)$ *and* $\Delta\tilde{p}_0^{\gamma}(\gamma_0, \beta_0)$ *denote the parts that correspond to the differentiation with respect to* β *and* γ, *respectively. Then* $\Delta\tilde{p}_0^{\beta}(z_i, \gamma_0, \beta_0)$ *is nonsingular.*

(b) There exists a function $p_1(z_i; \gamma_0, \beta_0)$ *with* $E[\|p_1(z_i; \gamma_0, \beta_0)\|^2] < \infty$ *such that*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{p}_n(z_i; \gamma_0, \beta_h) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_1(z_i; \gamma_0, \beta_0) = o_p(1).$$

ii. $\eta(w_i, \gamma)$ *and* $B_n(w_{2i}; \gamma, \beta)$ *are continuously differentiable in* γ *and* (γ, β), *respectively, with derivatives* $\Delta^{\gamma}\eta(w_i, \gamma)$ *and* $\Delta B_n(w_{2i}; \gamma, \beta)$ *respectively—the first assumption strengthens the "in probability" Lipschitz condition 2(iv).*

iii. Let $\tilde{D}_n(w_i; \gamma, \beta) = B_n(w_{2i}; \gamma, \beta)[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)]a(w_{2i})$. *The previous condition, along with Assumption 2(i) imply that* $\tilde{D}_n(w_i; \gamma, \beta)$ *is continuously differentiable in* (γ, β); *denote this derivative by* $\Delta\tilde{D}_n(w_i; \gamma, \beta)$. *We assume that* $\Delta\tilde{D}_n(w_i; \gamma, \beta)$ *has the property that for any sequence* (γ^*, β^*) *that converges in probability to* (γ_0, β_0), $\Delta\tilde{D}_n(w_i; \gamma^*, \beta^*)$ *converges in probability to a matrix* $\Delta\tilde{D}_0(\gamma_0, \beta_0)$. *Let* $\Delta\tilde{D}_0^{\beta}(\gamma_0, \beta_0)$ *and* $\Delta\tilde{D}_0^{\gamma}(\gamma_0, \beta_0)$ *be the parts that corresponds to* β *and* γ, *respectively.*

iv. For some function $D(w_i; \gamma_0, \beta_0)$ *with* $E[\|D(w_i; \gamma_0, \beta_0)\|^2] < \infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{D}_n(w_i; \gamma_0, \beta_h) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}D(w_i; \gamma_o, \beta_0) = o_p(1).$$

Note that $E[\tilde{D}_n(w_i; \gamma, \beta)] = 0$ and $\Delta D_n(w_i; \gamma, \beta) = 0$ for any γ and β. Thus, we must have $E[D(w_i; \gamma_0, \beta_0)] = 0$ and $\Delta\tilde{D}_0^{\beta}(\gamma_0, \beta_0) = 0$. Let $\Delta\tilde{p}_0^{\gamma}(\gamma_0, \beta_0)$, $\Delta\tilde{p}_0^{\beta}(\gamma_0, \beta_0)$, and $p_1(z_i; \gamma_0, \beta_0)$ be as defined in Assumption 16(i). The following is the main asymptotic normality theorem in this section.

THEOREM 4. *Suppose* $\hat{\beta}$ *is a consistent estimator of* β, $\sqrt{n}(\hat{\gamma} - \gamma_0) = n^{-1/2}\sum_{i=1}^{n}\upsilon_i + o_p(1)$ *and* $\hat{G}_n(\hat{\gamma}, \hat{\beta}) = o_p(n^{-1/2})$. *If Assumptions 1–3, 6, 7, and 11–16 are satisfied, then*

$$\sqrt{n}(\hat{\beta} - \beta_h) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_p(1),$$

where

$$\psi_i = -\Delta\tilde{p}_0^{\beta}(\gamma_0,\beta_0)^{-1}\times\big[\big(\Delta\tilde{D}_0^{\gamma}(\gamma_0,\beta_0)+\Delta\tilde{p}_0^{\gamma}(\gamma_0,\beta_0)\big)\upsilon_i+2p_1(z_i;\gamma_0,\beta_0)+2D(w_i;\gamma_0,\beta_0)\big].$$

Furthermore, assuming υ_i, $p_1(z_i;\gamma_0,\beta_0)$, *and* $D(w_i;\gamma_0,\beta_0)$ *are jointly i.i.d. with* $E[\upsilon_i]=0$ *and* $E[\|\upsilon_i\|^2]<\infty$,

$$\sqrt{n}(\hat{\beta}-\beta_h)\xrightarrow{d}\mathcal{N}(0,E[\psi_i\psi_i']).$$

PROOF. In the Appendix, we show that if Assumptions 1–3, 6, 7, and 11–14 are satisfied and $\{z_i, i=1,\dots,n\}$ is an i.i.d. sample,

$$\begin{aligned}\hat{G}_n(\gamma,\beta)&=G_n(\gamma,\beta)+\frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n}B_n(w_{2i};\gamma,\beta)[\tau(w_i,\gamma)-\eta(w_i,\gamma)-\mu(w_i,\gamma)]a(w_{2i})\\&\quad+\tilde{e}_n(z_i,z_j;\gamma,\beta)\\&\equiv G_n(\gamma,\beta)+\frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n}\tilde{D}_n(w_i;\gamma,\beta)+\tilde{e}_n(z_i,z_j;\gamma,\beta),\end{aligned}$$

where $\sup_{B,\Gamma}|\tilde{e}_n(z_i,z_j;\gamma,\beta)|=o_p(n^{-1/2})$. Note that $E[\tilde{D}_n(w_i;\gamma,\beta)]=0$ for any γ and β. Therefore, $\sup_{B,\Gamma}|E[\hat{G}_n(\gamma,\beta)]-E[G_n(\gamma,\beta)]|=o_p(n^{-1/2})$. This shows that the additional bias on $\hat{\beta}$ introduced by having to estimate the unknown function $\mu(\cdot)$ nonparametrically is at most of order $o_p(n^{-1/2})$, a consequence of using bias-reducing kernels in the construction of $\hat{\mu}_n(\cdot)$. This allows us to center the distribution of $\hat{\beta}$ around β_h. By the usual projection–decomposition, we have

$$\begin{aligned}G_n(\gamma,\beta)&\equiv\binom{n}{2}^{-1}\sum_{i<j}p_n(z_i,z_j;\gamma,\beta)\\&=p_{0n}(\gamma,\beta)+\frac{2}{n}\sum_i p_{1n}(z_i;\gamma,\beta)+\binom{n}{2}^{-1}\sum_{i<j}p_{2n}(z_i,z_j;\gamma,\beta),\end{aligned}$$

where p_{2n} is defined implicitly above. By Assumptions 11 and 16, $\{p_{2n}(z_i,z_j;\gamma,\beta)\}$ is Euclidean in a set of the form $\Theta_c\equiv\{(\gamma,\beta):\|(\gamma-\gamma_0,\beta-\beta_0)\|\le c\}$ for some constant c, and satisfies $E[\sup_{\Theta_c}p_{2n}(z_i,z_j;\gamma,\beta)^2]<\infty$. Applying theorem 3 of Sherman (1994) to the function $h^L p_{2n}(z_i,z_j;\gamma,\beta)$ we obtain

$$\sup_{\Theta_c}\binom{n}{2}^{-1}\sum_{i<j}h_n^L p_{2n}(z_i,z_j;\gamma,\beta)=O_p\left(\frac{1}{n}\right).$$

Consequently,

$$\begin{aligned}
&\hat{G}_n(\gamma, \beta) \\
&\quad = p_{0n}(\gamma, \beta) + \frac{2}{n}\sum_i p_{1n}(z_i; \gamma, \beta) + \frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n} \tilde{D}_n(w_i; \gamma, \beta) + \tilde{\tilde{e}}_n(z_i, z_j; \gamma, \beta) \\
&\quad \equiv \frac{1}{n}\sum_i \tilde{p}_n(z_i; \gamma, \beta) + \frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n} \tilde{D}_n(w_i; \gamma, \beta) + O_p\left(\frac{1}{nh_n^L}\right) + o_p(n^{-1/2}),
\end{aligned}$$

where the last equality follows from Assumption 6 and the Euclidean property. Denote $\theta \equiv (\beta, \gamma)'$. A first-order approximation yields

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_h) = &-\left(\frac{1}{n}\sum_i \Delta\tilde{p}_n^{\beta}(z_i; \theta^*) + \frac{1}{n}\sum_i \Delta\tilde{D}_n^{\beta}(z_i; \theta^{**})\right)^{-1} \\
&\times \left[\left(\frac{1}{n}\sum_i \Delta\tilde{p}_n^{\gamma}(z_i; \theta^*) + \frac{1}{n}\sum_i \Delta\tilde{D}_n^{\gamma}(z_i; \theta^{**})\right)\sqrt{n}(\hat{\gamma} - \gamma_0)\right. \\
&\quad + \frac{2}{\sqrt{n}}\sum_i p_{1n}(z_i, \gamma_0, \beta_h) + \frac{2(n-2)}{n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{D}_n(w_i; \gamma_0, \beta_h) \\
&\quad \left. + o_p(1) - \sqrt{n}\hat{G}_n(\hat{\gamma}, \hat{\beta})\right].
\end{aligned}$$

Note that $p_{0n}(z_i, \gamma_0, \beta_h) = 0$ by definition of β_h. Given this, the result follows by noting that $\frac{1}{n}\sum_i \Delta\tilde{D}_n^{\beta}(z_i; \theta^{**}) \xrightarrow{p} \Delta\tilde{D}_0^{\beta}(\gamma_0, \beta_0) = 0$. ■

Under additional assumptions, Lemma 3—below—provides more precise expressions for the asymptotic variance described above. In a number of applications the true value of γ is known. If this is the case, the following result follows immediately from Theorem 4.

COROLLARY 2. *If the true value of γ is known—e.g., if $\mu(w, \gamma) = \tau(w) - E[\eta(w) \mid w_2]$, with $\tau(\cdot)$ and $\eta(\cdot)$ known, then if $\hat{\beta}$ is a consistent estimator of β, $G_n(\hat{\beta}) = o_p(n^{-1/2})$ and Assumptions 1–3, 6, 7, and 11–16 are satisfied, we have*

$$\sqrt{n}(\hat{\beta} - \beta_h) \xrightarrow{d} \mathcal{N}(0, E[\psi_i \psi_i']),$$

where $\psi_i = -2\Delta\tilde{p}_0^{\beta}(\beta_0)^{-1} p_1(z_i; \beta_0) - 2\Delta\tilde{p}_0^{\beta}(\beta_0)^{-1} D(w_i; \beta_0)$.

3.7. *Verifying Some of the Conditions.* Theorem 4 makes some high level assumptions. In this section, we present some results that are useful in verifying

these assumptions. The following lemma, which follows immediately from lemma 1 in Honoré and Powell (1994), is useful for verifying that $\hat{G}_n(\hat{\gamma}, \hat{\beta}) = o_p(n^{-1/2})$.

LEMMA 1. *If the true parameter value $\beta_0 \in \mathbb{R}^K$ is an interior point in the parameter space, and*

i. *$S(v_i, v_j, \beta)$ is left-and-right differentiable in each component of β in some open neighborhood of the true parameter β_0,*

ii. *in an open neighborhood B_0 of β_0,*

$$\sup_{\beta \in B_0} \sum_{i<j} \mathbb{1}\left\{ \frac{\partial^- s(v_i, v_j, \beta)}{\partial \beta_\ell} \neq \frac{\partial^+ s(v_i, v_j, \beta)}{\partial \beta_\ell} \right\} = O_p(1); \quad \ell = 1, \ldots, K,$$

iii. *in an open neighborhood of β_0,*

$$\left| \frac{\partial^- s(v_i, v_j, \beta)}{\partial \beta_\ell} - \frac{\partial^+ s(v_i, v_j, \beta)}{\partial \beta_\ell} \right| \leq h(v_i, v_j); \quad \ell = 1, \ldots, K,$$

for some function h with $E[h(v_i, v_j)^{1+\delta}] < \infty$ for some δ; and

iv. *K is bounded, then*

$$\hat{G}_n(\hat{\gamma}, \hat{\beta}) = o_p\big(n^{-2+2/(1+\delta)} h_n^{-L}\big).$$

Next we provide some sufficient conditions under which Assumption 16 is satisfied. We will employ the usual dominance conditions. Define

$$\begin{aligned}
\ell_t(z_i, a, b) &= E[t(v_i, v_j, b) \mid z_i, \mu(w_j, \gamma_0) = a] \rho(a) f_\mu(a) \\
\bar{\ell}_t(z_i, a_1, a_2, b) &= E[t(v_i, v_j, b) \mid z_i, x_j = a_1, \mu(w_j, \gamma_0) = a_2] \bar{\rho}(a_1, a_2) f_{\mu \mid x}(a_2; a_1) \\
\ell_{t_1}(z_i, a, b) &= E\left[\frac{\partial(\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0))}{\partial \gamma} t(v_i, v_j, b) \mid z_i, \mu(w_j, \gamma_0) = a \right] \\
&\quad \times \rho(a) f_\mu(a) \\
\bar{\ell}_{t_1}(z_i, a_1, a_2, b) &= E\left[\frac{\partial(\mu(w_i, \gamma_0) - \mu(w_j, \gamma_0))}{\partial \gamma} t(v_i, v_j, b) \mid z_i, x_j = a_1, \right. \\
&\qquad \left. \mu(w_j, \gamma_0) = a_2 \right] \bar{\rho}(a_1, a_2) f_{\mu \mid x}(a_2; a_1).
\end{aligned}$$

The following condition will correspond to the case in which the exclusion restriction in Assumption 5 holds:

ASSUMPTION 17. *ℓ_t is differentiable with respect to its second and third argument, ℓ_{t_1} is differentiable with respect to its second argument, and there exists a function g with $E[g(z_i)^2] < \infty$ such that*

$$\begin{aligned}\text{Max}\Big\{&\big|\ell_t^{(2)}(z_i,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|,\big|\ell_t^{(3)}(z_i,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|,\\&\big|\ell_{t_1}^{(2)}(z_i,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|\Big\}\\&\le g(z_i),\quad\text{and}\quad\lim_{\|\psi\|\to\infty}K(\psi)\cdot\ell_t^{(2)}(v_i,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)=0.\end{aligned}$$

If Assumption 5′ holds, we modify Assumption 17 in the following way:

ASSUMPTION 17′. *$\bar{\ell}_t$ is differentiable with respect to its third and fourth arguments, $\bar{\ell}_{t_1}$ is differentiable with respect to its third argument, and there exists a function $\bar{g}$ with $E[\bar{g}(z_i)^2]<\infty$ such that*

$$\begin{aligned}\text{Max}\Big\{&\big|\bar{\ell}_t^{(2)}(z_i,x_j,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|,\big|\bar{\ell}_t^{(3)}(z_i,x_j,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|,\\&\big|\bar{\ell}_{t_1}^{(2)}(z_i,x_j,\mu(w_i,\gamma_0)-h_n\psi,\beta_0)\big|\Big\}\\&\le g(z_i),\lim_{\|\psi\|\to\infty}K(\psi)\cdot\bar{\ell}_t^{(2)}\big(z_i,x_j,\mu(w_i,\gamma_0)-h_n\psi,\beta_0\big)=0.\end{aligned}$$

We have the following result.

LEMMA 2. *Let*

$$\begin{aligned}&p_0^{\beta}(\gamma_0,\beta_0)\\&\quad=E\big[\ell_t^{(3)}(z_i,\mu(w_i,\gamma_0),\beta_0)\big],\quad p_0^{\gamma}(\gamma_0,\beta_0)=-E\big[\ell_{t_1}^{(2)}(z_i,\mu(w_i,\gamma_0),\beta_0)\big]\\&\bar{p}_0^{\beta}(\gamma_0,\beta_0)\\&\quad=E\big[\bar{\ell}_t^{(3)}(z_i,x_j,\mu(w_i,\gamma_0),\beta_0)\big],\quad \bar{p}_0^{\gamma}(\gamma_0,\beta_0)=-E\big[\bar{\ell}_{t_1}^{(2)}(z_i,x_j,\mu(w_i,\gamma_0),\beta_0)\big].\end{aligned}$$

Then,

i. *Under Assumptions 4, 5, 6, 7, and 17, $p_{0n}^{\beta}(\gamma_0,\beta_0)\to p_0^{\beta}(\gamma_0,\beta_0)$ and $p_{0n}^{\gamma}(\gamma_0,\beta_0)\to p_0^{\gamma}(\gamma_0,\beta_0)$;*
ii. *Under Assumptions 4′, 5′, 6, 7, and 17′, $p_{0n}^{\beta}(\gamma_0,\beta_0)\to\bar{p}_0^{\beta}(\gamma_0,\beta_0)$ and $p_{0n}^{\gamma}(\gamma_0,\beta_0)\to\bar{p}_0^{\gamma}(\gamma_0,\beta_0)$.*

The previous result implies that Assumption 16(i.a) is satisfied. The next lemma provides sufficient conditions for Assumptions 16(i.b) and 16(iv) to hold.

LEMMA 3. *Let $\tilde{D}_n$ and $\Delta^{\beta}\tilde{D}_n$ be as defined in Assumption 16(iii). Suppose that $p_{1n}(z_i,\gamma_0,\cdot)$ is continuously differentiable in a neighborhood $N(\beta_0)$ of β_0, and that there exists a function $h(z_i)$ with $E[\|h(z_i)\|^2]<\infty$ such that*

$$\text{Max}\big\{\big\|\Delta^{\beta}p_{1n}(z_i,\gamma_0,b)\big\|,\big\|\Delta^{\beta}\tilde{D}_n(w_{2i};\gamma_0,b)\big\|\big\}\le h(z_i)\,\forall b\in N(\beta_0).$$

Then,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{p}_n(z_i;\gamma_0,\beta_h)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_{1n}(z_i;\gamma_0,\beta_0)=o_p(1)$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{D}_n(w_i;\gamma_0,\beta_h)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{D}_n(w_i;\gamma_0,\beta_0)=o_p(1).$$

For the first result, note that $p_{0n}(\gamma_0,\beta_h)=0$ *by definition of* β_h. *Therefore, under Assumptions 4–7 and 17*,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{p}_n(z_i;\gamma_0,\beta_h)-\ell_t(z_i,\mu(w_i,\gamma_0),\beta_0)]=o_p(1)$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[\tilde{D}_n(w_i;\gamma_0,\beta_h)-E\big[\ell_t^{(2)}(zi,\mu(w_i,\gamma_0),\beta_0)\,\big|\,w_{2i}\big]$$
$$\cdot\big(\mu(w_i,\gamma_0)-\tau(w_i,\gamma_0)-\eta(w_i,\gamma_0)\big)\big]=o_p(1),$$

and under Assumptions 4′, 5′, 6, 7, and 17′,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{p}_n(z_i;\gamma_0,\beta_h)-E[\bar{\ell}_t(z_i,x_j,\mu(w_i,\gamma_0),\beta_0)\,|\,z_i]]=o_p(1)$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[\tilde{D}_n(w_i;\gamma_0,\beta_h)-E\big[\bar{\ell}_t^{(2)}(zi,x_j,\mu(w_i,\gamma_0),\beta_0)\,\big|\,w_{2i}\big]$$
$$\cdot(\mu(w_i,\gamma_0)-\tau(w_i,\gamma_0)-\eta(w_i,\gamma_0))\big]=o_p(1).$$

The previous result provides more precise expressions for the asymptotic distribution results in Theorem 4 and Corollary 2. They are also helpful in showing how to estimate the corresponding standard errors.

3.8. *Bias Reduction.* The asymptotic normality result for $\hat{\beta}$ in Theorem 4 centers the asymptotic distribution of $\hat{\beta}$ at the pseudo-true value β_h. As the proof of Theorem 4 shows, the "contribution" to the bias of $\hat{\beta}$ derived from the need to estimate the unknown function $\mu(\cdot)$ nonparametrically is at most of order $o_p(n^{-1/2})$. This is a result of using a higher-order bias reducing kernel $H(\cdot)$ in the construction of $\hat{\mu}_n(\cdot)$, along with the results of Theorem 1. Nevertheless, the pseudo-true value β_h need not converge to the true value β_0 at a rate faster than $\sqrt{n}$, because of the interaction of the bandwidth sequence h_n and the kernel $K(\cdot)$ in the estimation criterion $T_n(\cdot)$; as in Honoré and Powell (2005), this would be the case even if the control variable $\mu(w_i,\gamma_0)$ were observable. The usual approach to ensuring that $\sqrt{n}(\beta_h-\beta_0)=o(1)$ would use a higher-order bias reducing kernel $K(\cdot)$ to ensure $\sqrt{n}$-consistency, but such a requirement would be unattractive for

the kind of estimators proposed here.[8] The resulting negativity of the kernel function for some data points could compromise the convexity of the corresponding minimand, complicating both the asymptotic theory (through an additional compactness restriction) and computation of the estimator. An alternative to the use of higher-order kernels was proposed by Honoré and Powell (2005), which was based upon the familiar jackknife approach. Specifically, assuming the pseudo-true value β_h is a sufficiently smooth function of the bandwidth h_n, it is possible to construct a linear combination $\hat{\beta}$ of different estimators $\hat{\beta}$ of β_0 (involving different choices of the bandwidth h_n, each of which satisfies our assumptions) for which the corresponding linear combination β_n^* of pseudo-true values satisfies $\sqrt{n}(\beta_n^* - \beta_0) = o(1)$; furthermore, since the different estimators have the same linear representation (to order $o(n^{-1/2})$), the "jackknifed" estimator $\hat{\beta}$ will have the same asymptotic distribution as each $\hat{\beta}$, i.e., $\sqrt{n}(\hat{\beta} - \beta_h^*) = \sqrt{n}(\hat{\beta} - \beta_h) + o_p(1)$. It follows that

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta_0) &= \sqrt{n}\big(\hat{\beta} - \beta_n^*\big) + \sqrt{n}\big(\beta_n^* - \beta_0\big) = \sqrt{n}(\hat{\beta} - \beta_h) + o_p(1) + o(1) \\ &= \sqrt{n}(\hat{\beta} - \beta_h) + o_p(1),\end{aligned}$$

so the jackknifed estimator $\hat{\beta}$ will be $\sqrt{n}$-consistent and asymptotically normal, with the same asymptotic distribution as given in Theorem 4. Algebraic details of the construction of the jackknifed estimator $\hat{\beta}$ can be found in Honoré and Powell (2005), Section 3.3.

3.9. *Consistency and Asymptotic Normality without Exclusion Restrictions.* We now present conditions under which we can replicate the previous results without Assumptions 5 (or 5′), which were introduced in order to preserve consistency of $\hat{\beta}$ in the presence of trimming. Take a sequence of trimming sets $\mathcal{W}_{2_n}$. Basically, we could drop the exclusion restrictions if we were able to make the trimming set converge asymptotically to $\mathbb{S}(w_2)$ while preserving the results of Theorem 1 for the entire sequence $\mathcal{W}_{2_n}$. We now outline a set of conditions under which this is possible.

Take a sequence $\varsigma_n \to 0$. Assume that the trimming function is given by

$$a_n(w_2) = \begin{cases} \phi_n(w_2) > 0 & \text{if } \hat{f}_{w_{2_n}}(w_2) > \varsigma_n, \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{f}_{w_{2_n}}(\cdot)$ is the nonparametric estimator of $f_{w_2}(\cdot)$ used in the construction of $\hat{\mu}_n(\cdot)$. Assume that $\phi_n(\cdot)$ is bounded, strictly positive, and M times differentiable with bounded derivatives for all n, and $\phi_n(\cdot) \to c$ for some $c > 0$. Define $\mathcal{W}_{2n} = \{u \in \mathbb{R}^{L_2} : f_{w_2}(u) \geq \varsigma_n\}$ and $\bar{w}_n = \sup_{\mathcal{W}_{2n}} \|u\|$.

[8] Using higher-order kernels to achieve $\sqrt{n}$-consistency would be an attractive option if the pairwise-difference estimator has a closed form expression. See, for example, Aradillas-Lopez (2006).

ASSUMPTION 18.

i. All the properties stated for the set $\mathcal{W}_2$ in Assumptions 1 and 2 are satisfied by any compact subset in the interior of $\mathbb{S}(w_2)$.

ii. Strengthen Assumption 6 and assume that $n^{1-2\varepsilon} b_n^{L_2} h_n^{2(L+2)} \varsigma_n^2 \to \infty$ for some $\varepsilon > 0$.

iii. We will modify Assumption 3(iii) and assume that ς_n converges to zero sufficiently slow relative to b_n such that $b_n t + w_2 \in \text{interior}\{\mathbb{S}(w_2)\} \forall t \in \mathcal{H}, w_2 \in \mathcal{W}_{2n}, n \in \mathbb{N}$. Let $\rho_n(\mu(w_i, \gamma_0)) = E[a_n(w_{2i}) \mid \mu(w_i, \gamma_0)]$. We will generalize Assumption 4(v) and assume that $\rho_n(\mu(w_i, \gamma_0))$ is continuous and strictly positive for all $w_{2i} \in \mathcal{W}_{2n}$. We also assume that the tails of $f_{w_2}(\cdot)$ are such that $\ln(\bar{w}_n) = o_p(n^{\varepsilon})$, where ε is defined in the previous assumption.

The next result establishes consistency of $\hat{\beta}$ without Assumption 5 and it also shows that the asymptotic normality results still hold.

THEOREM 5. *Suppose we use the sequence of trimming functions $a_n(\cdot)$ described above. We have the following results:*

i. If Assumptions 1–4, 6–10, and 18 are satisfied, then $\hat{\beta}$ satisfies the consistency results of Theorems 2 and 3.

ii. Suppose Assumption 14 is satisfied for any compact set in the interior of $\mathbb{S}(w_2)$. Then, if Assumptions 1–3, 6–7, 11–13, and 15–16, and 18 are satisfied, the conclusions of Theorem 4 are true with $v_n(z_i, z_j; \gamma, \beta)$ and $p_n(z_i, z_j; \gamma, \beta)$ replaced by

$$\bar{v}_n(z_i, z_j; \gamma, \beta) = c * \frac{1}{h_n^L} K\left(\frac{\hat{\mu}_n(w_i, \gamma) - \hat{\mu}_n(w_j, \gamma)}{h_n}\right) t(v_i, v_j, \beta)$$

$$\bar{p}_n(z_i, z_j; \gamma, \beta) = c * \frac{1}{h_n^L} K\left(\frac{\mu(w_i, \gamma) - \mu(w_j, \gamma)}{h_n}\right) t(v_i, v_j, \beta)$$

with $c > 0$ being the limit of the sequence of functions $a_n(\cdot)$.

The main steps of the proof are included in the Appendix.

4. EXAMPLE: DISCRETE CHOICE MODEL WITH ENDOGENEITY

Let us revisit more formally the example briefly described in Section 2.5.2. Suppose we have

$$y_i = \mathbb{1}\{x_i' \beta_0 + \varepsilon_i \geq 0\},$$

and let $x_i = (x_{1i}', x_{2i}')'$, $\beta = (\beta_1', \beta_2')'$, and $w_i = (x_{1i}', w_{2i}')'$, where x_{2i} is a subvector of w_{2i}. The subvector $w_{1i} \equiv x_{1i}$ of regressors is assumed to be endogenous, and can be characterized by a (partially known) reduced form

$$x_{1i} = E[x_{1i} \mid w_{2i}] + \mu_i,$$

where $E[x_{1i} \mid w_{2i}]$ is of unknown functional form. Using the notation from Section 3, we have $\mu_i \equiv \mu(w_i) = x_{1i} - E[x_{1i} \mid w_{2i}]$ and $\tau(w_i, \gamma) \equiv x_{1i} \equiv \eta(w_i, \gamma)$ for this example. Following Blundell and Powell (2004), suppose we model endogeneity in this model by assuming that the dependence between the structural error term ε_i on the vector of regressors x_i and the "instrumental variables" w_{2i} is completely characterized by the reduced-form residuals μ_i; specifically, suppose we model the structural error term ε_i as additively separable in the reduced-form error term μ_i and a logistic error term ζ_i,

$$\varepsilon_i = g(\mu_i) + \zeta_i,$$

as in Blundell and Smith (1989, which also takes $g(\cdot)$ to be linear). Assuming ζ_i is independent of w_i, so that

$$\varepsilon_i | x_i, w_i \sim \varepsilon_i | \mu_i,$$

the structural model for the binary variable y_i can be rewritten as

$$y_i = \mathbb{1}\{x_i'\beta_0 + g(\mu_i) + \zeta_i \geq 0\},$$

which reduces to the partially linear logit model discussed in Section 2.2 and analyzed in Ai and McFadden (1997) and Honoré and Powell (2005), albeit with an unknown regressor μ_i in the nonparametric component. Suppose w_{2i} is a continuously distributed random vector with dimension L_2, and let $H(\cdot)$ be the kernel function described in Section 3.2. Define

$$R_n(\omega_2) = \frac{1}{nh_n^{L_2}} \sum_{j=1}^{n} x_{1j} H_{b_n}(w_{2j} - \omega_2), \quad \hat{f}_{w_{2n}}(\omega_2) = \frac{1}{nh_n^{L_2}} \sum_{j=1}^{n} H_{b_n}(w_{2j} - \omega_2), \quad \text{and}$$

$$\hat{\mu}_n(w_i) = x_{1i} - \frac{R_n(w_{2i})}{\hat{f}_{w_{2n}}(w_{2i})};$$

the estimator $\hat{\mu}_n(w_i)$ fits the description of Section 3.2 with $\tau(w_i; \gamma) = w_{1i} \equiv x_{1i} = \eta(w_{1i}, w_{2i}, \gamma)$. The conditions of Assumption 2 will be satisfied if $\mu(w_i)$ is M times differentiable with bounded derivatives, $E[w_{1i}^2 \mid w_{2i}]$ exists and is a continuous function of w_{2i}, and $E[w_{1i}^4] < \infty$. In addition, if the smoothness conditions of Assumption 1 concerning the density of w_{2i} are satisfied, along with the kernel and bandwidth properties of Assumption 3, then $\hat{\mu}_n(w_i)$ will satisfy the results of Theorem 1. Define

$$L_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{\substack{i<j \\ y_i \neq y_j}} K\left(\frac{\hat{\mu}_i - \hat{\mu}_j}{h}\right)$$

$$\times [y_i \ln(1 + \exp\{(x_j - x_i)'b\}) + y_j \ln(1 + \exp\{(x_i - x_j)'b\})] a(w_{2i}) a(w_{2j}).$$

Following the discussion in Section 2.2.1 and Equations (4) and (5), we propose to estimate β by minimizing $L_n(b)$. The need to add a trimming function to (5) follows from the detailed discussion in Section 3. The problem

$\hat{\beta} = \text{argmin}_b L_n(b)$ fits the framework analyzed in Section 3 with $s(v_i, v_j, b) = \mathbb{1}\{y_i \neq y_j\} \cdot [y_i \ln(1 + \exp\{(x_j - x_i)'b\}) + y_j \ln(1 + \exp\{(x_i - x_j)'b\})]$; in this setting, it is important to verity if the model fits the exclusion restrictions 5 or 5′. Since $y_i = \mathbb{1}\{x_i'\beta_0 + g(\mu_i) + \zeta_i \geq 0\}$, where $\zeta_i \sim$ i.i.d. logistic and independent of w_i, it follows that $y_j \mid x_j, \mu(w_j), w_{2j} \sim y_j \mid x_j, \mu(w_j)$ and thus if $v_i \equiv (y_i, x_i')'$, we have $E[s(v_i, v_j, b) \mid v_i, x_j, \mu(w_j), w_{2j}] = E[s(v_i, v_j, b) \mid v_i, x_j, \mu(w_j)]$ and the exclusion restriction in Assumption 5′ is satisfied. Assumption 4′ will be satisfied if the density of μ_i conditional on x_i is a continuous function of ζ_i. Assumption 10′ will be satisfied if $E[s_i(v_i, v_j; b) \mid x_i, x_j, \mu_i = \mu_j]$ is uniquely minimized at $b = \beta_0$; given our previous assumptions, this will hold if x_i has full rank. Provided that Assumptions 6, 7, and 9 hold, this would yield consistency of our estimator, and if Assumptions 11–13, 15, and 16 hold, then $\hat{\beta}$ would have the asymptotic distribution described in Corollary 2.[9]

4.1. *A Monte Carlo Study.* Having described the large sample properties of the proposed estimator in Section 3, we devote this subsection to evaluate the performance of a simple implementation of our estimation procedure for the case of a discrete choice model with endogeneity. Using the same notation as above, we have $x_{2i} \sim \mathcal{N}(0, 1)$, $w_{2i} \sim U[-1, 1]$, $\mu_i \sim \sqrt{2}\mathcal{N}(0, 1)$, and $\zeta_i \sim$ logistic; all these variables are independent of each other. x_{1i} and ε_i are given by

$$x_{1i} = \underbrace{\frac{(w_{2i} - 1)^2}{2} - \frac{w_{2i}^3}{4} + \frac{w_{2i}^4}{10} - \frac{\exp(w_{2i})}{(1 + \exp(w_{2i}))} + \sin(4w_{2i})}_{= E[x_{1i} \mid w_{2i}]} + \mu_i,$$

$$\varepsilon_i = \underbrace{\frac{2\mu_i}{\pi} \arctan(\mu_i)}_{= g(\mu_i)} + \zeta_i.$$

The functions chosen are smooth but clearly nonlinear in nature. We have

$$y_i = \mathbb{1}\{-1 + x_{1i} + x_{2i} + \varepsilon_i \geq 0\},$$

so $\beta_{1_0} = 1$ and $\beta_{2_0} = 1$. We focus only on the slope coefficients β_1, β_2 because the intercept is not involved in the identifying moment conditions.[10] We are interested in the properties of a quick, simple implementation of our methods. We estimate $E[x_{1i} \mid w_{2i}]$ simply by using a local polynomial of order six and we choose the standard normal density as the kernel $K(\cdot)$. Our focus here is on the bandwidth h_n. Let R_n denote the interquartile range of $(\hat{E}[x_1 \mid w_{2i}])_{i=1}^n$, let $\hat{\sigma}_{E[x_1|w_2]}$ denote the sample standard deviation of $(\hat{E}[x_1 \mid w_{2i}])_{i=1}^n$, and define $A_n = \min\{\hat{\sigma}_{E[x_1|w_2]}, R_n/1.34\}$. Fix a scalar $c \in \mathbb{R}$ and define $\hat{\beta}(c)$ as the estimator obtained by using $h_n = cA_n n^{-1/5}$. Silverman's "rule of thumb" density estimation bandwidth when the reference

[9] In this model, $\eta(\cdot)$ is not a function of w_{2i} and γ_0 is trivially known. This makes a number of the conditions studied in Section 3 (e.g., Assumptions 8 and 14) irrelevant.

[10] Recall that the function $g(\cdot)$ is assumed to be unknown.

TABLE 1
SIMULATION RESULTS FOR $\hat{\beta}_1$ (TRUE VALUE IS $\beta_{1_0} = 1$). NUMBER OF SIMULATIONS = 1000

	$n = 150$				$n = 450$			
	$\lvert\text{Bias}(\hat{\beta}_1)\rvert$	$\widehat{MSE}(\hat{\beta}_1)$	$\hat{\beta}_{1_{(0.025)}}$	$\hat{\beta}_{1_{(0.975)}}$	$\lvert\text{Bias}(\hat{\beta}_1)\rvert$	$\widehat{MSE}(\hat{\beta}_1)$	$\hat{\beta}_{1_{(0.025)}}$	$\hat{\beta}_{1_{(0.975)}}$
$\hat{\beta}_1^I$	0.10326	0.37230	0.04832	2.41428	0.02729	0.07156	0.53087	1.58677
$\hat{\beta}_1^{II}$	0.03594	0.10319	0.52705	1.78713	0.00576	0.02136	0.72904	1.29874
$\hat{\beta}_1^{III}$	0.06242	0.16345	0.36515	1.98202	0.01182	0.03691	0.65136	1.42428
	$n = 700$				$n = 1000$			
	$\lvert\text{Bias}(\hat{\beta}_1)\rvert$	$\widehat{MSE}(\hat{\beta}_1)$	$\hat{\beta}_{1_{(0.025)}}$	$\hat{\beta}_{1_{(0.975)}}$	$\lvert\text{Bias}(\hat{\beta}_1)\rvert$	$\widehat{MSE}(\hat{\beta}_1)$	$\hat{\beta}_{1_{(0.025)}}$	$\hat{\beta}_{1_{(0.975)}}$
$\hat{\beta}_1^I$	0.00626	0.04195	0.59200	1.42109	0.00992	0.02735	0.69016	1.33034
$\hat{\beta}_1^{II}$	0.00631	0.01288	0.77812	1.22448	0.00852	0.00975	0.80565	1.19510
$\hat{\beta}_1^{III}$	0.00244	0.02276	0.71927	1.30665	0.00293	0.01622	0.75708	1.25162

population is a standard normal is a special case, with $c = 0.90$ (see equation 3.31, p. 48 in Silverman, 1986).

One of our goals in this section is to evaluate the sensitivity of our estimator to the actual bandwidth chosen; we proceed as follows. Let $c_1 = 0.4$ and $c_j = c_{j-1} + 0.5$ for $j \geq 2$. We use a grid of scalar values $(c_1, \ldots, c_M)$ with $M = 5$ (i.e., $c_M = 2.4$). We estimate $\hat{\beta}$ in three different ways, all of which would have the same asymptotic linear representation and $\sqrt{N}$-distribution.[11]

$$\hat{\beta}^I = \hat{\beta}(c_1), \quad \hat{\beta}^{II} = \hat{\beta}(c_M), \quad \hat{\beta}^{III} = \frac{1}{M}\sum_{\ell=1}^{M} \hat{\beta}(c_\ell), \tag{21}$$

so $\hat{\beta}^I$ chooses the small bandwidth, $\hat{\beta}^{II}$ chooses the large bandwidth, and $\hat{\beta}^{III}$ takes a simple average over the bandwidths in our grid. We trimmed out the observations with the highest 5% of values for $|\hat{E}[x_1 \mid w_{2i}]|$. Tables 1 and 2 present the results of 1000 simulations for each one of the estimators described above.

As we should expect, our simulation results show that the actual choice of bandwidth has a relatively more important effect on smaller samples. In this particular case, using the larger bandwidth in our arbitrary grid yielded better results than those of the smaller bandwidth. The difference became less relevant for larger sample sizes. The results also illustrate that taking a simple average of the estimators in the bandwidth grid is a good "rule of thumb" to safeguard against the problem of finding the "correct bandwidth" to commit to. In fact, the absolute bias of $\hat{\beta}^{III}$ was smaller than those of $\hat{\beta}^I$ and $\hat{\beta}^{II}$ for sample sizes $N = 700$ and 1000.

The individual mean squared errors decreased steadily with the sample size in all cases, and their magnitudes also became closer across the three estimators

[11] Any weighted average of the resulting estimators would have the same Bahadur representation as any individual estimator (see Theorem 1), as long as the weights add up to one.

TABLE 2

SIMULATION RESULTS FOR $\hat{\beta}_2$ (TRUE VALUE IS $\beta_{2_0} = 1$). NUMBER OF SIMULATIONS = 1000

	$n = 150$				$n = 450$			
	$\lvert\text{Bias}(\hat{\beta}_2)\rvert$	$\widehat{MSE}(\hat{\beta}_2)$	$\hat{\beta}_{2_{(0.025)}}$	$\hat{\beta}_{2_{(0.975)}}$	$\lvert\text{Bias}(\hat{\beta}_2)\rvert$	$\widehat{MSE}(\hat{\beta}_2)$	$\hat{\beta}_{2_{(0.025)}}$	$\hat{\beta}_{2_{(0.975)}}$
$\hat{\beta}_2^I$	0.09622	0.20209	0.43348	2.04395	0.01974	0.02944	0.71601	1.37653
$\hat{\beta}_2^{II}$	0.04243	0.12874	0.45047	1.79982	0.00587	0.02497	0.70021	1.31226
$\hat{\beta}_2^{III}$	0.06454	0.14804	0.43502	1.89152	0.00696	0.02667	0.71132	1.32944
	$n = 700$				$n = 1000$			
	$\lvert\text{Bias}(\hat{\beta}_2)\rvert$	$\widehat{MSE}(\hat{\beta}_2)$	$\hat{\beta}_{2_{(0.025)}}$	$\hat{\beta}_{2_{(0.975)}}$	$\lvert\text{Bias}(\hat{\beta}_2)\rvert$	$\widehat{MSE}(\hat{\beta}_2)$	$\hat{\beta}_{2_{(0.025)}}$	$\hat{\beta}_{2_{(0.975)}}$
$\hat{\beta}_2^I$	0.00360	0.01573	0.76393	1.26773	0.01001	0.01085	0.82532	1.22939
$\hat{\beta}_2^{II}$	0.01445	0.01395	0.75836	1.22905	0.00293	0.00989	0.81105	1.20510
$\hat{\beta}_2^{III}$	0.00452	0.01462	0.76070	1.24420	0.00159	0.01021	0.81711	1.22166

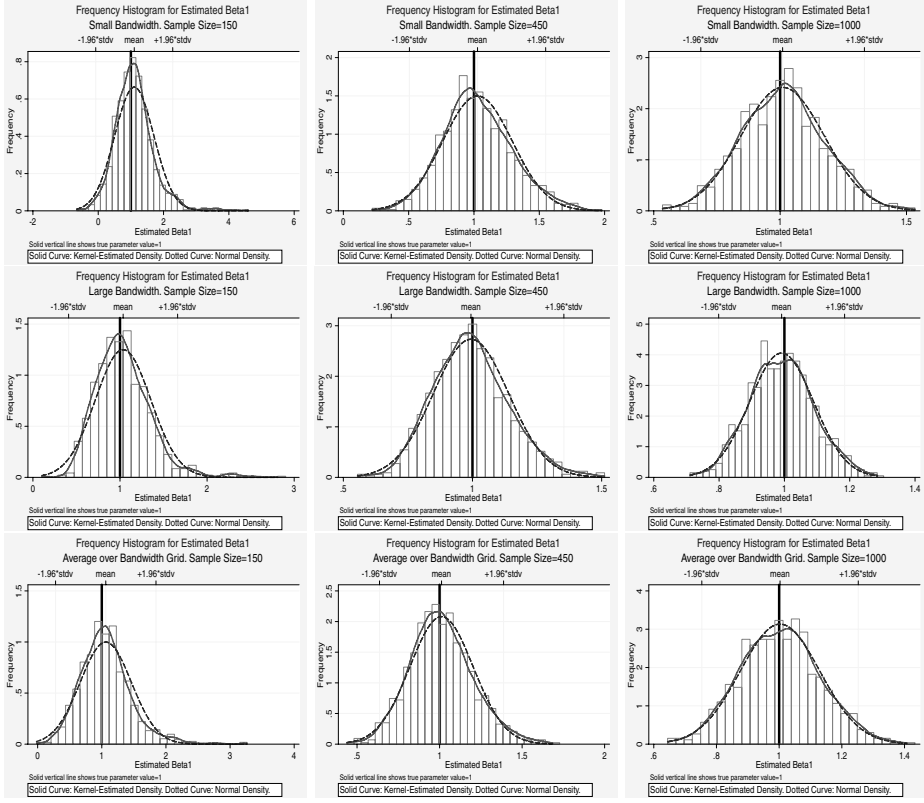

FIGURE 1

HISTOGRAM AND DENSITIES FOR SIMULATED ESTIMATES OF β_1. ROWS SUMMARIZE RESULTS FOR $\hat{\beta}_1^I$, $\hat{\beta}_1^{II}$, AND $\hat{\beta}_1^{III}$, RESPECTIVELY. COLUMNS SUMMARIZE RESULTS FOR $N = 150, 450$, AND 1000, RESPECTIVELY

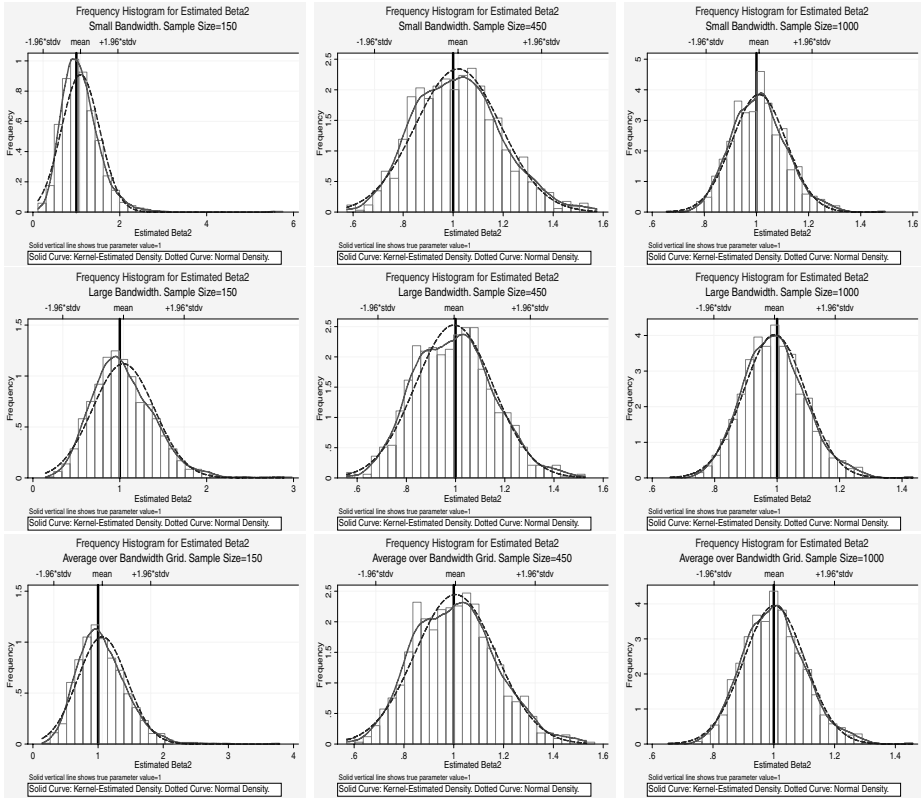

FIGURE 2

HISTOGRAM AND DENSITIES FOR SIMULATED ESTIMATES OF β_2. ROWS SUMMARIZE RESULTS FOR $\hat{\beta}_2^I$, $\hat{\beta}_2^{II}$, AND $\widehat{\beta}_2^{III}$, RESPECTIVELY. COLUMNS SUMMARIZE RESULTS FOR $N = 150, 450$, AND 1000, RESPECTIVELY

as the sample size increased. As Figures 1 and 2 show, the asymptotic Normal approximation becomes more precise as the sample size increases for our three estimators. For smaller sample sizes ($N = 150$ and 450), a normal approximation appears relatively more accurate for $\hat{\beta}_1$, the coefficient of the endogenous regressor, than for $\hat{\beta}_2$, the coefficient of the exogenous regressor.[12]

The normal approximation in both cases appears indistinguishably accurate for $N = 1000$. In the majority of cases, a normal approximation seemed more accurate for $\hat{\beta}^{III}$ than for the other two estimators. As with all other results, the difference seemed less obvious as the sample size was increased. Overall, taking a simple average over the estimators that result from a grid of bandwidths seems to lead to a nice balance between bias and dispersion. In our case, this was true even for a simple average using uniform weights.

[12] Note however that the absolute bias and the MSE of $\hat{\beta}_1$ were slightly larger than those of $\hat{\beta}_2$ in the majority of cases.

5. CONCLUDING REMARKS

Econometric models amenable to pairwise-differencing estimation arise in a variety of contexts ranging from sample selection and/or endogeneity to microeconomic models with rational expectations with or without strategic interactions. As this article showed, even if the control function involved in the identifying moment restriction is of unknown functional form, $\sqrt{N}$-consistent estimation is possible. As we argued at length, special care must be placed on the issues of trimming and its implications on identification. As it is always the case in semi or nonparametric estimation procedures, the choice of bias-reducing techniques is up to the researcher. Bias reduction could be done through the density estimator (using bias-reducing kernels) or through the bandwidth (taking advantage of the fact that any estimator that uses a valid bandwidth sequence will have the same asymptotic linear representation). We advocate the latter approach here. A simple Monte Carlo study showed that even using a naive, simple average of estimators over a grid of bandwidths could serve as a simple way of achieving a good balance between bias and dispersion without the need to search to the "correct" finite sample bandwidth. The asymptotic standard errors we found would be valid for such an estimator. As the sample size grows, the actual bandwidth chosen becomes relatively less important.

APPENDIX

A.1. *Steps to Prove Theorem 4.* Define

$$S_{1n}(z_i, z_j; \gamma, \beta) = \frac{1}{h_n^{L+1}} K^{(1)}\left(\frac{\mu(w_i, \gamma) - \mu(w_j, \gamma)}{h_n}\right) t(v_i, v_j, \beta) a(w_{2i}) a(w_{2j})$$

$$S_{2n}(z_i, z_j; \gamma) = \frac{[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_j, \gamma)]}{b_n^{L_2} f_{w_2}(w_{2j})} H_{b_n}(w_{2i} - w_{2j}).$$

and let

$$\begin{aligned}
&T_{1n}(z_i, z_j; \gamma, \beta) \\
&\quad = S_{1n}(z_i, z_j; \gamma, \beta)'[S_{2n}(z_i, z_i; \gamma, \beta) - S_{2n}(z_i, z_j; \gamma, \beta) + S_{2n}(z_j, z_i; \gamma, \beta) \\
&\qquad\qquad - S_{2n}(z_j, z_j; \gamma, \beta)] \\
&T_{2n}(z_i, z_j, z_k; \gamma, \beta) \\
&\quad = S_{1n}(z_i, z_j; \gamma, \beta)'[S_{2n}(z_k, z_i; \gamma, \beta) - S_{2n}(z_k, z_j; \gamma, \beta)] + S_{1n}(z_i, z_k; \gamma, \beta)' \\
&\qquad \times [S_{2n}(z_j, z_i; \gamma, \beta) - S_{2n}(z_j, z_k; \gamma, \beta)] + S_{1n}(z_j, z_k; \gamma, \beta)' \\
&\qquad \times [S_{2n}(z_i, z_j; \gamma, \beta) - S_{2n}(z_i, z_k; \gamma, \beta)].
\end{aligned}$$

Note that $T_{1n}(z_i, z_j; \gamma, \beta)$ and $T_{2n}(z_i, z_j, z_k; \gamma, \beta)$ are symmetric in i, j and in i, j, k, respectively. We have the following result.

CLAIM 1. *If Assumptions 1–3, 6, 7, and 11 are satisfied and* $\{z_i, i = 1, \ldots, n\}$ *is an i.i.d. sample, then*

$$\begin{aligned}
&\hat{G}_n(\gamma, \beta) \\
&\quad = G_n(\gamma, \beta) + \binom{n}{2}^{-1} \frac{1}{n} \sum_{i<j} T_{1n}(z_i, z_j; \gamma, \beta) + \binom{n}{2}^{-1} \frac{1}{n} \sum_{i<j<k} T_{2n}(z_i, z_j, z_k; \gamma, \beta) \\
&\qquad + c_n(z_i, z_j; \gamma, \beta),
\end{aligned}$$

where $\sup_{\substack{i,j \\ B,\Gamma}} |c_n(z_i, z_j; \gamma, \beta)| = o_p(n^{-1/2})$.

PROOF. Using Assumption 7, we have

$$\begin{aligned}
&\frac{1}{h_n^L}\left[K\left(\frac{\hat{\mu}_n(w_i, \gamma) - \hat{\mu}_n(w_j, \gamma)}{h_n}\right) - K\left(\frac{\hat{\mu}_n(w_i, \gamma) - \hat{\mu}_n(w_j, \gamma)}{h_n}\right)\right] \\
&\quad \leq \frac{1}{h_n^{L+1}} K^{(1)}\left(\frac{\mu(w_i, \gamma) - \mu(w_j, \gamma)}{h_n}\right)' \\
&\qquad \times \left([\hat{\mu}_n(w_i, \gamma) - \mu(w_i, \gamma)] - [\hat{\mu}_n(w_i, \gamma) - \mu(w_i, \gamma)]\right) \\
&\qquad + \frac{2L}{h_n^{L+2}} \left\| K^{(2)}\left(d_{ij}^*\right)\right\| \cdot \left[\|\hat{\mu}_n(w_i, \gamma) - \mu(w_i, \gamma)\|^2 + \|\hat{\mu}_n(w_j, \gamma) - \mu(w_j, \gamma)\|^2\right].
\end{aligned}$$

Assumptions 1–3, 6, 7, and 11 and the nature of the trimming function $a(\cdot)$ imply that

$$\begin{aligned}
&\sup_{\substack{i,j \\ B,\Gamma}} \frac{\left\| K^{(2)}\left(d_{ij}^*\right)\right\|}{h_n^{L+2}} \cdot \left[\|\hat{\mu}_n(w_i, \gamma) - \mu(w_i, \gamma)\|^2\right] \cdot \left| t(v_i, v_j, \beta) \right| \cdot a(w_{2i}) a(w_{2j}) \\
&\quad \leq \frac{C}{n h_n^{L+2}} O_p\left(n^{1-\delta} b_n^{L_2}\right)^{-1} \left|F(z_i, z_j)\right| a(w_{2i}) a(w_{2j})
\end{aligned}$$

for any $\delta > 0$ and some $C > 0$. Using Assumption 6, there exists a $\delta > 0$ such that $(n^{1-\delta} h_n^{L+2} b_n^{L_2})^{-1} = o(n^{-1/2})$. From Assumption 11 and the properties of $a(\cdot)$, we have $E[\{F(z_i, z_j) a(w_{2i}) a(w_{2j})\}^2] = O(n)$. Using lemma A.3 in Ahn and Powell, we have

$$\begin{aligned}
c_{1n}(z_i, z_j) &\equiv \frac{C}{n h_n^{L+2}} O_p\left(n^{1-\delta} b_n^{L_2}\right)^{-1} \binom{n}{2}^{-1} \sum_{i<j} |F(z_i, z_j)| a(w_{2i}) a(w_{2j}) \\
&= \frac{1}{n^{1-\delta} h_n^{L+2} b_n^{L_2}} O_p(1)
\end{aligned}$$

for all $\delta > 0$. Choosing the value of δ in Assumption 6, we obtain $c_{1n}(z_i, z_j) = o_p(n^{-1/2})$. If Assumptions 1–3 hold, Theorem 1 yields

$$\begin{aligned}
&\frac{1}{h_n^{L+1}} K^{(1)}\left(\frac{\mu(w_i,\gamma)-\mu(w_j,\gamma)}{h_n}\right)'(\hat{\mu}_n(w_i,\gamma)-\mu(w_i,\gamma))t(v_i,v_j,\beta)\cdot a(w_{2i})a(w_{2j}) \\
&\quad = \frac{1}{h_n^{L+1}} K^{(1)}\left(\frac{\mu(w_i,\gamma)-\mu(w_j,\gamma)}{h_n}\right)' t(v_i,v_j,\beta)\cdot a(w_{2i})a(w_{2j}) \\
&\qquad \times \left\{\sum_{k=1}^{n} \frac{[\tau(w_i,\gamma)-\eta(w_{1k},w_{2i},\gamma)-\mu(w_i,\gamma)]}{f_{w_2}(w_{2i})nb_n^{L_2}} H_{b_n}(w_{2k}-w_{2i}) + \xi_n(w_{2i},\gamma)\right\},
\end{aligned}$$

where $\sup_{\substack{v\in\mathcal{W}_2 \\ \gamma\in\Gamma}} \|\xi_n(\omega,\gamma)\| = O_p(n^{\delta-1}b_n^{-L_2})$ for any $\delta > 0$. We have

$$\begin{aligned}
&\frac{1}{h_n^{L+1}}\binom{n}{2}^{-1}\sum_{i<j}\sup_{\substack{i,j \\ B,\Gamma}}\left\| K^{(1)}\left(\frac{\mu(w_i,\gamma)-\mu(w_j,\gamma)}{h_n}\right)\right\| |t(v_i,v_j,\beta)| \\
&\quad \cdot a(w_{2i})a(w_{2j})\|\xi_n(w_{2i},\gamma)\| \\
&\quad \le \frac{C}{nh_n^{L+1}} O_p\left(n^{1-\delta}b_n^{L_2}\right)^{-1}\binom{n}{2}^{-1}\sum_{i<j}|F(z_i,z_j)|a(w_{2i})a(w_{2j}) \equiv c_{2n}(z_i,z_j)
\end{aligned}$$

for any $\delta > 0$. By Assumption 6 and the same argument used above, we have $c_{2n}(z_i, z_j) = o_p(n^{-1/2})$. Grouping the terms in the sum, we obtain the result of the claim, with $\sup_{\substack{i,j \\ B,\Gamma}} |c_n(z_i,z_j;\gamma,\beta)\| = D\cdot(c_{1n}(z_i,z_j)+c_{2n}(z_i,z_j))$ for some constant $D > 0$. ■

CLAIM 2. *If Assumptions 1–3, 6, 7, and 11–14 hold and $\{z_i, i = 1,\dots,n\}$ is an i.i.d. sample, then*

$$\begin{aligned}
&\binom{n}{2}^{-1}\frac{1}{n}\sum_{i<j<k} T_{2n}(z_i,z_j,z_k;\gamma,\beta) \\
&\quad = \frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n} B_n(w_{2i};\gamma,\beta)[\tau(w_i,\gamma)-\eta(w_i,\gamma)-\mu(w_i,\gamma)]a(w_{2i}), \\
&\qquad + \tilde{c}_n(z_i,z_j;\gamma,\beta),
\end{aligned}$$

where $\sup_{B,\Gamma}|\tilde{c}_n(z_i,z_j;\gamma,\beta)| = o_p(n^{-1/2})$.

PROOF. Recall from Assumption 1 that $f_w(w_1, w_2)$ denotes the joint density of w_{1i} and w_{2i}. Using iterated expectations, we have

$$
\begin{aligned}
&E[S_{1n}(z_i, z_j; \gamma, \beta)' S_{2n}(z_k, z_i; \gamma, \beta) \mid z_i] \\
&= E\left[\frac{1}{h_n^{L+1}} K^{(1)}\left(\frac{\mu(w_i, \gamma) - \mu(w_j, \gamma)}{h_n}\right)' t(v_i, v_j, \beta) a(w_{2i}) a(w_{2j}) \right. \\
&\qquad \left. \times \frac{1}{b_n^{L_2}} \iint \frac{[\tau(w_i, \gamma) - \eta(u, w_{2i}, \gamma) - \mu(w_i, \gamma)]}{f_{w_2}(w_{2i})} H\left(\frac{v - w_{2i}}{b_n}\right) f_w(u, v)\, du\, dv\right].
\end{aligned}
$$

Define $Q_\ell \equiv \{(q_1, \ldots, q_L) \in \mathbb{N}^{L_2} : q_1 + \cdots + q_{L_2} = \ell\}$ and $\Upsilon_\ell(u, v) = \sum_{Q_\ell} \times \frac{\partial^\ell f_w(u,v)}{\partial w_{2_1}^{q_1} \cdots \partial w_{2_{L_2}}^{q_{L_2}}}$. From Assumption 2(i), there exists $C > 0$ such that $\Upsilon_\ell(u, v) < C$ for all u, v and $\ell = 1, \ldots, M$. From Assumptions 1(ii) and 3(i), the following approximation is valid:

$$
\begin{aligned}
&\int H(\psi) f_w(u, b_n\psi + w_{2i}) d\psi \\
&\quad = f_w(u, w_{2i}) + b_n^M \frac{(-1)^M}{M!} \int \sum_{Q_M} \left(\psi_1^{q_1} \cdots \psi_{L_2}^{q_{L_2}}\right) \Upsilon\left(u, w_{2i} + b_n^* \psi\right) H(\psi)\, d\psi,
\end{aligned}
$$

with $b_n^* \in (0, b_n)$. Assumptions 1(ii) and 3(i) imply that $\sup_u \int |\sum_{Q_M} \times (\psi_1^{q_1} \cdots \psi_{L_2}^{q_{L_2}}) \Upsilon(u, w_{2i} + b_n^* \psi) H(\psi)| d\psi < D$ for some $D > 0$. Therefore, we have

$$
\sup_{B,\Gamma} E[S_{1n}(z_i, z_j; \gamma, \beta)' S_{2n}(z_k, z_i; \gamma, \beta) \mid z_i] \le C \frac{b_n^M}{h_n^{L+1}} E[F(z_i, z_j) a(w_{2i}) a(w_{2j})]
$$

for some $C > 0$. Using Assumptions 6 and 11, $\sup_{B,\Gamma} E[S_{1n}(z_i, z_j; \gamma, \beta)' \times S_{2n}(z_k, z_i; \gamma, \beta) \mid z_i] = o_p(n^{-1/2})$. Using iterated expectations and steps parallel to those used above, we obtain the same result for $E[S_{1n}(z_i, z_j; \gamma, \beta)' \times S_{2n}(z_k, z_j; \gamma, \beta) | z_i]$, $E[S_{1n}(z_i, z_k; \gamma, \beta)' S_{2n}(z_j, z_i; \gamma, \beta) | z_i]$, and $E[S_{1n}(z_i, z_k; \gamma, \beta)' \times S_{2n}(z_j, z_k; \gamma, \beta) \mid z_i]$. Define

$$
\tilde{D}_n(v, u; \gamma, \beta) = B_n(v; \gamma, \beta)\big[\tau(u, v, \gamma) - \eta(u, v, \gamma) - \mu(v, \gamma)\big] a(v).
$$

Using iterated expectations, we have

$$
\begin{aligned}
&E[S_{1n}(z_j, z_k; \gamma, \beta)' S_{2n}(z_i, z_j; \gamma, \beta) \mid z_i] \\
&\quad = E\left[B_n(w_{2j}; \gamma, \beta) \frac{[\tau(w_i) - \eta(w_i, \gamma) - \mu(w_j, \gamma)]}{b_n^{L_2} f_{w_2}(w_{2j})} a(w_{2j}) H_{b_n}(w_{2j} - w_{2i}) \mid z_i\right] \\
&\quad = \int \tilde{D}_n(w_{1i}, b_n\psi + w_{2i}; \gamma, \beta) H(\psi) d\psi.
\end{aligned}
$$

From Assumption 12, $a(\cdot)$ is M times differentiable almost everywhere—the boundary of $\mathcal{W}_2$ is a set of Lebesgue-measure zero in $\mathbb{R}^{L_2}$. Therefore, Assumptions 2(i), 12, and 14 imply that with probability one, $\tilde{D}_n(u, v; \gamma, \beta)$ is M times differentiable with respect to v with bounded derivatives for all $\gamma \in \Gamma$ and $\beta \in B$. Let the

set Q_ℓ be as defined above and denote $\Lambda_\ell(u, v; \gamma, \beta) = \sum_{Q_\ell} \frac{\partial^\ell \tilde{D}_n(u,v;\gamma,\beta)}{\partial v_1^{q_1} \cdots \partial v_{L_2}^{q_{L_2}}}$. An Mth order approximation yields

$$E[S_{1n}(z_j, z_k; \gamma, \beta)' S_{2n}(z_i, z_j; \gamma, \beta) \mid z_i]$$
$$= \tilde{D}_n(w_{1i}, w_{2i}; \gamma, \beta) + b_n^M \frac{(-1)^M}{M!} \int \sum_{Q_M} (\psi_1^{q_1} \cdots \psi_{L_2}^{q_2}) \Lambda(w_{1i}, w_{2i} + b_n^* \psi; \gamma, \beta)\, d\psi.$$

Assumptions 2(i), 3(i), 12, and 14 yield $\sup_{B,\Gamma} | \int \sum_{Q_M} (\psi_1^{q_1} \cdots \psi_{L_2}^{q_2}) \Lambda(w_{1i}, w_{2i} + b_n^* \psi; \gamma, \beta)\, d\psi | < C$ for some $C > 0$. Therefore, for some $D > 0$,

$$\sup_{B,\Gamma} | E[S_{1n}(z_j, z_k; \gamma, \beta)' S_{2n}(z_i, z_j; \gamma, \beta) \mid z_i] - B_n(w_{2i}; \gamma, \beta)$$
$$\times [\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)] a(w_{2i}) | \le D b_n^M = o(n^{-1/2}),$$

where the last equality follows from Assumption 6—or its weaker version, Assumption 3(ii). Using iterated expectations and following parallel steps, we have

$$\sup_{B,\Gamma} | E[S_{1n}(z_j, z_k; \gamma, \beta)' S_{2n}(z_i, z_k; \gamma, \beta) | z_i] + B_n(w_{2i}; \gamma, \beta)$$
$$\times [\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)] a(w_{2i}) | = o(n^{-1/2}).$$

The change in sign with respect to the previous result is a direct consequence of the properties of $K^{(1)}(\cdot)$ stated in Assumption 7. Combining all these results together, we get

$$E[T_{2n}(z_i, z_k, z_k; \gamma, \beta) \mid z_i]$$
$$= 2 B_n(w_{2i}; \gamma, \beta)[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)] a(w_{2i}) + \tilde{c}_n(z_i, z_j; \gamma, \beta),$$

where $\sup_{B,\Gamma} | \tilde{c}_n(z_i, z_j; \gamma, \beta) | = o_p(n^{-1/2})$. Note that

$$E[B_n(w_{2i}; \gamma, \beta)[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)] a(w_{2i})] = 0.$$

Using Assumptions 3(i), 6, 7, and 11, the properties of the trimming set, we have

$$\sup_{B,\Gamma} E\big[\| S_{1n}(z_i, z_j; \gamma, \beta)' S_{2n}(z_i, z_j; \gamma, \beta) \|^2\big] \le \frac{C}{h_n^{2(L+1)} b_n^{2L_2}}$$
$$\times \sup_{\Gamma} E\left[\| F(z_i, z_j)[\tau(w_i, \gamma) - \eta(w_{1i}, w_{2j}, \gamma) - \mu(w_j, \gamma)] \|^2 \left(\frac{a(w_{2i}) a(w_{2j})}{f_{w_2}(w_{2j})} \right)^2 \right].$$

Using Assumptions 11 and 6, the right-hand side is $O(n)$. This result also holds for all the remaining components of $T_{2n}(z_i, z_j, z_k; \gamma, \beta)$. Therefore, we obtain

$$\sup_{B,\Gamma} E\big[\| T_{2n}(z_i, z_j, z_k; \gamma, \beta) \|^2\big] = O(n),$$

and lemma A.3 in Ahn and Powell yields

$$\binom{n}{3}\sum_{i<j<k} T_{2n}(z_i, z_j, z_k; \gamma, \beta) = \frac{3}{n}\sum_{i=1}^{n} E[T_{2n}(z_i, z_k, z_k; \gamma, \beta) \mid z_i] + \tilde{d}_n(z_i, z_j; \gamma, \beta)$$

$$= \frac{6}{n}\sum_{i=1}^{n} B_n(w_{2i}; \gamma, \beta)[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)]a(w_{2i}) + \tilde{d}_n(z_i, z_j; \gamma, \beta)$$

$$+ o_p(n^{-1/2}),$$

where $\sup_{B,\Gamma} |\tilde{d}_n(z_i, z_j; \gamma, \beta)| = o_p(n^{-1/2})$. The result from the Claim follows immediately by noting that $\frac{1}{n}\binom{n}{2}^{-1} = \frac{1}{3}\frac{n-2}{n}\binom{n}{3}^{-1}$. ■

CLAIM 3. *If Assumptions 1–3, 6, 7, and 11–14 are satisfied and* $\{z_i, i = 1, \ldots, n\}$ *is an i.i.d. sample, then*

$$\hat{G}_n(\gamma, \beta) = G_n(\gamma, \beta) + \frac{2(n-2)}{n}\frac{1}{n}\sum_{i=1}^{n} B_n(w_{2i}; \gamma, \beta)$$

$$\times[\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)]a(w_{2i}) + \tilde{e}_n(z_i, z_j; \gamma, \beta),$$

where $\sup_{B,\Gamma} |\tilde{e}_n(z_i, z_j; \gamma, \beta)| = o_p(n^{-1/2})$.

PROOF. The same argument we used to show that $\sup_{B,\Gamma} E[\| T_{2n}(z_i, z_j, z_k; \gamma, \beta)\|^2] = O(n)$ also yields $\sup_{B,\Gamma} E[\| T_{1n}(z_i, z_j; \gamma, \beta)\|^2] = O(n)$. Consequently, $\sup_{B,\Gamma}\binom{n}{2}\sum_{i<j} \| T_{1n}(z_i, z_j; \gamma, \beta)\| = O_p(1)$ and therefore $\sup_{B,\Gamma} \times \frac{1}{n}\binom{n}{2}\sum_{i<j} T_{1n}(z_i, z_j; \gamma, \beta) = o_p(n^{-1/2})$. Given this, the result follows from Claims 1 and 2. ■

PROOF OF THEOREM 5. We will outline the general steps of the proof. We begin by noting that our assumptions are consistent with lemma 25 in Ichimura (2004), which implies that

$$\Pr\big(\mathbb{1}\{f_{w_{2n}}(w_{2i})\} - \mathbb{1}\{f_{w_2}(w_{2i})\} \neq 0 \text{ for at least one } w_{2i}\big) \longrightarrow 0.$$

This result is very useful since our trimming set $\mathcal{W}_{2n}$ is defined in terms of $f_{w_{2n}}(w_{2i})$. Given this, we now present the key components that lead to consistency. As we did previously, define

$$\mathcal{T}_n(\gamma_0, b) = \binom{n}{2}^{-1}\sum_{i<j}\frac{1}{h_n^L} K\left(\frac{\mu_(w_{2i}, \gamma_0) - \mu(w_j, \gamma_0)}{h_n}\right) s(v_i, v_j; b)a_n(w_{2i})a_n(w_{2j}).$$

Using Assumptions 4(i), 6, 7, 18(iii), the properties of $a_n(\cdot)$, and Lebesgue's Dominated Convergence Theorem, we obtain

$$E[\mathcal{T}_n(\gamma_0, b)] \to E[f_{\mu(w_2,\gamma_0)}(\mu(w_i, \gamma_0))\kappa_s(v_i, \mu(w_i, \gamma_0); b)] \equiv T_0(\gamma_0, b),$$

which depends on w_{2i} only through $\mu(w_i, \gamma_0)$. Therefore, consistency will rely exclusively on Assumption 10, namely that $E[s(v_i, v_j; b) \mid \mu(w_j, \gamma_0) = \mu(w_i, \gamma_0)]$ is uniquely minimized at $b = \beta_0$, without the need of Assumption 5.

If Assumptions 1 and 2 are satisfied for any compact subset in the interior of $\mathbb{S}(w_2)$ and if the condition $\log(\bar{w}_n) = o_p(n^{\varepsilon})$ holds, we can extend Theorem 1 to show that

(a) $\sup_{\substack{v \in \mathcal{W}_{2n} \\ \gamma \in \Gamma}} (n^{1-\delta} b_n^{L_2})^{1/2} \|\hat{\mu}_n(v, \gamma) - \mu(v, \gamma)\| = O_p(1)$ for any $\delta > 0$.

(b) $\hat{\mu}_n(v, \gamma) - \mu(v, \gamma) = \frac{1}{f_{w_2}(v)} \frac{1}{n b_n^{L_2}} \sum_{i=1}^{n} [\tau(w_i, \gamma) - \eta(w_{1i}, v, \gamma) - \mu(v, \gamma)] \times H_{b_n}(w_{2i} - v) + \xi_n(\omega, \gamma)$,

where $\sup_{\substack{v \in \mathcal{W}_{2n} \\ \gamma \in \Gamma}} \|\xi_n(\omega, \gamma)\| = O_p(n^{\delta-1} b_n^{-L_2})$ for any $\delta > 0$. Following the same steps as in Sections 3.3.1 and 3.4, we use this result to establish pointwise and uniform convergence of $T_n(\hat{\gamma}, b)$ to $T_0(\gamma_0, b)$ and thus establish consistency of $\hat{\beta}$.

To establish asymptotic normality, we rely once again on the extension of Theorem 1 to the set $\mathcal{W}_{2n}$. The remaining key step is to note that

$$\sup_{B,\Gamma} E\big[\|S_{1n}(z_i, z_j; \gamma, \beta)' S_{2n}(z_i, z_j; \gamma, \beta)\|^2\big] \leq \frac{C}{h_n^{2(L+1)} b_n^{2L_2} \varsigma_n^2}$$

$$\times \sup_{\Gamma} E\left[\|F(z_i, z_j)[\tau(w_i, \gamma) - \eta(w_{1i}, w_{2j}, \gamma) - \mu(w_j, \gamma)]\|^2 \left(\frac{a(w_{2i}) a(w_{2j})}{f_{w_2}(w_{2j})}\right)^2\right].$$

Using Assumptions 3(i), 7, 11, and 18(ii), the right-hand side is $O(n)$. Using the same steps as we did in Claim 3, this result leads to the conclusion that

$$\begin{aligned}\hat{G}_n(\gamma, \beta) &= G_n(\gamma, \beta) + \frac{2(n-2)}{n} \frac{1}{n} \sum_{i=1}^{n} B_n(w_{2i}; \gamma, \beta) \\ &\quad \times [\tau(w_i, \gamma) - \eta(w_i, \gamma) - \mu(w_i, \gamma)] a(w_{2i}) + \tilde{e}_n(z_i, z_j; \gamma, \beta) \\ &\equiv G_n(\gamma, \beta) + \frac{2(n-2)}{n} \frac{1}{n} \sum_{i=1}^{n} \tilde{D}_n(w_i; \gamma, \beta) + \tilde{e}_n(z_i, z_j; \gamma, \beta).\end{aligned}$$

The final result follows from the exact same steps used to prove Theorem 4, and the fact that $a_n(\cdot) \to c$. ■

REFERENCES

AHN, H., AND J. L. POWELL, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics* 58 (1993), 3–29.

AI, C., AND D. L. MCFADDEN, "Estimation of Some Partially Specified Nonlinear Models," *Journal of Econometrics* 76 (1997), 1–37.

ARADILLAS-LOPEZ, A., "Semiparametric Estimation of a Simultaneous Game with Incomplete Information," Mimeo, Princeton University, 2005.

——, "Pairwise Difference Estimation of Incomplete Information Games," Mimeo, Princeton University, 2006.

BLUNDELL, R., AND J. L. POWELL, "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71 (2004), 581–913.

——, AND R. J. SMITH, "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models," *Review of Economic Studies* 56 (1989), 37–58.

DAS, M., W. K. NEWEY, AND F. VELLA, "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies* 70 (2003), 33–58.

ENGLE, R. F., C. W. J. GRANGER, J. RICE, AND A. WEISS, "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association* 81 (1986), 310–20.

HONG, H., AND M. SHUM, "Pairwise Difference Estimator of a Dynamic Optimization Model," Mimeo, Duke University, 2004.

HONORÉ, B. E., "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60 (1992), 533–65.

——, AND J. L. POWELL, "Pairwise Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics* 64 (1994), 241–78.

——, AND ——, "Pairwise Difference Estimation of Nonlinear Models," in D. W. K. Andrews and J. H. Stock, eds., *Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* (Cambridge: Cambridge University Press, 2005), 520–53.

ICHIMURA, H., "Computation of Asymptotic Distribution for Semiparametric GMM Estimators," Mimeo, University College London, 2004.

——, AND L. F. LEE, "Semiparametric Estimation of Multiple Index Models," in W. A. Barnett, J. L. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics. Proceedings of the Fifth International Symposium in Economic Theory and Econometrics* (Cambridge: Cambridge University Press, 1991).

NEWEY, W. K. AND D. L. MCFADDEN, "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. L. McFadden eds., *Handbook of Econometrics, Vol. IV* (Amsterdam: Elsevier, 1994).

———, J. L. POWELL, AND F. VELLA, "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica* 67 (1999), 565–603.

ROBINSON, P., "Root-N Consistent Semiparametric Regression," *Econometrica* 56 (1988), 931–54.

SHERMAN, R., "U-Processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory* 10 (1994), 372–95.

SILVERMAN, B. W., *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall, 1986).