Center for Research on Economic and Social Theory

and

Department of Economics

Working Paper Series

Semiparametric Estimation of Simultaneous Equation Microeconometric Models with Index Restrictions

Lung-fei Lee

April, 1993
Number 93-05

DEPARTMENT OF ECONOMICS
University of Michigan
Ann Arbor, Michigan 48109-1220

Semiparametric Estimation of Simultaneous Equation Microeconometric Models with Index Restrictions

by

Lung-fei Lee

Department of Economics
The University of Michigan
Ann Arbor, MI 48109

April, 1993

Abstract

This article introduces semiparametric methods for the estimation of simultaneous equation microeconometric models with index restrictions. The methods are motivated by some semiparametric minimum distance procedure, which unifies the estimation of both regression-type models and simultaneous equation models in a general framework without emphasis on the construction of instrumental variables. The methods can be applied, for examples, to the estimation of simultaneous equation sample selection models, endogenous switching regression models, Amemiya's simultaneous equation limited dependent variables models, and simultaneous equation disequilibrium markets models. The equations can be nonlinear simultaneous equations. Both single equation and system estimation methods are introduced. Optimal weighting procedures are introduced. The estimators are $\sqrt{n}$-consistent and asymptotically normal. For the estimation of nonparametric regression and some sample selection models where the variances of disturbances are also functions of same indices, the optimal weighted estimator attains Chamberlain's efficient bound for models with conditional moment restrictions. The weighted estimator is also shown to be optimal within a class of semiparametric instrumental variables estimators.

1991 JEL classification numbers: C14, C24, C34.

Keywords:
Semiparametric estimation, microeconometrics, simultaneous equations, nonlinear simultaneous equations, index restrictions, sample selection models, limited dependent variables, disequilibrium market models, semiparametric regression models, conditional moment restrictions, semiparametric estimation, semiparametric instrumental variables, efficiency bound, asymptotic efficiency.

Correspondence Address:
Lung-fei Lee, Department of Economics, Lorch Hall, The University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220.

Semiparametric Estimation of Simultaneous Equation Microeconometric Models with Index Restrictions

by Lung-fei Lee*

1. Introduction

This article concerns about estimation of simultaneous equation microeconometric models, such as simultaneous equation sample selection model, simultaneous equation models with limited dependent and/or qualitative dependent variables with index restrictions. The literature of semiparametric econometrics has mostly considered estimation of regression-type equations but not explicitly simultaneous equation models with limited and qualitative dependent variables with a few excepts [Newey (1985), Powell (1987) and Lee (1991)]. Newey (1985) considered a simultaneous equation model which contains a single limited dependent variable and the structural equation is linear in the explanatory endogenous variables (which are continuous and not limited). Powell (1987) and Lee (1991) considered the estimation of sample selection models by some semiparametric instrumental variables methods. This article considers simultaneous equation microeconometric models which allow nonlinear simultaneous equation structures and suggests general semiparametric methods for the estimation of such models. The proposed estimation method is motivated by minimum distance methods for the estimation of models of quantal responses. It has broad applications for both parametric and semiparametric models. This article will focus on the estimation of semiparametric models.

To motivate the basic idea, consider the classical minimum distance (MD) estimation methods [Berkson (1944), Taylor (1953), Rothenberg (1973), and Chamberlain (1982), among others]. In the context of quantal response, the minimum distance method was developed for models with many observations of responses for each value of the independent variable. In a parametric binary choice model, one specifies a parametric probability function $F(x,\beta)$, i.e., $\text{Prob}(y=1|x)=F(x,\beta)$. For each value of the independent variable, the corresponding frequency of responses provides an unrestricted estimate, $\hat{p}(x_i)$, of the response probability. The minimum distance estimator of β is derived by the minimization:

$$\min_{\beta}\sum_{i=1}^{m}\omega(x_i)(\hat{p}(x_i)-F(x_i,\beta))^2,$$

where m is the number of distinct values of x and $\omega(x_i)$ is some weighting function. For disaggregated data with continuous explanatory variables, one has to use the method of maximum likelihood [see, e.g., Amemiya (1981)]. However, with the development of nonparametric methods, nonparametric regression functions can be consistently estimated without grouped data, even though that its rate of convergence can be slower than the usual rate of convergence for the grouped data case. The main idea in nonparametric regression estimation is local smoothing, in that, at each value of the regressor, its neighboring points are used to construct a 'frequency' estimate. As $E(y|x)=\text{Prob}(y=1|x)$ in the binary response model, $\text{Prob}(y=1|x)$ can be estimated by a nonparametric procedure. For a random sample of size n, suppose that $E_n(y|x)$ is a nonparametric regression estimate of $\text{Prob}(y=1|x)$ at x. Then a generalization of the MD method is

$$\min_{\beta}\sum_{i=1}^{n}\omega(x_i)(E_n(y|x_i)-F(x_i,\beta))^2,$$

where $\omega(x_i)$ is an appropriate weight and n is the sample size.

Although the proposed estimation method has its merit for the estimation of parametric models, the main interest in this article concerns estimation of semiparametric microeconometric models. In Section 2, we will point out its relevance for the estimation of semiparametric models. Simultaneous equation sample selection models, simultaneous equation models with limited dependent variables, and simultaneous equation disequilibrium market models provide some of the interesting examples. Another related model is a semiparametric regression model with a Box-Cox transformation on the dependent variable. An interesting feature of this estimation method is that it unifies the estimation of simultaneous equation models in

* I appreciate having financial support from NSF under grant SBR-9223325 for my research. I would like to thank two anonymous referees, Professor Peter Robinson and Professor C. Kao for their valuable comments and suggestions on an earlier version of my article.

a general framework without emphasis on the construction of appropriate instrumental variables for estimation. Section 3 provides the asymptotic properties of consistency and asymptotic distribution of the semiparametric estimator. Section 4 discusses the possibility of more efficient estimation procedures which take into account variance structure of the semiparametric models. Generalized semiparametric minimum distance estimators are introduced. We compare efficiency gains of system estimation as compared with single equation estimation. Semiparametric efficiency bound for conditional moment restrictions on some semiparametric models has been studied in Chamberlain (1992). The semiparametric models considered in Chamberlain (1992) include the semiparametric regression model of Engle et al. (1986) and Robinson (1988), and a model of sample section of known index. In Section 5, we compare our estimators and their asymptotic variances with the Chamberlain efficiency bounds for those models. When the disturbances of those models are also functions of indices, our generalized semiparametric estimators can achieve such efficiency bounds. In Section 6, we investigate some general semiparametric instrumental variables methods for the estimation of semiparametric simultaneous equation models. It is shown that the generalized semiparametric minimum distance estimator is asymptotically the best semiparametric instrumental variables estimator. Final conclusions are drawn in Section 7. Some of the basic properties of nonparametric estimates of unknown functions and proofs of asymptotic properties of our estimators are collected in the Appendix.

2. Simultaneous Equations Models with Index Restrictions and Semiparametric Minimum Distance Estimation

There are many simultaneous equation models in the microeconometrics literature, which include simultaneous equation sample selection models, endogenous switching regression models, disequilibrium market models, and simultaneous equation models with qualitative and limited dependent variables. In this section, we will introduce the semiparametric minimum distance method for the estimation of semiparametric models with index restrictions. We show how conditional moment equations, which form the estimating functions for our semiparametric estimation method, can be derived from the various microeconometric models.

As a general framkework, we consider estimation of θ in the following system of equations:

$$E[f(z,\beta(\theta))|x] = E[f(z,\beta(\theta))|x\delta(\theta)], \tag{2.1}$$

where x is a finite dimensional (row) vector of exogenous variables, z is a finite dimensional vector consisting of endogenous and/or some exogenous variables in x, f is a (column) vector-valued function with known functional form, $x\delta$ is an m-dimensional (row) vector of indices, and β and δ can be functions of θ (a finite dimensional column vector). The latter specification captures possible parameter constraints on β and δ. The $E(f(z,\beta)|x)$ can be estimated by a nonparametric kernel regression. Suppose that x is a k-dimensional vector of continuous random variables and $K(\cdot)$ is a kernel function on R^k with a bandwidth sequence $\{a_n\}$. Let (z_i,x_i), $i=1,\cdots,n$, be a random sample of size n. Define $A_n(x_i,\theta) = \frac{1}{(n-1)a_n^k}\sum_{j\neq i}^n f(z_j,\beta)K\left(\frac{x_i-x_j}{a_n}\right)$ and $B_n(x_i) = \frac{1}{(n-1)a_n^k}\sum_{j\neq i}^n K\left(\frac{x_i-x_j}{a_n}\right)$. The $E(f(z,\beta)|x)$ at $x=x_i$ can be estimated by

$$E_n(f(z,\beta)|x_i) = A_n(x_i,\theta)/B_n(x_i). \tag{2.2}$$

Similarly, let J be a m-dimensional kernel function with a bandwidth sequence $\{b_n\}$. Define $A_{Jn}(x_i,\theta) = \frac{1}{(n-1)b_n^m}\sum_{j\neq i}^n f(z_j,\beta)J\left(\frac{x_i\delta-x_j\delta}{b_n}\right)$ and $B_{Jn}(x_i) = \frac{1}{(n-1)b_n^m}\sum_{j\neq i}^n J\left(\frac{x_i-x_j}{b_n}\right)$. The $E(f(z,\beta)|x\delta)$ at $x\delta = x_i\delta$ can be estimated by

$$E_n(f(z,\beta)|x_i\delta) = A_{Jn}(x_i,\theta)/B_{Jn}(x_i,\theta). \tag{2.3}$$

To simplify notation, let

$$G_n(x_i,\theta) = E_n(f(z,\beta)|x_i) - E_n[f(z,\beta)|x_i\delta],$$

and $G(x_i,\theta) = E(f(z,\beta)|x_i) - E[f(z,\beta)|x_i\delta]$. A semiparametric minimum distance (SMD) estimation method can be defined as

$$\min_{\theta\in\Theta}\sum_{i=1}^n G_n'(x_i,\theta)W(x_i)G_n(x_i,\theta), \tag{2.4}$$

where Θ is the parameter space of θ, $W(x_i)$ is some weighting matrix at x_i, and n is the sample size. The appropriate weighting matrix shall take into account trimming of the tail distributions of x so as to control for the erratic behavior of denominators in kernel nonparametric regression functions in (2.2) and (2.3) [see Robinson (1988), Powell (1987), Ichimura and Lee (1991)]. We make the general weighting in (2.4) by an arbitrary $W(x)$ to illustrate its role in the asymptotic distribution of the SMD estimator. In a subsequent section, we discuss the optimal weighting for some of the semiparametric models.

The equation (2.1) can be derived from a specific model. The following semiparametric simultaneous equation models are interesting examples.

Example 1: *Simultaneous equation sample selection models*

Consider a system of latent equations $g(y^*,x,\beta) = u$, where the value y of the vector y^* can be observed only if the choice criterion $x\delta > \epsilon$ is satisfied. The disturbances u and ϵ are correlated. The x is a vector of exogenous variables. The g system can be a simultaneous linear or nonlinear equations system. An empiricial example of this model is Heckman (1974). With the observed sample y, it implies an equations system with unknown functions:

$$E[g(y,x,\beta)|I=1,x] = E(u|I=1,x), \tag{2.5}$$

where I is a dichotomous indicator with $I=1$ indicating that $x\delta > \epsilon$. It implies also that

$$E[g(y,x,\beta)|I=1,x\delta] = E(u|I=1,x\delta). \tag{2.6}$$

Under the index assumption that the joint distribution of u and ϵ conditional on x may depend on x only through $x\delta$,

$$E(u|I=1,x) = E(u|x\delta>\epsilon,x) = E(u|I=1,x\delta). \tag{2.7}$$

The unknown function $E(u|I=1,x)$ in (2.5) and $E(u|I=1,x\delta)$ in (2.6) can be eliminated by taking the difference of (2.5) and (2.6) using the identity (2.7):

$$E[g(y,x,\beta)|I=1,x] = E[g(y,x,\beta)|I=1,x\delta]. \tag{2.8}$$

These moment equations in (2.8) are valid for both the truncated sample selection model and the censored sample selection model. For the truncated sample selection model where only sample observations on y and x and the event $I=1$ are available, (2.8) can be used for the SMD estimation. For the censored sample selection model, the events of either $I=1$ or $I=0$ are also observed. For this case, additional information is available in that

$$E(I|x) = E(I|x\delta), \tag{2.9}$$

The censored sample selection can be estimated with (2.8) and (2.9). Alternatively, (2.8) and (2.9) imply that

$$E[Ig(y,x,\beta)|x] = E[Ig(y,x,\beta)|x\delta]. \tag{2.10}$$

The (2.9) and (2.10) can be used together for the SMD estimation.

For an endogenous switching sample selection model, there are observed outcome equations for each alternative. Consider, for example, a model with two choices. Let $I=1$ indicate the first alternative and $I=0$ for the second alternative. There are outcome equations for each possible alternative: $g_1(y_1,x,\beta_1) = u_1$ for alternative 1, and $g_2(y_2,x,\beta_2) = u_2$ for alternative 2. This model implies the following conditional moment equations:

$$E[g_1(y_1,x,\beta_1)|I=1,x] = E[g_1(y_1,x,\beta_1)|I=1,x\delta], \tag{2.11}$$

$$E[g_2(y_2,x,\beta_2)|I=0,x] = E[g_2(y_2,x,\beta_2)|I=0,x\delta], \tag{2.12}$$

and (2.9). These equations can be used together for the SMD estimation. For models with polytomous choices, the estimation can be extended to incorporate equations for all alternatives. For the polytomous case, the $x\delta$ will be a vector of indices.

Example 2: *Multi-market Disequilibrium Models*

Multi-market disequilibrium models have been specified in Ito (1980) and Gourieroux et al. (1980). Ito's two-market disequilibrium system is specified as

$$y^d = \tilde{y}^d + \alpha_1(l - \tilde{l}^s)$$
$$y^s = \tilde{y}^s + \alpha_2(l - \tilde{l}^d)$$
$$l^d = \tilde{l}^d + \beta_1(y - \tilde{y}^s)$$
$$l^s = \tilde{l}^s + \beta_2(y - \tilde{y}^d),$$

where y^d, l^d, y^s and l^s are effective demand and supply, $\tilde{y}^d$, $\tilde{l}^d$, $\tilde{y}^s$ and $\tilde{l}^s$ are notional demand and supply, and the observed dependent variables are $y = \min\{y^s, y^d\}$ and $l = \min\{l^s, l^d\}$. The notional demand and suply equations are $\tilde{y}^d = x\alpha_d + u_1$, $\tilde{y}^s = x\alpha_s + u_2$, $\tilde{l}^d = x\beta_d + v_1$, and $\tilde{l}^s = x\beta_s + v_2$. For this two-market model, there are four regimes of excess demand and/or excess supply. In most applications, regime classification information may not be available. For analysis, it is convenient to define latent regime indicators I_1 and I_2 in that $I_1 = 1$ if and only if $y^d < y^s$, otherwise, $I_1 = 0$; and $I_2 = 1$ if and only if $l^s < l^d$, otherwise, $I_2 = 0$. The implied equations for each regime have been derived in Ito (1980). Let $y^* = \tilde{y}^d - \tilde{y}^s$ and $l^* = \tilde{l}^s - \tilde{l}^d$.

Regime ($I_1 = 1, I_2 = i$): This regime occurs if and only if $y^* < \alpha_2 l^*$ and $l^* < \beta_1 y^*$. The equations for the observed dependent variables are $y = \tilde{y}^d$ and $l = \tilde{l}^s$.

Regime ($I_1 = 0, I_2 = 0$): This regime occurs if and only if $y^* > \alpha_1 l^*$ and $l^* > \beta_2 y^*$. The observed equations are $y = \tilde{y}^s$ and $l = \tilde{l}^d$.

Regime ($I_1 = 1, I_2 = 0$): This regime occurs if and only if $y^* < \alpha_1 l^*$ and $l^* > \beta_1 y^*$. The observed equations are $y = \tilde{y}^s + (y^* - \alpha_1 l^*)/(1 - \alpha_1\beta_1)$ and $l = \tilde{l}^s + (\beta_1 y^* - l^*)/(1 - \alpha_1\beta_1)$.

Regime ($I_1 = 0, I_2 = 1$): This regime occurs if and only if $y^* > \alpha_2 l^*$ and $l^* < \beta_2 y^*$. The observed equations are $y = \tilde{y}^d + (\alpha_2 l^* - y^*)/(1 - \alpha_2\beta_2)$ and $l = \tilde{l}^d + (l^* - \beta_2 y^*)$.

From these regime characterizations, the regime probabilities are functions of two indices $x(\alpha_d - \alpha_s)$ and $x(\beta_d - \beta_s)$ because the regime inequalities involve only the variables y^* and l^* and some unknown parameters.[1] Combining regimes, since $I_1 = I_1 I_2 + I_1(1 - I_2)$,

$$\begin{aligned} y &= I_1 y^d + (1 - I_1) y^s \\ &= \tilde{y}^s + I_1(\tilde{y}^d - \tilde{y}^s) + \alpha_1 I_1(1 - I_2)(l - \tilde{l}^s) + \alpha_2(1 - I_1) I_2(l - \tilde{l}^d) \\ &= \tilde{y}^s + I_1 y^* + \frac{\alpha_1}{1 - \alpha_1\beta_1} I_1(1 - I_2)(\beta_1 y^* - l^*) + \frac{\alpha_2}{1 - \alpha_2\beta_2}(1 - I_1) I_2(l^* - \beta_2 y^*), \end{aligned}$$

because $I_1 I_2(l - \tilde{l}^s) = 0$ and $(1 - I_1)(1 - I_2)(l - \tilde{l}^d) = 0$. It follows that

$$E(y|x) = x\alpha_s + E(I_1 y^*|x) + \frac{\alpha_1}{1 - \alpha_1\beta_1} E[I_1(1 - I_2)(\beta_1 y^* - l^*)|x] + \frac{\alpha_2}{1 - \alpha_2\beta_2} E[(1 - I_1) I_2(l^* - \beta_2 y^*)|x].$$

Let $x\delta = (x(\alpha_d - \alpha_s), x(\beta_d - \beta_s))$. By index restrictions, $E(I_1 y^*|x) = E(I_1 y^*|x\delta)$, $E[I_1(1 - I_2)(\beta_1 y^* - l^*)|x] = E[I_1(1 - I_2)(\beta_1 y^* - l^*)|x\delta]$ and $E[(1 - I_1) I_2(l^* - \beta_2 y^*)|x] = E(1 - I_1) I_2(l^* - \beta_2 y^*)|x\delta]$. By eliminating these unknown functions,

$$E(y - x\alpha_s|x) = E(y - x\alpha_s|x(\alpha_d - \alpha_s), x(\beta_d - \beta_s)). \quad (2.13)$$

Similarly,

$$E(l - x\beta_d|x) = E(l - x\beta_d|x(\alpha_d - \alpha_s), x(\beta_d - \beta_s)). \quad (2.14)$$

The (2.13) and (2.14) can be used for the SMD estimation.[2] The α_1, α_2, β_1 and β_2 are not directly identifiable from (2.13) and (2.14). However, if this disequilibrium model has been derived from behavioral models (fixed price equilibrium models), all the parameters will be functions of basic structural parameters of utility

[1] The listing of regimes above demonstrates that the relevant indices in estimation are the two indices $x(\alpha_d - \alpha_s)$ and $x(\beta_d - \beta_s)$ instead of the indices $x\alpha_d$, $x\alpha_s$, $x\beta_d$, and $x\beta_s$.

[2] There are symmetry properties in (2.13) and (2.14). The (2.13) implies $E(y - x\alpha_d|x) = E(y - x\alpha_d|x(\alpha_d - \alpha_s), x(\beta_d - \beta_s))$ and vice versa. The (2.14) is equivalent to $E(l - x\beta_s|x) = E(l - x\beta_s|x(\alpha_d - \alpha_s), x(\beta_d - \beta_s))$.

and production functions. All the structural parameters are captured in the notional demand and supply equations [see, e.g., Lee (1986)]. For such models, these spill-over effects will be estimable.

Example 3: *Tobit Simultaneous Equation Model*

A Tobit simultaneous equation model has been introduced in Amemiya (1974):

$$y_l^* = \sum_{j \neq l}^{m} \gamma_{lj} y_j + x\beta_l + u_l, \qquad l = 1, \cdots, m,$$

where the observed dependent variables are $y_l = \max\{y_l^*, 0\}$, $l = 1, \cdots, m$. For this model, there are several different reduced form equations. For the model to be a well defined probability model (coherency), the regime conditions have to be mutually exclusive and exhaustive so that the regime probabilities sum to one. Let Γ be the $m \times m$ matrix which has ones on the diagonal and $-\gamma_{ij}$ in the (i, j)th place. The model is coherent if and only if every principal minor of Γ is positive [Amemiya (1974)]. The joint event (regime) probabilities are functions of $x\beta_1, \cdots, x\beta_m$, which are the indices for this model. For exposition and notational simplicity, consider a two-equation model:

$$\begin{aligned} y_1^* &= y_2\gamma_1 + x\beta_1 + u_1, \\ y_2^* &= y_1\gamma_2 + x\beta_2 + u_2. \end{aligned} \quad (2.15)$$

For model identification, it is assumed that there is at least one exogenous variable in each structural equation not included in the other [Ameimya (1974), Assumption 3.3]. There are four regimes; namely, (*i*). $y_1 > 0, y_2 > 0$, (*ii*). $y_1 > 0, y_2 = 0$, (*iii*). $y_1 = 0, y_2 > 0$, and (*iv*). $y_1 = 0, y_2 = 0$. The model coherency condition is that $1 - \gamma_1\gamma_2 > 0$. Define four mutually exclusive and exhaustive dichotomous indicators I_l, $l = 1, \cdots, 4$, for the four regimes. Let $y = (y_1, y_2)$, $y^* = (y_1^*, y_2^*)$, $u = (u_1, u_2)$, and $B = (\beta_1, \beta_2)$.

For the regime (*i*), let $\Gamma_1 = \begin{pmatrix} 1 & -\gamma_2 \\ -\gamma_1 & 1 \end{pmatrix}$. The reduced form equations are $y^* = xB\Gamma_1^{-1} + u\Gamma_1$. The $y\Gamma_1 = xB + u$ is observed if $xB\Gamma_1^{-1} + u\Gamma_1 \geq 0$. The indices are xB. Since $E(u|I_1 = 1, x) = E(u|I_1 = 1, xB)$, it implies that $E(y\Gamma_1 - xB|I_1 = 1, x) = E(y\Gamma_1 - xB|I_1 = 1, xB)$. As the xB on both sides cancels each other, the conditional moment equations are simply

$$E(y\Gamma_1|I_1 = 1, x) = E(y\Gamma_1|I_1 = 1, xB). \quad (2.16)$$

For the regime (*ii*), let $\Gamma_2 = \begin{pmatrix} 1 & -\gamma_2 \\ 0 & 1 \end{pmatrix}$. The reduced form equations are $y^* = xB\Gamma_2^{-1} + u\Gamma_2^{-1}$. The $y_1 = x\beta_1 + u_1$ is observed if $xB\Gamma_2^{-1} + u\Gamma_2^{-1} \geq 0$. The conditional moment equation is

$$E(y_1|I_2 = 1, x) = E(y_1|I_2 = 1, xB). \quad (2.17)$$

For the regime (*iii*), let $\Gamma_3 = \begin{pmatrix} 1 & 0 \\ -\gamma_1 & 1 \end{pmatrix}$. The conditional moment equation is

$$E(y_2|I_3 = 1, x) = E(y_2|I_3 = 1, xB). \quad (2.18)$$

For the regime (*vi*), the reduced form equations are $y^* = xB + u$. Except for the regime indicator, no observations on the continuous dependent variables are available.

In addition to the above conditional moment equations, moment equations for the regime probabilities are:

$$E(I_l|x) = E(I_l|xB), \qquad l = 1, 2, 3. \quad (2.19)$$

These conditional moment and probabilities equations can be used jointly for estimation.

Alternatively, the regimes (*i*) and (*ii*) can be combined such that $y_1 = y_2\gamma_1 + (I_1 + I_2)x\beta_1 + (I_1 + I_2)u$. Since $E((I_1 + I_2)u|x) = E((I_1 + I_2)u|xB)$ and $E(I_1 + I_2|x) = E(I_1 + I_2|xB)$, it follows that

$$E(y_1 - y_2\gamma_1|x, I_1 + I_2 = 1) = E(y_1 - y_2\gamma_1|xB, I_1 + I_2 = 1), \quad (2.20)$$

and

$$E(I_1 + I_2|x) = E(I_1 + I_2|xB). \tag{2.21}$$

As $I_1 + I_2 = 1$ if and only if $y_1 > 0$, the above moment equations provide information for the estimation of the structural equation of y_1. Similarly, the regimes (*i*) and (*iii*) can be combined for the estimation of the structural equation of y_2. The corresponding moment equations are

$$E(y_2 - y_1\gamma_2|x, I_1 + I_3 = 1) = E(y_2 - y_1\gamma_2|xB, I_1 + I_3 = 1), \tag{2.22}$$

and

$$E(I_1 + I_3|x) = E(I_1 + I_3|xB). \tag{2.23}$$

These moment equations can be used for estimation.

The above examples provide some important models that can be estimated by the SMD method. In a subsequent section, we will provide more detailed analysis of the simultaneous equation sample selection models. Many more microeconometric models may be estimated by this method. There is some similarity of this method with the semiparametric nonlinear least squares method introduced originally in Ichimura (1987) and extended in Ichimura and Lee (1991). The model considered in Ichimura and Lee (1991) is an index model motivated by a single equation (truncated) sample selection model:

$$y = x\beta + \tau(x\delta) + v, \tag{2.24}$$

where $\tau(\cdot)$ is an unknown function of an index $x\delta$. The semiparametric nonlinear least squares (SNLS) method introduced in Ichimura and Lee (1991) is

$$\min_{\theta} \sum_{i=1}^{n} \left(y_i - x_i\beta(\theta) - E_n[y - x\beta(\theta)|x_i\delta(\theta)]\right)^2,$$

where $E_n(y - x\beta|x\delta)$ provides a nonparametric estimate of $\tau(x\delta)$. This approach is motivated by replacing the unkown function τ by a nonparametric function in an intermediate step and the final estimation is done as an M estimation method. Similar ideas are used in Robinson (1988), among others. The SMD approach takes an alternative view. Conditional on y being observed, the sample selection equation implies that $E(y|x) = x\beta + \tau(x\delta)$ and $E(y|x\delta) = E(x|x\delta)\beta + \tau(x\delta)$. The unknown τ is then eliminated by taking the difference to have the estimating moment equation $E(y - x\beta|x) = E(y - x\beta|x\delta)$. However, as it can be shown in subsequent sections, these two approaches provide asymptotically equivalent estimates for (2.24). For the estimation of (2.24), the SNLS may be preferred as it is relatively computationally simpler. The SMD approach involves the computation of $E_n(y|x)$. On the other hand, the SMD is applicable to the estimation of simultaneous equation models and models with implicit function where the dependent variable y can not be separated out, while the SNLS is not applicable to such models. An illustrative example is the sample selection model (2.24) with a Box-Cox transformation on y: $(y^\lambda - 1)/\lambda = x\beta + \tau(x\delta) + v$. The SNLS method $\min_{\theta,\lambda} \sum_{i=1}^n \{(y_i^\lambda - 1)/\lambda - E_n[((y^\lambda - 1)/\lambda)|x_i\delta(\theta)] - [x_i - E_n(x|x_i\delta)]\beta(\theta)\}^2$ would not be consistent [see Amemiya (1985) for the regression case without sample selection]. The relevant SMD estimation is $\min_{\theta,\lambda} \sum_{i=1}^n \left(E_n[(y^\lambda - 1)/\lambda - x\beta(\theta)|x_i] - E_n[(y^\lambda - 1)/\lambda) - x\beta(\theta)|x_i\delta(\theta)]\right)^2$, which can be shown to be consistent. As a remark, we note that Powell (1991) has considered estimation of limited dependent variable model with the Box-Cox transformation under quantile restrictions but the semiparametric estimation of sample selection models with the Box-Cox transformation has not been considered in the literature.

For the estimation of a single linear simultaneous equation sample selection model:

$$y = z\beta + \tau(x\delta) + v, \tag{2.25}$$

where z contains endogenous variables. Given a consistent estimate of δ, Powell (1987) has suggested a generalization of Robinson's semiparametric least squares [Robinson (1988)] to a two-stage semiparametric instrumental variable approach for the estimation of β in (2.16). In a subsequent article, Lee (1991) has considered in some detail the identification problem for a linear simultaneous equation sample selection system

and derived an asymptotically efficient two-step semiparametric instrumental variable method (conditional on a given consistent estimate of δ) for the estimation of (2.25). The SMD method in this article differs from these articles in that one does not need to care about the construction of appropriate instrumental variables for estimation. It can be easily extended to the estimation of nonlinear simultaneous equation semiparametric models. It can be designed for the estimation of a single equation and/or a system of equations.

3. Semiparametric Minimum-distance Estimation

In this section, we discuss asymptotic properties of the SMD method. Detailed proofs of the results can be found in the Appendix A3. The underlying regularity conditions and assumptions are summarized in the Appendix A2.

Let $\hat{\theta}_I$ denote the SMD estimate from (2.4). Let θ_o be the true parameter vector, and let β_o and δ_o denote, respectively, β and δ evaluated at $\theta = \theta_o$. For any possible value θ of θ_o in Θ, $E(.|x\delta)$ denotes $E(.|x\delta(\theta))$, a conditional expectation conditional on $x\delta(\theta)$ for a given value θ, for simplicity. Furthermore, $E(.|x_i)$ stands for $E(.|x)$ evaluated at $x = x_i$, and $E(.|x_i\delta)$ for $E(.|x\delta)$ evaluated at $x\delta = x_i\delta$. All the expectation operations are taken with respect to the true data generating process at θ_o. The following propositions show that, under proper identification conditions and regularity conditions, $\hat{\theta}_I$ is $\sqrt{n}$-consistent. To simplify notation, denote

$$G_{\theta'}(x, \theta_o) = \frac{\partial E(f(z, \beta_o)|x)}{\partial \theta'} - \frac{\partial E(f(z, \beta_o)|x\delta_o)}{\partial \theta'},$$

and

$$u = f(z, \beta_o) - E(f(z, \beta_o)|x)$$

throughout this article. u is the disturbance of the model. As a convention, for any function, say, $f(z, \theta)$ of θ, $f'_\theta(z, \theta)$ with a subscript θ will denote its gradient $\frac{\partial f'(z,\theta)}{\partial \theta}$ with respect to θ.

Proposition 3.1 *Under Assumptions 1-6 and the identification condition that, for any $\theta \neq \theta_o$, $G(x, \theta) \neq 0$ with positive probability on X where $X = \{x|W(x) \neq 0\}$, $\hat{\theta}_I$ is a consistent estimator of θ_o and*

$$\sqrt{n}(\hat{\theta}_I - \theta_o) = -\left\{E\left(G'_\theta(x, \theta_o)W(x)G_{\theta'}(x, \theta_o)\right)\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [C_1(x_i) - C_2(x_i\delta_o)]u_i + o_p(1), \tag{3.1}$$

where $C_1(x_i) = G'_\theta(x_i, \theta_o)W(x_i)$, and $C_2(x_i\delta_o) = E\{G'_\theta(x, \theta_o)W(x)|x_i\delta_o\}$. Consequently, $\sqrt{n}(\hat{\theta}_I - \theta_o) \xrightarrow{D} N(0, \Omega_I)$, where

$$\Omega_I = \left\{E\left[G'_\theta(x, \theta_o)W(x)G_{\theta'}(x, \theta_o)\right]\right\}^{-1} \Sigma_I \left\{E\left[G'_\theta(x, \theta_o)W(x)G_{\theta'}(x, \theta_o)\right]\right\}^{-1} \tag{3.2}$$

and $\Sigma_I = E\left[\left(C_1(x) - C_2(x\delta_o)\right)\mathrm{Var}(u|x)\left(C_1(x) - C_2(x\delta_o)\right)'\right]$.

We note that [Ichimura and Lee (1991), Lemma 4]

$$\begin{aligned} G_{\theta'}(x, \theta_o) &= E(f_{\theta'}(z, \beta_o)|x) - E(f_{\theta'}(z, \beta_o)|x\delta_o) \\ &\quad - \nabla' E(f(z, \beta_o)|x\delta_o)\left(\frac{\partial \delta'(\theta_o)x'}{\partial \theta'} - E\left(\frac{\partial \delta'(\theta_o)x'}{\partial \theta'}x' \Big| x\delta_o\right)\right), \end{aligned} \tag{3.3}$$

where $\nabla E(.|x\delta_o)$ denotes the gradient of $E(.|x\delta_o)$ with respect to its argument vector $x\delta_o$.

It is interesting to point out an implication of Proposition 3.1 for the special model with $f(z, \beta) = y - f_1(x, \beta) - f_2(z, \beta)$, where f_1 depends only on x. For this case, $G(x, \theta_o) = E(y|x) - f_1(x, \beta_o) - E(f_2(z, \beta_o)|x) - E(f(z, \beta_o)|x\delta_o)$, and the model is equivalent to

$$y = f_1(x, \beta_o) + f_2(z, \beta_o) + E(f(z, \beta_o)|x\delta_o) + u. \tag{3.4}$$

An alternative estimation approach for this model can be

$$\min_{\theta\in\Theta}\sum_{i=1}^{n}[y_i - f_1(x_i,\beta) - E_n(f_2(z,\beta)|x_i) - E_n(f(z,\beta)|x_i\delta)]' W(x_i) \\ \times [y_i - f_1(x_i,\beta) - E_n(f_2(z,\beta)|x_i) - E_n(f(z,\beta)|x_i\delta)] . \qquad (3.5)$$

For the model where the term $f_2(z,\beta_o)$ vanishes, the estimation method (3.5) will be the SNLS method originally introduced in Ichimura (1987). By arguments similar to the proof of Proposition 3.1, the SNLS estimator $\hat{\theta}_{NL}$ of θ from (3.5) is consistent and asymptotically normal. For any possible value θ, denote

$$r(x,\theta) = f_1(x,\beta) + E[f_2(z,\beta)|x] + E[f(z,\beta)|x\delta],$$

and $r_{\theta'}(x,\theta) = \frac{\partial r(x,\theta)}{\partial \theta'}$. The $\hat{\theta}_{NL}$ has the same asymptotic distribution as the SMD estimator $\hat{\theta}_I$ for this model, which can be seen from the following proposition.

Proposition 3.2 *Under Assumptions 1-6 and the identification condition that, for any $\theta \neq \theta_o$, $r(x,\theta) \neq r(x,\theta_o)$ with positive probability on X, the SNLS estimator $\hat{\theta}_{NL}$ from (3.5) is consistent, and*

$$\sqrt{n}(\hat{\theta}_{NL}-\theta_o) = \{E[r'_\theta(x,\theta_o)W(x)r_{\theta'}(x,\theta_o)]\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{r'_\theta(x_i,\theta_o)W(x_i) - E([r'_\theta(x,\theta_o)W(x)|x_i\delta_o)\}u_i + o_p(1). \qquad (3.6)$$

The limiting distribution of $\hat{\theta}_{NL}$ for (3.3) is the same as the limiting distribution of $\hat{\theta}_I$ in Proposition 3.1 because $E(f'_{1\theta}(x,\beta_o)|x) = f'_{1\theta}(x,\beta_o)$.

4. Generalized Semiparametric Estimation

We see from Proposition 3.1 that the asymptotic distribution of SMD estimator depends on the weighting function $W(x_i)$. If $W(x)$ is selected as functions of $x\delta_o$, the asymptotic distribution for the corresponding SMD estimator has simpler expression because $C_2(x\delta_o)$ will be zero as $E\{G'_\theta(x,\theta_o)W(x\delta_o)|x_i\delta_o\} = E\{G'_\theta(x,\theta_o)|x_i\delta_o\}W(x_i\delta_o) = 0$, by (3.3). In terms of the issue of selecting the optimal weighting function, one may suspect that the appropriate weight function shall be the inverse of the variance function of u. Unfortunately, if the conditional variance function of u (conditional on x) is a general function of x (not a function of $x\delta_o$), such a weighting function does not necessarily improve efficiency because the presence of the additional term C_2, which captures the effect of using nonparametric functions to replace the correct regression functions as an intermediate step in estimation. However, in many microeconometric models such as the simultaneous equation models introduced before, the conditional variances may depend on x only through their dependence on $x\delta_o$. With such heteroskedastic variance models, the optimal weighting function is indeed the inverse of the variance function because the complicated $C_2(x\delta_o)$ can be zero. In general, if the (limiting) weighting function depends only on $x\delta$, $C_2(x\delta_o)$ will be zero, that can be regarded as an orthogonality property of an estimation procedure. Thus for models where heteroskedastic variances are functions of $x\delta_o$, more efficient estimator may be derived by some generalized semiparametric procedures.

Let $\hat{\theta}$ be a $\sqrt{n}$-consistent estimator of θ, for example, $\hat{\theta}_I$ from the previous SMD estimation. Define a nonparametric variance function:

$$V_n(x_i\hat{\delta}) = E_n[f(z,\hat{\beta})f'(z,\hat{\beta})|x_i\hat{\delta}] - E_n[f(z,\hat{\beta})|x_i\hat{\delta}]E_n[f'(z,\hat{\beta})|x_i\hat{\delta}], \qquad (4.1)$$

where $\hat{\delta}$ and $\hat{\beta}$ are $\delta(\theta)$ and $\beta(\theta)$ evaluated at $\hat{\theta}$. This nonparametric variance function is an estimate of the variance function $V(x\delta_o)$ where

$$V(x\delta_o) = \text{Var}[f(z,\beta_o)|x\delta_o] \qquad (4.2)$$

is the variance function $f(z,\beta_o)$ conditonal on $x\delta_o$. This is so no matter whether $\text{Var}(f(z,\beta_o)|x)$ equals $\text{Var}(f(z,\beta_o)|x\delta_o)$ or not. The feasible weighting matrix $W_n(x_i\hat{\delta})$ will be $W_n(x_i\hat{\delta}) = V_n^{-1}(x_i\hat{\delta})$. Denote

$$G_{n,\theta'}(x_i,\theta) = \frac{\partial E_n(f(z,\beta(\theta))|x_i)}{\partial\theta'} - \frac{\partial E_n[f(z,\beta(\theta))|x_i\delta(\theta)]}{\partial\theta'}.$$

Our suggested generalized semiparametric minimum distance (GSMD) estimator is a two-step estimator based on a generalized semiparametric nonlinear least squares procedure:

$$\hat{\theta}_G = -\Big\{\sum_{i=1}^{n}t_n(x_i,\hat{\theta})G'_{n,\theta}(x_i,\hat{\theta})W_n(x_i\hat{\delta})G_{n,\theta'}(x_i,\hat{\theta})\Big\}^{-1} \\ \times\sum_{i=1}^{n}t_n(x_i,\hat{\theta})G'_{n,\theta}(x_i,\hat{\theta})W_n(x_i\hat{\delta})\left(G_n(x_i,\hat{\theta}) - G_{n,\theta'}(x_i,\hat{\theta})\hat{\theta}\right), \qquad (4.3)$$

where $t_n(x,\hat{\theta})$ is a trimming function which goes to 1 as n goes to infinity. The function of $t_n(x,\hat{\theta})$ is to trim the erratic behavior of nonparametric regression and variance functions at their tails. The trimming becomes less severe as n goes to infinity. The detailed description of this trimming is in the Appendix A4. This two-step GSMD estimator is motivated by the GSMD method :

$$\min_\theta\sum_{i=1}^{n}t_n(x_i,\hat{\theta})G'_n(x_i,\theta)W_n(x_i\hat{\delta})G_n(x_i,\theta).$$

However, for parameter value of θ not in the neighborhood of $\hat{\theta}$, the trimming by $t_n(x_i,\hat{\theta})$ does not control properly the tails of the nonparametric regressions at θ. The two-step estimator overcomes this difficulty. If $t_n(x_i,\hat{\theta})$ were replaced by $t_n(x_i,\theta)$ in the minimization method, the global minimum would be be zero at values of θ such that all the observations are trimmed out. One possible remedy is to introduce penalty functions for trimmed observations. However, the two-step method seems simpler.

Proposition 4.1 *Under Assumptions 1-5 and 7 and the identification condition that, for any $\theta \neq \theta_o$, $E[f(z,\beta(\theta))|x] - E[f(z,\beta(\theta))|x\delta(\theta)] \neq 0$ with positive probability, $\hat{\theta}_G$ is a consistent, asymptotically normal estimator with*

$$\sqrt{n}(\hat{\theta}_G - \theta_o) \longrightarrow N(0,\Omega_G), \qquad (4.4)$$

where $\Omega_G = D^{-1}(\theta_o)\Sigma_G D^{-1}(\theta_o)$ *with* $D(\theta_o) = E\left[G'_\theta(x,\theta_o)V^{-1}(x\delta_o)G_{\theta'}(x,\theta_o)\right]$, *and*

$$\Sigma_G = E\left[G'_\theta(x,\theta_o)V^{-1}(x\delta_o)\text{Var}(u|x)V^{-1}(x\delta_o)G_{\theta'}(x,\theta_o)\right].$$

Furthermore, if $\text{Var}(u|x) = \text{Var}(u|x\delta_o)$, *then* $\Omega_G = D^{-1}(\theta_o)$.

One can compare the limiting variance matrix Ω_G with the variance matrix Ω_I in Proposition 3.1 for the model where $\text{Var}(u|x) = \text{Var}(u|x\delta_o)$. For such a model

$$\Sigma_I = E([C_1(x) - C_2(x\delta_o)]V(x\delta_o)[C_1(x) - C_2(x\delta_o)]').$$

Since $E(G'_\theta(x,\theta_o)|x\delta_o) = 0$ by (3.3),

$$E(G'_\theta(x,\theta_o)W(x)G_{\theta'}(x,\theta_o)) = E(G'_\theta(x,\theta_o)C'_1(x)) = E(G'_\theta(x,\theta_o)[C_1(x) - C_2(x\delta_o)]').$$

Hence Ω_I can be written as

$$\Omega_I = \{E(G'_\theta(x,\theta_o)[C_1(x) - C_2(x\delta_o)]')\}^{-1}E([C_1(x) - C_2(x\delta_o)]V(x\delta_o)[C_1(x) - C_2(x\delta_o)]') \\ \times\{E([C_1(x) - C_2(x\delta_o)]G_{\theta'}(x,\theta_o))\}^{-1}. \qquad (4.5)$$

Comparing this Ω_I with Ω_G by the generalized Schwartz inequality, we see that $\Omega_I \geq \Omega_G$. Hence $\hat{\theta}_G$ is asymptotically efficient relative to $\hat{\theta}_I$ as it uses the optimal weighting.

When f is a system of equations, the above estimation method is a system estimation method. For the estimation of the classical linear (nonlinear) simultaneous equation model, it is well-known that the three stage least squares (nonlinear three stage least squares) estimator is effcient relative to the two stage least squares (nonlinear two stage least squares) estimator. As a system estimation method, one may expect that a system GSMD estimator will be asymptotically efficient relative to estimators derived from single equation methods. For the case that $\mathrm{Var}(f(z,\theta_o)|x) = \mathrm{Var}(f(x,\theta_o)|x\delta_o)$, there is indeed possible efficiency gain in the system GSMD method. To see that, let f consist of two components such that $f'(z,\beta) = (f_1'(z,\beta), f_2'(z,\beta))$. Suppose that θ_1 is the parameter vector in the first set of moment equations:

$$E[f_1(z,\beta(\theta_{o,1}))|x] = E[f_1(z,\beta(\theta_{o,1}))|x\delta(\theta_{o,1})], \tag{4.6}$$

and θ_2 is the parameter vector in the remaining moment equations

$$E[f_2(z,\beta(\theta_{o,2}))|x] = E[f_2(z,\beta(\theta_{o,2}))|x\delta(\theta_{o,2})]. \tag{4.7}$$

Consider first the case that θ_1 and θ_2 are distinct parameters. The θ_1 can be estimated by using only the moment equations (4.6) of f_1. Let $\hat{\theta}_{1,s}$ be the GSMD estimator using f_1 alone. Let $\hat{\theta}_{1,G}$ be the corresponding GSMD estimator using both f_1 and f_2 in (4.6) and (4.7) as a system. Let $V(x\delta_o) = \begin{pmatrix} V_{11}(x\delta_o) & V_{12}(x\delta_o) \\ V_{21}(x\delta_o) & V_{22}(x\delta_o) \end{pmatrix}$ be the corresponding partitioned variance matrix, and let $V^{-1}(x\delta_o) = \begin{pmatrix} V^{11}(x\delta_o) & V^{12}(x\delta_o) \\ V^{21}(x\delta_o) & V^{22}(x\delta_o) \end{pmatrix}$ be its inverse matrix. Let $A_1(x) = E(f'_{1\theta_1}(z,\beta_o)|x) - \frac{\partial E(f_1'(z,\beta_o)|x\delta_o)}{\partial\theta_1}$, and $A_2(x) = E(f'_{2\theta_2}(z,\beta_o)|x) - \frac{\partial E(f_2'(z,\beta_o)|x\delta_o)}{\partial\theta_2}$. From Proposition 4.1, the limiting variance matrix of $\sqrt{n}(\hat{\theta}_{1,s} - \theta_{o,1})$ is $\Omega_{1,s}$ where

$$\Omega_{1,s}^{-1} = E[A_1(x)V_{11}^{-1}(x\delta_o)A_1'(x)],$$

and the limiting variance matrix of $\sqrt{n}(\hat{\theta}_{1,G} - \theta_{o,1})$ is $\Omega_{1,G}$, where

$$\begin{aligned}\Omega_{1,G}^{-1} = {} & E[A_1(x)V^{11}(x\delta_o)A_1'(x)] \\ & - E[A_1(x)V^{12}(x\delta_o)A_2'(x)]\{E[A_2(x)V^{22}(x\delta_o)A_2'(x)]\}^{-1}E[A_2(x)V^{21}(x\delta_o)A_1'(x)].\end{aligned}$$

Since $V^{11}(x\delta_o) - V_{11}^{-1}(x\delta_o) = V^{12}(x\delta_o)\{V^{22}(x\delta_o)\}^{-1}V^{21}(x\delta_o)$, the Schwartz inequality implies that

$$\begin{aligned}\Omega_{1,G}^{-1} - \Omega_{1,s}^{-1} = {} & E[A_1(x)V^{12}(x\delta_o)\{V^{22}(x\delta_o)\}^{-1}V^{21}(x\delta_o)A_1'(x)] \\ & - E[A_1(x)V^{12}(x\delta_o)A_2'(x)]\{E[A_2(x)V^{22}(x\delta_o)A_2'(x)]\}^{-1}E[A_2(x)V^{21}(x\delta_o)A_1'(x)] \geq 0.\end{aligned}$$

This shows that $\hat{\theta}_{1,G}$ is efficient relative to $\hat{\theta}_{1,s}$. For the case that θ_1 and θ_2 contain some common parameters, the system GSMD estimator which incorporates common parameter constraints will be more efficient than the GSMD estimator of (θ_1, θ_2) without imposing the constraints. To see that for our semiparametric estimators, let θ_1 and θ_2 be function of parameters η. Then

$$\frac{\partial E(f'(z,\beta_o)|x)}{\partial\eta} - \frac{\partial E(f'(z,\beta_o)|x\delta_o)}{\partial\eta} = \left(\frac{\partial\theta_1'}{\partial\eta}, \frac{\partial\theta_2'}{\partial\eta}\right)A(x),$$

where $A(x) = \begin{pmatrix} A_1(x) & 0 \\ 0 & A_2(x) \end{pmatrix}$. Let $\hat{\eta}_G$ be the generalized semiparametric estimator of η and let $\hat{\theta}_{1,G}^{(c)} = \theta_1(\hat{\eta}_G)$ and $\hat{\theta}_{2,G}^{(c)} = \theta_2(\hat{\eta}_G)$ be the corresponding constrained estimators of θ_1 and θ_2. By Rao's delta method, the asymptotic variance matrix of $(\hat{\theta}_{1,G}^{(c)}, \hat{\theta}_{2,G}^{(c)})$ is

$$\begin{pmatrix} \frac{\partial\theta_1}{\partial\eta'} \\ \frac{\partial\theta_2}{\partial\eta'} \end{pmatrix} \left\{ \left(\frac{\partial\theta_1'}{\partial\eta}, \frac{\partial\theta_2'}{\partial\eta}\right) E[A(x)V^{-1}(x\delta_o)A'(x)] \begin{pmatrix} \frac{\partial\theta_1}{\partial\eta'} \\ \frac{\partial\theta_2}{\partial\eta'} \end{pmatrix} \right\}^{-1} \left(\frac{\partial\theta_1'}{\partial\eta}, \frac{\partial\theta_2'}{\partial\eta}\right).$$

The asymptotic variance of the unconstrained estimator $(\hat{\theta}_{1,G}, \hat{\theta}_{2,G})$ is $\{E(A(x)V^{-1}(x\delta_o)A'(x))\}^{-1}$. By the Schwartz inequality, the asymptotic variance of $(\hat{\theta}_{1,G}, \hat{\theta}_{2,G})$ is apparently larger than the asymptotic variance of $(\hat{\theta}_{1,G}^{(c)}, \hat{\theta}_{2,G}^{(c)})$.

For the sample selection and limited dependent variables models, some of the conditional moment restrictions can be expressed as conditional moments conditional on the occurrence of regimes. It is desirable to generalize the semiparametric estimation methods to cover such cases. Suppose there are L mutually exclusive and exhaustive regimes in a model. Let I_l be a dichotomous indictor of the regime l, $l = 1, \cdots, L$ such that $\sum_{l=1}^L I_l = 1$. For the regime l, the conditional moment equations are

$$E(f_l'(z,\beta_o)|x, I_l = 1) = E(f_l'(z,\beta_o)|x\delta_o, I_l = 1). \tag{4.8}$$

In addition to these conditional moments, the regime probabilities satisfy also the index restriction,

$$E(I_l|x) = E(I_l|x\delta_o), \quad l = 1, \cdots, L. \tag{4.9}$$

To unify notations, denote $f_0'(z,\beta_o) = (I_1, \cdots, I_{L-1})$ and $I_0 = 1$. The moment restrictions (4.8) and (4.9) can be summaried as

$$E(f_l'(z,\beta_o)|x, I_l = 1) = E(f_l'(z,\beta_o)|x\delta_o, I_l = 1), \quad l = 0, 1, \cdots, L. \tag{4.10}$$

The disturbances in this model are $u_l = f_l(z,\beta_o) - E(f_l(z,\beta_o)|x, I_l = 1)$, $l = 0, 1, \cdots, L$. It is apparent that $E(I_l u_l I_j u_j'|x) = 0$ and $E(I_l u_l I_s'|x) = 0$ for all $j \neq l$, $j, l, s = 1, \cdots, L$, i.e., all the relevant disturbances are mutually uncorrelated. The $E(f_l(z,\beta)|x_i, I_l = 1)$ can be estimated by

$$E_n(f_l(z,\beta)|x_i, I_l = 1) = \sum_{j\neq i}^n I_{lj} f_l(z_j,\beta) K\left(\frac{x_i - x_j}{a_n}\right) \Big/ \sum_{j\neq i}^n I_{lj} K\left(\frac{x_i - x_j}{a_n}\right). \tag{4.11}$$

Since

$$\begin{aligned}\frac{1}{(n-1)a_n^k}\sum_{j\neq i}^n I_{lj} f_l(z_j,\beta) K\left(\frac{x_i - x_j}{a_n}\right) & \xrightarrow{p} E(I_l f_l(z,\beta)|x_i)h(x_i) \\ & = E(f_l(z,\beta)|x_i, I_l = 1)E(I_l|x_i)h(x_i),\end{aligned}$$

and $\frac{1}{(n-1)a_n^k}\sum_{j\neq i}^n I_{lj} K\left(\frac{x_i - x_j}{a_n}\right) \xrightarrow{p} E(I_l|x_i)h(x_i)$ where $h(x)$ is the marginal density of x, $E_n(f_l(z,\beta)|x_i, I_l = 1)$ is a consistent estimate of $E(f_l(z,\beta)|x_i, I_l = 1)$. Similarly, $E(f_l(z,\beta)|x_i\delta, I_l = 1)$ can be estimated by

$$E_n(f(z,\beta)|x_i\delta, I_l = 1) = \sum_{j\neq i}^n I_{lj} f_l(z_j,\beta) J\left(\frac{x_i\delta - x_j\delta}{b_n}\right) \Big/ \sum_{j\neq i}^n I_{lj} J\left(\frac{x_i\delta - x_j\delta}{b_n}\right). \tag{4.12}$$

Let $V_l(x\delta_o)$ denote $\mathrm{Var}(u_l|x\delta_o, I_l = 1)$ for $l = 0, 1, \cdots, L$. Let

$$V_{n,l}(x_i\hat{\delta}) = E_n[f_l(z,\hat{\beta})f_l'(z,\hat{\beta})|x_i\hat{\delta}, I_l = 1] - E_n(f_l(z,\hat{\beta})|x_i\hat{\delta}, I_l = 1)E_n(f_l'(z,\hat{\beta})|x_i\hat{\delta}, I_l = 1), \tag{4.13}$$

Let $W_{n,l}(x\hat{\delta}) = V_{n,l}^{-1}(x\hat{\delta})$ be the weighting matrices for $l = 0, \cdots, L$. Denote

$$G_n^{(l)}(x_i,\theta) = E_n(f_l(z,\beta)|x_i, I_l = 1) - E_n(f_l(z,\beta)|x_i\delta, I_l = 1),$$

and $G_{n,\theta'}^{(l)}(x_i,\theta) = \frac{\partial E_n(f_l(z,\hat{\beta})|x_i, I_l=1)}{\partial\theta'} - \frac{\partial E_n(f_l(z,\hat{\beta})|x_i\hat{\delta}, I_l=1)}{\partial\theta'}$. The GSMD estimator of θ for this model can be defined as

$$\begin{aligned}\hat{\theta}_{c,G} = {} & -\left\{\sum_{i=1}^n t_n(x_i,\hat{\theta}) \sum_{l=0}^L I_{li} G_{n,\theta}^{(l)\prime}(x_i,\hat{\theta}) W_{n,l}(x_i\hat{\delta}) G_{n,\theta'}^{(l)}(x_i,\theta)\right\}^{-1} \\ & \times \sum_{i=1}^n t_n(x_i,\hat{\theta}) \sum_{l=0}^L I_{li} G_{n,\theta}^{(l)\prime}(x_i,\hat{\theta}) W_{n,l}(x_i\hat{\delta}) (G_n^{(l)}(x_i,\hat{\theta}) - G_{n,\theta'}^{(l)}(x_i,\hat{\theta})\hat{\theta}),\end{aligned} \tag{4.14}$$

where the trimming functions will be applied to all the conditional nonparametric functions and variances. Denote $G_{\theta'}^{(l)}(x,\theta) = \frac{\partial E(f_l(z,\beta)|x, I_l=1)}{\partial\theta'} - \frac{\partial E(f_l(z,\beta)|x\delta, I_l=1)}{\partial\theta'}$. With similar arguments for the proof of Proposition 4.1, we have the following asymptotic properties:[3]

Corollary 4.1 *For the conditional moment restrictions model, under the identification condition that, for any $\theta \neq \theta_o$, $E(f_l(z,\beta(\theta))|I_l = 1, x) - E(f_l(z,\beta(\theta))|I_l = 1, x\delta(\theta)) \neq 0$ with positive probability for some l, $l \in \{0, 1, \cdots, L\}$, $\hat{\theta}_{c,G}$ is a consistent, asymptotically normal estimator and*

$$\sqrt{n}(\hat{\theta}_{c,G} - \theta_o) \longrightarrow N(0, \Omega_{c,G}), \tag{4.15}$$

where $\Omega_{c,G} = D_c^{-1}(\theta_o)\Sigma_{c,G} D_c^{-1}(\theta_o)$ with

$$D_c(\theta_o) = E\left(\sum_{l=0}^{L} I_l G_{\theta}^{(l)'}(x,\theta_o) V_l^{-1}(x\delta_o) G_{\theta'}^{(l)}(x,\theta_o)\right),$$

and

$$\Sigma_{c,G} = E\left(\sum_{l=0}^{L} I_l G_{\theta}^{(l)'}(x,\theta_o) V_l^{-1}(x\delta_o) \mathrm{Var}(u_l|x, I_l = 1) V_l^{-1}(x\delta_o) G_{\theta'}^{(l)}(x,\theta_o)\right).$$

Furthermore, if $\mathrm{Var}(u_l|x, I_l = 1) = \mathrm{Var}(u_l|x\delta_o, I_l = 1)$ for all $l = 0, \cdots, L$, then $\Omega_{c,G} = D_c^{-1}(\theta_o)$.

5. Semiparametric Regression and Sample Selection Models

In a recent article, Chamberlain (1992) has investigated semiparametric efficiency bound for semiparametric models with conditional moment restriction. The conditional moment restriction considered is

$$E[\rho(x, y, \theta_o, q_o(x_2))|x] = 0, \tag{5.1}$$

where x_2 is a subvector of x, $\rho(x, y, \theta, \tau)$ is a known function, but $q_o(x_2)$ is an unknown measurable mapping to a finite dimensional space. Chamberlain (1992) has derived an efficiency bound for θ of models with the conditional moment restriction (5.1). Several concrete examples are considered. Among them are a semiparametric regression model [Engle et al. (1986), Robinson (1988)] and a sample selection model. As these models are special cases of models considered in this article, it is worthy of investigating whether our estimates can attain Chamberlain's efficiency bounds for these models.

5.1. Semiparametric Regression Model

The semiparametric regression model [Engle et al. (1986), Robinson (1988)] is specified as

$$E(y|x) = x_1\theta_o + q_o(x_2), \tag{5.2}$$

where the regression function consists of a parametric component and a nonparametric component. For this model, $\rho(x, y, \theta, \tau) = y - x_1\theta - \tau$ in Chamberlain's framework. Let $\sigma^2(x)$ denote $\mathrm{Var}(y|x)$. The information bound [Chamberlain (1992), p. 569] for θ is

$$J = E\left(\frac{x_1'x_1}{\sigma^2(x)}\right) - E\left[E\left(\frac{x_1'}{\sigma^2(x)}\Big|x_2\right) E\left(\frac{x_1}{\sigma^2(x)}\Big|x_2\right) \Big/ E\left(\frac{1}{\sigma^2(x)}\Big|x_2\right)\right]. \tag{5.3}$$

If the model is homoskedastic with $\sigma^2(x) = \sigma^2$ a constant, then

$$J = \sigma^{-2}E\{[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]'\} = \sigma^{-2}E[\mathrm{Var}(x_1|x_2)]. \tag{5.4}$$

If the variance satisfies also an index restriction with $\sigma^2(x) = \sigma^2(x_2)$, then

$$J = E\{\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]'\}. \tag{5.5}$$

[3] The moment conditions in Assumptions 2, 4, 6, and 7 need, of course, be modified accordingly to conditional moments conditional on the additional regime occurrence in additional to x and $x\delta$.

This model is a special index model with $x\delta = x_2$. In our estimation framework, $f(z, \beta(\theta)) = y - x_1\theta$ and $G_\theta(x, \theta_o) = -[x_1 - E(x_1|x_2)]'$. With $W(x)$ in Proposition 3.1 being an identity matrix (and trimming function goes to unity everywhere) for the homoscedastic variance model, $\Sigma_I = \sigma^2 E([x_1 - E(x_1|x_2)]'[x_1 - E(x_1|x_2)])$ and

$$\Omega_I = \sigma^2\{E([x_1 - E(x_1|x_2)]'[x_1 - E(x_1|x_2)])\}^{-1},$$

which attains the variance bound of Chamberlain. For the model where $\sigma^2(x) = \sigma^2(x_2)$, as $V(x_2) = \mathrm{Var}(f(z, \beta_o)|x_2) = \mathrm{Var}(y|x)$, Proposition 4.1 implies that

$$\Omega_G = \{E(\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]')\}^{-1},$$

which attains the variance bound of Chamberlain for the index variance model.[4] However, for the general case where $\sigma^2(x)$ is a general function of x, neither $\hat{\theta}_I$ nor $\hat{\theta}_G$ attain the efficiency bound. Apparently our estimation methods have utilized only index restrictions for estimation.

5.2. Sample Selection Model with Known Index

The sample selection model considered in [Chamberlain (1992), p.568] is

$$y^* = x_1\theta_o + x_2\gamma_o + \epsilon, \tag{5.6}$$

and

$$I = \begin{cases} 1, & \text{if } g_o(x_2, v) \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5.7}$$

where $x = (x_1, x_2)$ and $y = (y_1, I)$, where $y_1 = Iy^*$, are observed; so y^* is observed only if $I = 1$. The function g_o depends on x only via x_2 but is otherwise unrestricted. The disturbances ϵ and v satisfy that $E(\epsilon|x, v) = E(\epsilon|v)$ and v is independent of x_1 conditional on x_2. This model is thus a sample selection model with index restriction where the indices are $x\delta = x_2$. This model implies that $E(y_1|x, I = 1) = x_1\theta_o + q(x_2)$, where $q(x_2) = x_2\gamma_o + E(u|x_2, I = 1)$. Thus

$$\rho(x, y, \theta, \tau) = I[y_1 - x_1\theta - \tau] \tag{5.8}$$

in Chamberlain's framework. Chamberlain points out that one may also extend the ρ function to include the restriction that $E(I|x) = E(I|x_2)$, so that

$$\rho(x, y, \theta, \tau) = \begin{Bmatrix} I[y_1 - x_1\theta - \tau_1] \\ I - \tau_2 \end{Bmatrix}. \tag{5.8$'$}$$

Chamberlain derives the efficiency bound for this model and has pointed out that the efficiency bound for θ is the same with either one of the above ρ. Let $\sigma^2(x)$ denote $\mathrm{Var}(y_1|x, I = 1)$. The efficiency bound for θ is

$$J = E\left\{E(I|x_2)\left[E\left(\frac{x_1'x_1}{\sigma^2(x)}\right) - E\left(\frac{x_1'}{\sigma^2(x)}\Big|x_2\right) E\left(\frac{x_1}{\sigma^2(x)}\Big|x_2\right) \Big/ E\left(\frac{1}{\sigma^2(x)}\Big|x_2\right)\right]\right\}. \tag{5.9}$$

For the case that $\sigma^2(x)$ depends on x only through x_2, $\sigma^2(x) = \sigma^2(x_2)$ and

$$J = E\{E(I|x_2)\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]'\}. \tag{5.10}$$

In our framework, this model implies that $E(I(y_1 - x_1\theta_o)|x) = E(I(y_1 - x_1\theta_o)|x_2)$ and $E(I|x) = E(I|x_2)$. Hence $f'(z, \beta) = (I(y - x_1\theta), I)$. For the model that $\sigma^2(x) = \sigma^2(x_2)$, since $I(y - x_1\theta_o) = I(y - x_1\theta_o - q(x_2)) + Iq(x_2)$,

$$\begin{aligned} \mathrm{Var}(I(y - x_1\theta_o)|x) &= \mathrm{Var}(I(y_1 - x_1\theta_o - q(x_2))|x_2) + q^2(x_2)\mathrm{Var}(I|x_2) \\ &= \sigma^2(x_2)E(I|x_2) + q^2(x_2)\mathrm{Var}(I|x_2), \end{aligned} \tag{5.11}$$

[4] As in Chamberlain's discussion, the variance restrictions were not assumed to be known; otherwise, they should be incorporated into the conditional moment restriction (5.1). The efficiency bounds were obtained by the SMD or GSMD estimators if the true (but unknown) variance restrictions happened to be there.

and

$$V(x_2) = \begin{pmatrix} \sigma^2(x_2)E(I|x_2) + q^2(x_2)\mathrm{Var}(I|x_2) & q(x_2)\mathrm{Var}(I|x_2) \\ q(x_2)\mathrm{Var}(I|x_2) & \mathrm{Var}(I|x_2) \end{pmatrix}. \tag{5.12}$$

As $E(f_\theta'(z,\beta_o)|x) - \frac{\partial E(f'(z,\beta_o)|x_2)}{\partial\theta} = [-E(I|x_2)(x_1 - E(x_1|x_2, I=1)), 0] = [-E(I|x_2)(x_1 - E(x_1|x_2)), 0]$ and the first diagonal element of $V^{-1}(x_2)$ is $\{E(I|x_2)\sigma^2(x_2)\}^{-1}$ by a formula of inverse of a partitioned matrix, Proposition 4.1 implies that

$$\begin{aligned}\Omega_G &= \{E([E(I|x_2)(x_1 - E(x_1|x_2), 0]V^{-1}(x_2)[E(I|x_2)(x_1 - E(x_1|x_2), 0]')\}^{-1} \\ &= \{E(E(I|x_2)\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]')\}^{-1},\end{aligned}$$

which attains the efficiency bound of Chamberlain for this model. It is interesting to note that if the moment equation $E(I|x) = E(I|x_2)$ were ignored but only the equation $E(I(y_1 - x_1\theta_o)|x) = E(I(y_1 - x_1\theta_o)|x_2)$ was used in estimation, i.e., $f(z,\beta) = I(y - x_1\theta)$, then the limiting variance matrix for the GSMD estimator of θ would be $\{E(\frac{E^2(I|x_2)}{\mathrm{Var}(I(y-x_1\theta_o)|x_2)}[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]')\}^{-1}$, which were larger than the variance bound from (5.10) for this model as seen from (5.11). The point is that even though the equation $E(I|x) = E(I|x_2)$ does not contain θ, it helps to improve efficiency for the estimation of θ when it is included in estimation (as in a seemingly unrelated regression framework). This point has been noted in Chamberlain (1992) in his efficiency analysis. The above methods provide an example.

It is of interest to note that if the conditional moment restriction $E(y - x_1\theta_o|x, I = 1) = E(y - x_1\theta_o|x\delta_o, I = 1)$ is used for estimation. Corollary 4.1 implies that

$$\begin{aligned}\Omega_{c,G} &= \{E(I\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)]'[x_1 - E(x_1|x_2)])\}^{-1} \\ &= \{E(E(I|x)\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)]'[x_1 - E(x_1|x_2)])\}^{-1} \\ &= \{E(E(I|x_2)\sigma^{-2}(x_2)[x_1 - E(x_1|x_2)]'[x_1 - E(x_1|x_2)])\}^{-1},\end{aligned}$$

which attains the Chamberlain efficiency bound for the index variance model. The addition of the conditional moment restriction $E(I|x) = E(I|x_2)$ does not help because $I - E(I|x)$ is uncorrelated with the disturbance $y - x_1\theta_o - E(y - x_1\theta_o|x, I = 1) = y - x_1\theta_o - q(x_2)$. This conditional moment approach differs from the approach of using the moment condition of $I(y - x_1\theta)$ in that the conditional moment approach uses nonparametric estimate of $q(x_2)$ directly but the previous moment approach uses the nonparametric estimate of $E(I|x)q(x_2)$. The variance of $I(y - x_1\theta_o)$ conditional on x is larger than the variance of $I(y - x_1\theta_o - q(x_2))$ conditional on x as the former contains the additional term $q^2(x_2)\mathrm{Var}(I|x_2)$ in (5.11).

5.3. Sample Selection Model with Unknown Index

The previous sample selection model assumes that the indices in the choice equation are x_2. A more general model assumes that the index in the conditional moment and variance is $x\delta(\theta_o)$ with unknown coefficients. In this paragraph, we derive the limiting covariance matrix for such a model. For generality, consider a structural simultaneous equation where

$$g(y^*, x, \beta_o) = \epsilon \tag{5.13}$$

is a potential outcome equation before selection. The estimation of θ can be based on the equations $E(g(z,\beta_o)|I = 1, x) = E(g(z,\beta_o)|I = 1, x\delta_o)$ and $E(I|x) = E(I|x\delta_o)$. Let $G_\theta^{(1)}(x,\theta_o) = [E(g_\theta(z,\beta_o)|I = 1, x) - E(g_\theta(z,\beta_o)|I = 1, x\delta_o) + \frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'\nabla q(x\delta_o)]$, where $\nabla q(x\delta_o)$ denotes the gradient of $q(x\delta_o)$ with respect to $x\delta_o$. The limiting variance $\Omega_{c,G}$ of $\hat\theta_{c,G}$ follows from Corollary 4.1:

$$\left\{E\left(\frac{E(I|x\delta_o)}{\sigma^2(x\delta_o)}G_\theta^{(1)}(x,\theta_o)G_\theta^{(1)'}(x,\theta_o) + \frac{\{\nabla E(I|x\delta_o)\}^2}{\mathrm{Var}(I|x\delta_o)}\frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'(x - E(x|x\delta_o))\frac{\partial\delta(\theta_o)}{\partial\theta'}\right)\right\}^{-1}.$$

For the semiparametric binary choice model, the semiparametric efficiency bound has been derived in Cosslett (1987) and Chamberlain (1986), which is $\frac{\{\nabla E(I|x\delta_o)\}^2}{\mathrm{Var}(I|x\delta_o)}\frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'(x - E(x|x\delta_o))\frac{\partial\delta(\theta_o)}{\partial\theta'}$. Comparing the inverse of the efficiency bound and the limiting variance matrix of the GSMD estimator, the limiting

variance matrix of the GSMD estimator is asymptotically more efficient. Even for the case that $\theta' = (\beta', \delta')$, as long as $\nabla q(x\delta_o)$ does not vanish with probability one, the GSMD estimator of δ (with proper normalization) will be efficient relative to semiparametric estimators derived for estimating only the binary choice equation. The system estimation of both the discrete choice equation and the outcome equation gains efficiency because the disturbances of the discrete choice and the outcome equation are correlated.[5] For a linear simultaneous equation sample selection model, Powell (1987) has derived a two-step semiparametric instrumental varable method for the estimation of the outcome equation conditional on a given $\sqrt{n}$-consistent estimate (first stage estimate) of the parameters in the choice equation. Lee (1991) generalized Powell's method to take into account heteroskedastic errors and autocorrelated errors introduced from the first stage estimate. These methods are single equation semiparametric instrumental variable method. As will be shown in the next section that the GSMD is asymptotically equivalent to the best semiparametric instrumental variable estimator, the GSMD is asymptotically efficient relative to the two-step estimator in Powell (1987).

For an endogenous switching regression model with two regimes and an outcome equation for each regime, the GSMD estimation can be based on two conditional moment equations and a choice probability equation. Let $g_1(y_1^*, x, \beta_o) = \epsilon_1$ and $g_2(y_2^*, x, \beta_o) = \epsilon_2$ be, respectively, the potential outcome equations for regime 1 ($I = 1$) and regime 2 ($I = 0$). The implied conditional moment equations are

$$E(g_1(z,\beta_o)|I = 1, x) = E(g_1(z,\beta_o)|I = 1, x\delta_o) \tag{5.14}$$

and

$$E(g_2(z,\beta_o)|I = 0, x) = E(g_2(z,\beta_o)|I = 0, x\delta_o). \tag{5.15}$$

Let $q_1(x\delta_o) = E(\epsilon_1|I = 1, x\delta_o)$ and $q_2(x\delta_o) = E(\epsilon_2|I = 0, x\delta_o)$. Let $\sigma_1^2(x\delta_o)$ denote the conditional variance of $g_1(z,\beta_o) - q_1(x\delta_o)$ conditional on $I = 1$ and x, and let $\sigma_2^2(x\delta_o)$ denote the conditional variance of $g_2(z,\beta_o) - q_2(x\delta_o)$ conditional on $I = 0$ and x. With (5.14) and (5.15) and $E(I = 1|x) = E(I = 1|x\delta_o)$, the limiting variance of $\hat\theta_{c,G}$ follows from Corollary 4.1 which is

$$\begin{aligned}\Bigg\{E\Bigg(&\frac{E(I|x\delta_o)}{\sigma_1^2(x\delta_o)}G_\theta^{(1)}(x,\theta_o)G_\theta^{(1)'}(x,\theta_o) + \frac{E(1-I|x\delta_o)}{\sigma_2^2(x\delta_o)}G_\theta^{(2)}(x,\theta_o)G_\theta^{(2)'}(x,\theta_o) \\ &+ \frac{\{\nabla E(I|x\delta_o)\}^2}{\mathrm{Var}(I|x\delta_o)}\frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'(x - E(x|x\delta_o))\frac{\partial\delta(\theta_o)}{\partial\theta'}\Bigg)\Bigg\}^{-1},\end{aligned}$$

where

$$G_\theta^{(1)}(x,\theta_o) = [E(g_{1\theta}(z,\beta_o)|I = 1, x) - E(g_{1\theta}(z,\beta_o)|I = 1, x\delta_o) + \frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'\nabla q_1(x\delta_o)],$$

and

$$G_\theta^{(2)}(x,\theta_o) = [E(g_{2\theta}(z,\beta_o)|I = 0, x) - E(g_{2\theta}(z,\beta_o)|I = 0, x\delta_o) + \frac{\partial\delta'(\theta_o)}{\partial\theta}(x - E(x|x\delta_o))'\nabla q_2(x\delta_o)].$$

6. Semiparametric Instrumental Variables Estimation

Instead of the SMD or GSMD methods, it is possible to suggest semiparametric instrumental variables methods for the estimation of semiparametric simultaneous equation models. The (2.1) can be rewritten as

$$f(z,\beta_o) = E(f(z,\beta_o)|x\delta_o) + u, \tag{6.1}$$

where u satisfies $E(u|x) = 0$. Denote $u_n(z_i, x_i, \theta) = f(z_i,\beta) - E_n(f(z,\beta)|x_i\delta)$ and $u_{n,\theta'}(z_i, x_i, \theta) = \frac{\partial u_n(z_i,x_i,\theta)}{\partial\theta'}$. Furthermore, $u(z_i, x_i, \theta) = f(z_i,\beta) - E(f(z,\beta)|x_i\delta)$ and $u_{\theta'}(z_i, x_i, \theta) = \frac{\partial u(z_i,x_i,\theta)}{\partial\theta'}$. Let w be a

[5] The sample selection model with unknown indices has not been covered in Chamberlain analysis of conditional moment restriction model, it is unknown what is the efficient bound for the conditional moment equations of this model.

matrix of instrumental variables (functions of x).[6] A semiparametric instrumental variable (SIV) estimator taking into account possible variance structure of u can be defined as

$$\begin{aligned}\hat{\theta}_{IV} = &- \left\{ \sum_{i=1}^{n} t_n(x_i,\hat{\theta}) u'_{n,\theta}(z_i,x_i,\hat{\theta}) V_n^{-1}(x_i\hat{\delta}) w_i \left(\sum_{i=1}^{n} t_n(x_i,\hat{\theta}) w_i' V_n^{-1}(x_i\hat{\delta}) w_i \right)^{-1} \right. \\ &\left. \sum_{i=1}^{n} t_n(x_i,\hat{\theta}) w_i' V_n^{-1}(x_i\hat{\delta}) u_{n,\theta'}(z_i,x_i,\hat{\theta}) \right\}^{-1} \\ &\times \sum_{i=1}^{n} t_n(x_i,\hat{\theta}) u'_{n,\theta}(z_i,x_i,\hat{\theta}) V_n^{-1}(x_i\hat{\delta}) w_i \left(\sum_{i=1}^{n} t_n(x_i,\hat{\theta}) w_i' V_n^{-1}(x_i\hat{\delta}) w_i \right)^{-1} \\ &\times \sum_{i=1}^{n} t_n(x_i,\hat{\theta}) w_i' V_n^{-1}(x_i\hat{\delta}) \left(u_n(z_i,x_i,\hat{\theta}) - u_{n,\theta'}(z_i,x_i,\hat{\theta})\hat{\theta} \right).\end{aligned} \tag{6.2}$$

This two-step SIV estimtor is motivated from the following semiparametric nonlinear (weighted) two-stage method:

$$\min_{\theta} \sum_{i=1}^{n} t_n(x_i,\hat{\theta}) u_n'(z_i,x_i,\theta) V_n^{-1}(x_i\hat{\delta}) w_i \left(\sum_{i=1}^{n} w_i' V_n^{-1}(x_i\hat{\delta}) w_i \right)^{-1} \sum_{i=1}^{n} w_i' V_n^{-1}(x_i\hat{\delta}) u_n(z_i,x_i,\theta).$$

The asymptotic distribution of this two-step SIV estimator can be derived with similar arguments in the previous proofs. Some of the details are in the Appendix A6. Let $A = E\left(u_\theta'(z,x,\theta_o) V^{-1}(x\delta_o) w\right)$ and $\Lambda = E(w' V^{-1}(x\delta_o) w)$. By uniform convergence of nonparametric functions,

$$\sqrt{n}(\hat{\theta}_{IV} - \theta_o) = -\{A\Lambda^{-1}A'\}^{-1} A\Lambda^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n(x_i,\theta_o) w_i' V_n^{-1}(x_i\delta_o) u_n(z_i,x_i,\theta_o) + o_p(1).$$

It follows from the U-statistic theory that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n(x_i,\theta_o) w_i' V_n^{-1}(x_i\delta_o) u_n(z_i,x_i,\theta_o) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (w_i - E(w|x_i\delta_o))' V^{-1}(x_i\delta_o) u_i + o_p(1).$$

Hence $\sqrt{n}(\hat{\theta}_{IV} - \theta_o) \xrightarrow{D} N(0, \Omega_{IV})$ where

$$\begin{aligned}\Omega_{IV} = &\{A\Lambda^{-1}A'\}^{-1} A\Lambda^{-1} E([w - E(w|x\delta_o)]' V^{-1}(x\delta_o) \mathrm{Var}(u|x\delta_o) V^{-1}(x\delta_o)[w - E(w|x\delta_o)]) \\ &\times \Lambda^{-1}A'\{A\Lambda^{-1}A'\}^{-1}.\end{aligned}$$

For the model where $\mathrm{Var}(u|x) = \mathrm{Var}(u|x\delta_o)$, Ω_{IV} becomes

$$\Omega_{IV} = \{A\Lambda^{-1}A'\}^{-1} A\Lambda^{-1} \Lambda_e \Lambda^{-1} A' \{A\Lambda^{-1}A'\}^{-1}, \tag{6.3}$$

where $\Lambda_e = E([w - E(w|x\delta_o)]' V^{-1}(x\delta_o)[w - E(w|x\delta_o)])$. From this result, it is interesting to note that better instrumental variables are $w - E_n(w|x\hat{\delta})$ instead of w, if $E(w|x\delta_o) \neq 0$. Using $w - E_n(w|x\hat{\delta})$ as instrumental variables, the limiting variance matrix of the corresponding SIV will be

$$\Omega_{IV,e} = \{A\Lambda_e^{-1}A'\}^{-1}, \tag{6.4}$$

[6] Alternatively, one may assume that w does not provide extra information in the presence of x in any conditional expectation function conditional on x and w. If that were not the case, x should be enlarged to include w.

because $A = E\left(u_\theta'(z,x,\theta_o)V^{-1}(x\delta_o)w\right) = E\left(u_\theta'(z,x,\theta_o)V^{-1}(x\delta_o)[w - E(w|x\delta_o)]\right)$ by (3.3). By the generalized Schwartz inequality, $\Omega_{IV,e}^{-1} \geq \Omega_{IV}^{-1}$. It remains to compare the GSMD estimator with any SIV estimator. Since w is a function of x and $E(u_\theta'(z,x,\theta_o)|x) = G_\theta'(x,\theta_o)$,[7]

$$\begin{aligned}E\left(u_\theta'(z,x,\theta_o)V^{-1}(x\delta_o)[w - E(w|x\delta_o)]\right) &= E\left\{E\left(u_\theta'(z,x,\theta_o)V^{-1}(x\delta_o)[w - E(w|x\delta_o)]\right) \Big| x\right\} \\ &= E(G_\theta'(x,\theta_o)V^{-1}(x\delta_o)[w - E(w|x\delta_o)]).\end{aligned}$$

It follows that $\Omega_{IV,e}$ can be rewritten as

$$\Omega_{IV,e} = \left\{E\left(G_\theta'(x,\theta_o)V^{-1}(x\delta_o)[w - E(w|x\delta_o)]\right)\Lambda_e^{-1}E\left([w - E(w|x\delta_o)]'V^{-1}(x\delta_o)G_{\theta'}(x,\theta_o)\right)\right\}^{-1}.$$

Comparing this variance matrix with Ω_G in Proposition 4.1 for the model $\mathrm{Var}(u|x) = \mathrm{Var}(u|x\delta_o)$ by the generalized Cauchy-Schwartz inequality, it is apparent that $\Omega_G^{-1} \geq \Omega_{IV,e}^{-1}$. To show that the GSMD is a SIV estimator, take

$$w = E(f_{\theta'}(z,\beta_o)|x) - \nabla' E(f(z,\beta_o)|x\delta_o)\left(\frac{\partial\delta'(\theta_o)x'}{\partial\theta'} - E\left(\frac{\partial\delta'(\theta_o)x'}{\partial\theta'}\Big|x\delta_o\right)\right), \tag{6.5}$$

then $w - E(w|x\delta_o) = G_{\theta'}(x,\theta_o)$ from (3.3) and $\Omega_{IV,e} = \Omega_G$. The optimal instrumental variables vector is $G_{\theta'}(x,\theta_o)$ (or some consistent estimates of it). The GSMD is asymptotically equivalent to the best SIV estimator for the model where $E(u|x) = E(u|x\delta_o)$ and $\mathrm{Var}(u|x) = \mathrm{Var}(u|x\delta_o)$.

7. Conclusion

This article has considered the estimation of broad classes of general nonlinear simultaneous equation models with qualitative and limited dependent variables, and models with sample selection subject to index restrictions. A general semiparametric estimation method, which is motivated by the classical minimum distance estimation method of quantal response models, is introduced. Conditional on the vector of exogenous variables, (conditional) moment equations can be derived from a structural microeconomic model. Such equations contain unknown functions because of unspecified distribution of distrubances in the model. The unknown functions can, however, be eliminated by exploring index restrictions in the model. After eliminating such unknown functions, the implied moment equations are expressed as differences of conditional moment equations conditional on all exogenous variables and conditional moment equations conditional on indices. The implied moment equations become the estimating functions for our approach. To estimate the structural parameters, the conditional moment equations are replaced by nonparametric regression functions and the structural parameters are derived by minimizing some average distances of the nonparametric regression functions. The estimators are shown to be $\sqrt{n}$-consistent and asymptotically normal. The GSMD estimators which take into account of unknown heteroskedastic disturbances with index restrictions are also introduced. The GSMD estimators are shown to be efficient relative to unweighted SMD estimators. The GSMD estimators which take into account of correlation across different structural equations are shown to be asymptotically efficient relative to single equation estimates. For the estimation of semiparametric regression models and some sample selection models with known indices, efficiency bounds have been derived in Chamberlain (1992). Our semiparametric estimators can attain the efficient bound when the heteroskedastic distrubances happen to satisfy also the same index restriction. For models with indices involving unknown parameters, semiparametric instrumental variables approach is introduced. The GSMD estimator is asymptotically equivalent to the best semiparametric instrumental variables estimator. An interesting feature of the proposed estimation approach is that it provides a unified estimation framework for the estimation of regression-type and simultaneous equation type models. For the estimation of simultaneous equation models, all one shall concern about for structural estimation is the set of appropriate exogenous variables used for conditioning. These methods are of particular interest for the estimation of nonlinear simultaneous equations models with sample selection or limited dependent variables.

[7] If w were not a function of x, the following equality will still hold if w does not provide extra information in the presence of x. This is so, since $E(u_\theta'(z,x,\theta_o)|x,w) = E(u_\theta'(z,x,\theta_o)|x)$ under such a circumstance.

Appendix

A1: Some Useful Propositions

The following propositions will be used repeatedly in subsequent proofs of asymptotic properties of our estimators. The first proposition is useful for establishing uniform convergence in probability of nonparametric regression functions with index restrictions and their first and second order derivatives. This uniform law of large numbers generalizes slightly the uniform law of large numbers in Ichimura and Lee (1991). The following three propositions summerize the bias order of nonparametric functions and their first and second derivatives. These biases are familiar results in the nonparametric regression literature and semiparametric econometrics literature [e.g., in Rao (1983), Robinson (1988), Powell et al. (1989), and Ichimura and Lee (1991)]. They are summarized here for convenient reference and are useful to justify some of the regularity conditions in Appendix A2 on our model. The remaining propositions will be useful for deriving the asymptotic distributions of our estimators.

Proposition A1.1 (A Uniform Law of Large Numbers) *Let $\{z_i\}$ be a sequence of i.i.d. random vectors. The measurable function $h(z,\theta,a_n)$ takes the form $h(z,\theta,a_n) = \frac{1}{a_n^d} h_1(z,\theta) h_2\left(z,\theta,\frac{s(z,\theta)}{a_n}\right)$, where $a_n = O(\frac{1}{n^p})$, $p > 0$, $d \geq 0$, $\theta \in \Theta$, and $s(z,\theta)$ is a finite dimensional vector-valued function. Suppose that the following conditions are satisfied :*

(1) Θ is a compact subset of a finite dimensional Euclidean space.

(2) The function $h_1(z,\theta)$ is differentiable with respect to θ. The rth order moment, where $r \geq 2$, of $\sup_{\theta\in\Theta} |h_1(z,\theta)|$ is finite. The first moment of $\sup_{\theta\in\Theta} |\frac{\partial h_1(z,\theta)}{\partial\theta}|$ exists and is finite.

(3) $|h_2| \leq c$ for some constant c.

(4) $E(h_1^2 h_2^2) = O(a_n^{\bar{d}})$ uniformly in $\theta \in \Theta$, for some $\bar{d}$.

(5) The functions $h_2(z,\theta,u)$, and $s(z,\theta)$ satisfy the bounded Lipschitz condition of order 1 with respect to θ and u.

If $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{2(1+1/r)d-\bar{d}} = \infty$, then $\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^n h(z_i,\theta,a_n) - E(h(z,\theta,a_n))\right| \xrightarrow{p} 0$. Furthermore, in addition to the above conditions, if $E(h(z,\theta,a_n))$ converges to a limit function $h_\infty(\theta)$ uniformly in $\theta\in\Theta$, then $\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^n h(z_i,\theta,a_n) - h_\infty(\theta)\right| \xrightarrow{p} 0$.

Proof: This theorem generalizes slightly the uniform law of large number in [Ichimura and Lee (1991), pp. 22-23] in that the the condition (2) is used to replace the original conditions 2 and 3 in that article.[8] In the orginal proof, the distribution of z was divided in an interior component and a tail component. The proof can be modified by dividing the distribution of $\sup_\theta |h_1(z,\theta)|$ into interior and tail components instead. With this slight modification, the original arguments for the proof in Ichimura and Lee (1991) will go through with little change. Q.E.D.

Proposition A1.2 *Let $K(v)$ be a function on R^m with a bounded support D such that $\int_D |K(v)|dv < \infty$. Let $t(z,\theta)$ be a continuous m-dimensional random vector. Suppose that $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$, where $g(t|\theta)$ is the density function of $t(z,\theta)$, is uniformly continuous in t, uniformly in (θ,z_i).[9] Then*

$$\lim_{n\to\infty}\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\frac{1}{a_n^m}K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)\Big|z_i\right] - E[c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta]g(t(z_i,\theta)|\theta)\right\| = 0.$$

Furthermore, if $K(v)$ is a function with zero moments up to the order s^*, i.e., $\int_D v_1^{i_1}\cdots v_m^{i_m}K(v)dv = 0$, for all $i_j \geq 0$, $j = 1,\dots,m$, $i_1+\cdots+i_m < s^*$ and $\int_D \|v\|^{s^*}|K(v)|dv < \infty$, and $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$ is differentiable on R^m to the order s^*, and the s^* order derivatives are uniformly bounded, then

$$\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\frac{1}{a_n^m}K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)\Big|z_i\right] - E[c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta]g(t(z_i,\theta)|\theta)\right\| = O(a_n^{s^*}).$$

[8] The original result is formulated to include U-statistics. Here we only need statistics with a single sum. The above proposition can also be modified for U-statistics.

[9] A function $g(t,z_i,\theta)$ is said to be uniformly continuous in t uniformly in (θ,z_i) if $\forall\epsilon > 0$, there exisits a $\delta > 0$ (may depend on ϵ only) such that whenever $\|t_1-t_2\| \leq \delta$, $\|g(t_1,z_i,\theta) - g(t_2,z_i,\theta)\| \leq \epsilon$ for all (z_i,θ).

Proposition A1.3 *Let $K(v)$ be a function on R^m with a bounded support D such that $K(v)$ goes to zero at the boundary of D and its gradient $\frac{\partial K(v)}{\partial v}$ is bounded. Suppose that $\frac{\partial}{\partial t}[E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)]$, where $g(t|\theta)$ is the density function $t(z,\theta)$, are uniformly continuous in t, uniformly in (z_i,θ). Then*

$$\lim_{n\to\infty}\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\frac{1}{a_n^{m+1}}\frac{\partial K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)}{\partial v}\Big|z_i\right] - \frac{\partial}{\partial t}[E(c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta)g(t(z_i,\theta)|\theta)]\right\| = 0.$$

Furthermore, if $K(v)$ has zero moments up to the order s^*, $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$ is differentiable at t everywhere to the order s^*+1, and these derivatives are uniformly bounded, then

$$\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\frac{1}{a_n^{m+1}}\frac{\partial K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)}{\partial v}\Big|z_i\right] - \frac{\partial}{\partial t}[E(c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta)g(t(z_i,\theta)|\theta)]\right\| = O(a_n^{s^*}).$$

Proposition A1.4 *Let $K(v)$ be a twice differentiable function on R^m with a bounded support D such that $K(v)$ and its gradient $\frac{\partial K(v)}{\partial v}$ go to zero at the boundary of D, and the gradient $\frac{\partial K(v)}{\partial v}$ and its hessian matrix $\frac{\partial^2 K(v)}{\partial v\partial v'}$ are bounded. Suppose that $\frac{\partial^2}{\partial t\partial t'}[E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)]$ are uniformly continuous in t, uniformly in (z_i,θ). Then*

$$\lim_{n\to\infty}\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\frac{1}{a_n^{m+2}}\frac{\partial^2 K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)}{\partial v\partial v'}\Big|z_i\right] - \frac{\partial^2}{\partial v\partial v'}[E(c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta)g(t(z_i,\theta)|\theta)]\right\| = 0.$$

Proposition A1.5 *Let $\{C_{j,n}(z_i)\}$, $j = 1,2$ be two sequences of measurable functions of an i.i.d. sample $\{z_i\}$. The $\{d_n(z_i)\}$ is a sequence of measurable functions with the property that either $E(|d_n(z)|) < \infty$ uniformly in n or $\sup_{z_i}|d_n(z_i)| = O_P(1)$. Suppose that, for each j,*

(1) $\sup_{z_i}|E(C_{j,n}(z)|z_i) - C_n^{(j)}(z_i)| = O(a_{j,n}^{s_j})$, for some measurable functions $C_n^{(j)}(z_i)$, and

(2) $\sup_{z_i}\text{var}(C_{j,n}(z)|z_i) = O(\frac{1}{n a_{j,n}^{r_j}})$, $j = 1,2$.

If $\lim_{n\to\infty}\sqrt{n}a_{j,n}^{r_j} = \infty$ and $\lim_{n\to\infty}\sqrt{n}a_{j,n}^{2s_j} = 0$ for $j = 1,2$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n d_n(z_i)\left|C_{1,n}(z_i) - C_n^{(1)}(z_i)\right|\cdot\left|C_{2,n}(z_i) - C_n^{(2)}(z_i)\right| \xrightarrow{p} 0.$$

Proof: This is Lemma 6 in Lee (1992). The proof follows from the Markov and Cauchy inequalities. See also [Ichimura and Lee (1991), Lemma 10]. Q.E.D.

Proposition A1.6 *Let $\{A_n(z_i)\}$ and $\{C_{l,n}(z_i)\}$. $l = 1,2$ be sequences of measurable functions of an i.i.d. sample $\{z_i\}$. Suppose that,*

(1) $\frac{1}{n}\sum_{i=1}^n |A_n(z_i)| = O_P(1)$;

(2) for each $l\in\{1,2\}$, $C_{l,n}(z_i) = \frac{1}{(n-1)}\sum_{j\neq i}^n \frac{1}{a_{l,n}^{d_l}} h_1^{(l)}(z_j,z_i) h_2^{(l)}(z_j,z_i,\frac{S_l(z_j,z_i)}{a_{l,n}})$ satisfies the conditions (1)-(5) of Proposition A1.1 with Θ being the compact support of z_i; and

(3) $\sup_{z_i}|E(C_{l,n}(z)|z_i) - h_n^{(l)}(z_i)| = O(a_{l,n}^{s_l})$ for some $s_l > 0$ and measurable functions $h_n^{(l)}(z)$.

If $\lim_{n\to\infty}\frac{n^{1-(1/t)(1+1/r)}}{\ln n} a_{l,n}^{2(1+1/r)d_l-\bar{d}_l} = \infty$ and $\lim_{n\to\infty} n a_{l,n}^{2ts_l} = 0$, where $t = t_1+t_2$, for $l = 1,2$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n A_n(z_i)|C_{1,n}(z_i) - h_n^{(1)}(z_i)|^{t_1}|C_{2,n}(z_i) - h_n^{(2)}(z_i)|^{t_2} \xrightarrow{p} 0.$$

Proof: Since $\sup_{z_i}\left|n^{-1/2}\sum_{i=1}^n A_n(z_i)|C_{1,n}(z_i) - h_n^{(1)}(z_i)|^{t_1}|C_{2,n}(z_i) - h_n^{(2)}(z_i)|^{t_2}\right| \leq \sqrt{n}\sup_{z_i}|C_{1,n}(z_i) - h_n^{(1)}(z_i)|^{t_1}\sup_{z_i}|C_{2,n}(z_i) - h_n^{(2)}(z_i)|^{t_2}\cdot\frac{1}{n}\sum_{i=1}^n|A_n(z_i)|$, it is sufficient to investigate the convergence of

$n^{\frac{t_1}{2t}} \sup_{z_i} |C_{j,n}(z_i) - h_n^l(z_i)|^{t_1}$. As $a_{l,n} = O(1/n^{p_l})$ in Proposition A1.1, $n = O(1/a_{l,n}^{1/p_l})$. Therefore, $n^{1/(2t)} C_{l,n}(z_i)$ is proportional to $\frac{1}{(n-1)} \sum_{j\neq i}^n \frac{1}{a_{l,n}^{d_l+1/(2tp_l)}} h_1^{(l)}(z_j, z_i) h_2^{(l)}(z_j, z_i, s_l(z_j, z_i)/a_{l,n})$. Proposition A1.1 implies that if $\frac{n}{\ln n} a_{l,n}^{2(1+1/r)(d_l+1/(2tp_l))-\bar{d}_l}$ goes to infinity, $n^{1/(2t)}[C_{l,n}(z_i) - E(C_{l,n}(z_i)|z_i)] \xrightarrow{p} 0$ uniformly in z_i. This rate requirement is equivalent to that $\frac{n^{1-(1+1/r)/t}}{\ln n} a_{l,n}^{2(1+1/r)d_l - \bar{d}_l}$ goes to infinity. Since

$$
\begin{aligned}
& n^{\frac{t_1}{2t}} \sup_{z_i} |C_{l,n}(z_i) - h_n^{(l)}(z_i)|^{t_1} \\
\leq & \left\{ \sup_{z_i} |n^{1/(2t)}[C_{l,n}(z_i) - E(C_{l,n}(z_i)|z_i)]| + \sup_{z_i} |n^{1/(2t)}[E(C_{l,n}(z_i)|z_i) - h_n^{(l)}(z_i)| \right\}^{t_1},
\end{aligned}
$$

the results follow. Q.E.D.

Proposition A1.7 *Let $\{z_i\}$ be an i.i.d. sample and $\{\Phi_n(z, a_n)\}$ be a sequence of vector-valued random functions with bandwidth $\{a_n\}$. Suppose that*

(1) there exist square integrable functions $q_j(z)$, $j = 1, 2$ such that $|E(\Phi_n(z, a_n)|z_j)| \leq q_j(z_j)$, $j = 1, 2$,

(2) $E(\Phi_n(z, a_n)) = O(a_n^s)$ and $\text{var}(\Phi_n(z, a_n)) = O(\frac{1}{a_n^r})$,

(3) $\lim_{n\to\infty} E(\Phi_n(z, a_n)|z_j) = \psi_j(z_j)$, a.e., for some measurable functions ψ_j, $j = 1, 2$, and

(4) $\lim_{n\to\infty} \sqrt{n} a_n^s = 0$ and $\lim_{n\to\infty} n a_n^r = \infty$.

If $\psi_1(z)$ and $\psi_2(z)$ are zero a.e., then $\frac{1}{\sqrt{n}(n-1)} \sum_{i=1}^n \sum_{j\neq i}^n \Phi_n(z_i, z_j, a_n) \xrightarrow{p} 0$.

On the other hand, if $\lim_{n\to\infty} \{[\psi_1(z) + \psi_2(z)][\psi_1(z) + \psi_2(z)]'\} = \Sigma$, then

$$
\frac{1}{\sqrt{n}(n-1)} \sum_{i=1}^n \sum_{j\neq i}^n \Phi_n(z_i, z_j, a_n) \xrightarrow{D} N(0, \Sigma).
$$

Proof: This result generalizes slightly the central limit theorem for U-statistics in Hoeffding (1948). It follows from a U-statistics result in Powell, Stock and Stoker (1989). It is Lemma 8 in Lee (1992). Q.E.D.

Proposition A1.8 *Suppose that K is an r-dimensional kernel function with a bounded support D such that $\int_D |K(v)| dv < \infty$ and with a bandwidth a_n. Let $A_n(s|x_i\delta_o) = \frac{1}{(n-1)a_n^r} \sum_{j\neq i}^n s_j K\left(\frac{x_i\delta_o - x_j\delta_o}{a_n}\right)$ and let $g(x\delta_o|\theta_o)$ be the density of $x\delta_o$. Denote $A(s|x\delta_o) = E(s|x\delta_o) g(x\delta_o|\theta_o)$. Let $f_n(s, x\delta_o)$ be measurable functions such that $\sup_n |f_n(s, x\delta_o)|$ is square integrable. Suppose that*

(1) $E(s|x\delta_o) g(x\delta_o|\theta_o)$ is uniformly continuous at $x\delta_o$,

(2) $E(f_n(s, x\delta_o)|x\delta_o) g(x\delta_o|\theta_o)$ is continuous in $x\delta_o$ a.e. uniformly in n,

(3) there exists a measurable function $h(x\delta_o)$ such that

$$
\left| E(f_n(s, x\delta_o)|x\delta_o) \right| \leq h(x\delta_o)
$$

with $E|h(x\delta_o) A(s|x\delta_o)| < \infty$ for large n, and

(4) $\lim_{n\to\infty} E(f_n(s, x\delta_o)|x\delta_o) = c(x\delta_o)$ a.e.

If $\lim_{n\to\infty} n a_n^r = \infty$ and $\lim_{n\to\infty} n a_n^{2s} = 0$, then

$$
\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n f_n(s_i, x_i\delta_o)[A_n(s|x_i\delta_o) - A(s|x_i\delta_o)] \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^n s_i c(x_i\zeta_o) g(x_i\zeta_o|\delta_o) - E\left[c(x\zeta_o) E(s|x\zeta_o) g(x\zeta_o|\delta_o)\right].
\end{aligned}
$$

Proof:

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^n f_n(s_i, x_i\delta_o)[A_n(s|x_i\delta_o) - A(s|x_i\delta_o, \theta_o)] = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j\neq i}^n \Phi_n(\bar{s}_i, \bar{s}_j, a_n),
$$

where $\bar{s} = (s, x)$ and

$$
\Phi_n(\bar{s}_i, \bar{s}_j, a_n) = f_n(s_i, x_i\delta_o)\left[\frac{1}{a_n^r} s_j K\left(\frac{x_i\delta_o - x_j\delta_o}{a_n}\right) - A(s|x_i\delta_o)\right].
$$

By (1), $\lim_{n\to\infty} E(A_n(s|x_i\delta_o)|x_i) = A(s|x_i\delta_o)$. Therefore, $E(\Phi_n(\bar{s}_i, \bar{s}_j, a_n)|\bar{s}_i)$ converges to zero. On the other hand,

$$
E(\Phi_n(\bar{s}_j, \bar{s}_i, a_n)|\bar{s}_i) = s_i E\left[f_n(s_j, x_j\delta_o)\frac{1}{a_n^r} K\left(\frac{x_j\delta_o - x_i\delta_o}{a_n}\right)\Big|x_i\right] - E\{E[f_n(s, x\delta_o)|x\delta_o] A(s|x\delta_o)\}.
$$

Since $\lim_{n\to\infty} \left\{E\left[f_n(s_j, x_j\delta_o)\frac{1}{a_n^r} K\left(\frac{x_j\delta_o - x_i\delta_o}{a_n}\right)\Big|x_i\right] - E[f_n(s, x\delta_o)|x_i\delta_o] g(x_i\delta_o|\theta_o)\right\} = 0$ by condition (2), it follows that, by conditions (3) and (4) and the Lebesque convergence theorem,

$$
\lim_{n\to\infty} E(\Phi_n(\bar{s}_j, \bar{s}_i, a_n)|\bar{s}_i) = s_i c(x_i\delta_o) g(x_i\delta_o|\theta_o) - E\left(c(x\delta_o) E(s|x\delta_o) g(x\delta_o|\theta_o)\right) \quad a.e.
$$

The result of the proposition follows as the projection of a U-statistics (see Proposition A1.7). Q.E.D.

A2: Regularity Conditions, Nonparametric Regression and Related Functions

This appendix collects regularity conditions for our model and points out some of the useful properties of nonparametric (index) regression functions and their first and second order derivatives. Assumption 1 below contains the basic regularity conditions of our model. Asumption 2 contains regularity conditions on the regression function of $f(z, \beta)$ on x and its first and second order derivatives. Assumption 3 describes some basic desirable properties for the kernel function K and its bandwidth a_n. Assumptions 4 and 5 are conditions for the nonparametric regression function of $f(z, \beta)$ on $x\delta$ and related functions and the kernel function J. Assumption 6 contains additional regularity conditions for the SMD estimator. Finally, regularity conditions for the GSMD estimator are in Assumption 7.

Assumption 1:

(1.1) Θ is a compact convex subset of a finite dimensional Euclidean space, and the true parameter vector θ_o is in the interior of Θ.

(1.2) The sample observations (x_i, z_i), $i = 1, \cdots, n$, are i.i.d.[10]

(1.3) x is a k-dimensional vector of continuous random variables with a density $h(x)$ and a compact support $S_x \subseteq R^k$.

(1.4) $\delta(\theta)$ is twice differentiable w.r.t. θ, and its first two order derivatives are bounded on Θ.

(1.5) For each $\theta \in \Theta$, $x\delta$ is an m-dimensional vector of continuous variates with a density $p(x\delta|\theta)$.

(1.6) f is a vector-valued measurable function with known form which satisfies the relation $E[f(z, \beta(\theta_o))|x] = E[f(z, \beta(\theta_o)|x\delta_o)$.

Assumption 2:

(2.1) f is differentiable w.r.t. θ to the third order. The rth *order moment, where* $r \geq 2$ of $\sup_\Theta \|f(z, \beta)\|$, $\sup_\Theta \|\frac{\partial f(z,\beta)}{\partial \theta_{j_1}}\|$, and $\sup_\Theta \|\frac{\partial^2 f(z,\beta)}{\partial \theta_{j_1}\theta_{j_2}}\|$, where θ_j's are components of θ, are finite. The first moment of $\sup_\Theta \|\frac{\partial^3 f(z,\beta)}{\partial \theta_{j_1}\theta_{j_2}\theta_{j_3}}\|$, for all j_1, j_2, j_3, exist.

(2.2) $h(x)$, $E(\|f(z, \beta)\|^2|x)$, $E\left(\|\frac{\partial f(z,\beta)}{\partial \theta_{j_1}}\|^2|x\right)$, and $E\left(\|\frac{\partial^2 f(z,\beta)}{\partial \theta_{j_1}\theta_{j_2}}\|^2|x\right)$ are bounded on $S_x \times \Theta$.

(2.3) $h(x)$, $E(f(z, \beta)|x)$, $E\left(\frac{\partial f(z,\beta)}{\partial \theta_{j_1}}|x\right)$, and $E\left(\frac{\partial^2 f(z,\beta)}{\partial \theta_{j_1}\theta_{j_2}}|x\right)$ are differentiable w.r.t. x to the order s_1^*. These s_1^* order derivatives are uniformly bounded on $S_x \times \Theta$.

Assumption 3:

(3.1) $K(v)$ is a continuous kernel function on R^k with a bounded support.

(3.2) $K(v)$ is a kernel with zero moments up to the order s_1^*, i.e., $\int v_1^{i_1} \cdots v_k^{i_k} K(v) dv = 0$ for all $0 \leq i_l$, $l = 1, \cdots, k$, and $1 \leq i_1 + \cdots + i_k < s_1^*$.

---

[10] x and z may contain common variables.

(3.3) The bandwidth sequence $\{a_n\}$ is a sequence of positive constants such that $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{2(1+1/r)k-k} = \infty$.

Assumption 4:

(4.1) The functions $p(x\delta|\theta)$, $E(\|x\|^4|x\delta)$, $E(\|f(z,\beta)\|^2|x\delta)$, $E(\|x\|^4\|f(z,\beta)\|^2|x\delta)$, $E\left(\|\frac{\partial f(z,\beta)}{\partial\theta_{j_1}}\|^2|x\delta\right)$, $E\left(\|\frac{\partial f(z,\beta)}{\partial\theta_{j_1}}\|^2\|x\|^2|x\delta\right)$, and $E\left(\|\frac{\partial^2 f(z,\beta)}{\partial\theta_{j_1}\theta_{j_2}}\|^2|x\delta\right)$, where θ_j's are components of θ, are bounded on $S_x \times \Theta$.

(4.2) $p(x\delta|\theta)$, $E(f(z,\beta)|x\delta)$, and $E\left(\frac{\partial f(z,\beta)}{\partial\theta_{j_1}}|x\delta,\theta\right)$, and $E\left(\frac{\partial^2 f(z,\beta)}{\partial\theta_{j_1}\theta_{j_2}}|x\delta\right)$ are differentiable w.r.t. $x\delta$ to the order s_2^*, and these s_2^* order derivatives are uniformly bounded on $S_x \times \Theta$. $E(x_i - x|x\delta, x_i)p(x\delta|\theta)$ and $E(f(z,\beta)(x_i - x)|x\delta, x_i)p(x\delta|\theta)$ are differentiable w.r.t. $x\delta$ to the order $s_2^* + 1$, and these $s_2^* + 1$ order derivatives are uniformly bounded on $S_x \times \Theta$.

(4.3) The function $E\left(\frac{\partial^2 f(z,\beta)}{\partial\theta_{j_1}\partial\theta_{j_2}}|x\delta\right)p(x\delta|\theta)$; the first order derivatives of $E(x_i - x|x\delta, x_i)p(x\delta|\theta)$, $E(f(z,\beta)(x_i - x)|x\delta, x_i)p(x\delta|\theta)$, and $E(\frac{\partial f(z,\beta)}{\partial\theta_{j_1}}(x_i - x)|x\delta, x_i)p(x\delta|\theta)$ w.r.t. $x\delta$; and the second order derivatives of $E((x_i - x)\otimes(x_i - x)|x\delta, x_i)p(x\delta|\theta)$ and $E(f(z,\beta)(x_i - x)\otimes(x_i - x)|x\delta, x_i)p(x\delta|\theta)$ w.r.t. $x\delta$ are uniformly continuous in $x\delta$, uniformly in $(x_i, \theta) \in S_x \times \Theta$.[11]

Assumption 5:

(5.1) $J(u)$ is a twice continuously differentiable kernel function on R^m with a bounded support. $\frac{\partial^2 J(u)}{\partial u \partial u'}$ satisfies the bounded Lipschitz condition of order 1 w.r.t. u.

(5.2) $J(u)$ is a kernel with zero moments up to the order s_2^* with $s_2^* > 2$.

(5.3) The bandwidth sequence $\{b_n\}$ is a sequence of constants such that $\lim_{n\to\infty} b_n = 0$ and

$$\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+2)-m} = \infty.$$

Assumption 6:

(6.1) The weighting function $W(x)$ is bounded.

(6.2) The set X where $X = \{x|W(x) \neq 0\}$, is contained in the interior of S_x.

(6.3) $h(x)$ is bounded away from zero on X, and $p(x\delta|\theta)$ is bounded away from zero uniformly on $X \times \Theta$.[12]

(6.4) $E(G'_\theta(x,\theta_o)W(x)G_{\theta'}(x,\theta_o))$ is nonsingular.

(6.5) The bandwidths a_n and b_n satisfy the rate of convergence that $\lim_{n\to\infty} na_n^{2k} = \infty$, $\lim_{n\to\infty} na_n^{2s_1^*} = 0$; and $\lim_{n\to\infty} nb_n^{2(m+2)} = \infty$ and $\lim_{n\to\infty} nb_n^{2s_2^*} = 0$.

Assumption 7:

(7.1) The rth moment of $\sup_{\theta\in\Theta} \|f(z,\beta)f'(z,\beta)\|$ exists, where $r \geq 2$.

(7.2) $E(f(z,\beta)f'(z,\beta)|x)$ is uniformly continuous on $S_x \times \Theta$.

(7.3) $V(x\delta_o)$ is positive definite for each $x \in S_x$.

(7.4) The matrix $E(G'_\theta(x,\theta_o)V^{-1}(x\delta_o)G_\theta(x,\theta_o))$ is nonsingular.

(7.5) The bandwidths a_n and b_n satisfy the following rates of convergence:

$$\lim_{n\to\infty} \frac{n^{1-\frac{1}{3}(1+\frac{1}{r})}}{\ln n} a_n^{2(1+\frac{1}{r})(k+3/2)-k} = \infty, \quad \lim_{n\to\infty} na_n^{2(k+4)} = \infty, \quad \lim_{n\to\infty} na_n^{4(s_1^*-2)} = 0, \quad \lim_{n\to\infty} na_n^{2s_1^*} = 0,$$

and

$$\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+3)-m} = \infty, \quad \lim_{n\to\infty} \frac{n^{1-\frac{1}{3}(1+\frac{1}{r})}}{\ln n} b_n^{2(1+\frac{1}{r})(m+2)-m} = \infty, \quad \lim_{n\to\infty} nb_n^{2(m+4)} = \infty,$$
$$\lim_{n\to\infty} nb_n^{4(s_2^*-2)} = 0, \quad \lim_{n\to\infty} nb_n^{2s_2^*} = 0.$$

[11] $\otimes$ refers to the Kronecker product.

[12] Assuming that the support S_x is a rectangle, a simple procedure to construct such an X is to truncate the variables in x. More sophisticated procedure is to trim some fixed proportion of the values in the tail distribution for each variable. For the latter, the asymptotic analysis will only be slightly complicated under LeCam's discretization device.

Assumptions 2, 3, 4, and 5 are used mainly to justify the uniform convergence of nonparametric functions and their derivatives to their desirable limiting functions. These assumptions provide the regularity conditions and bandwidth rates for the applications of Propositions A1.1-A1.4 in Appendix A1. The additional rates of convergence in Assumptions 6 and 7 are needed mainly for the applications of Propositions A1.5-A1.7 to derive the asymptotic distributions of the SMD and GSMD estimators. The rate requirements in Assumptions 3, 5, 6, and 7 of the bandwidth parameters a_n and b_n depend on the existence of moments of order r. The rate requirements in Assumption 7 are, in general, stronger than the requirements in other assumptions. This is so because the trimming of the regressors and indices are relaxing as sample size increases. The nonparametric functions need to converge fast enough to dominate the divergence of $\frac{1}{B_n(x_i)}$ and $\frac{1}{B_{Jn}(x_i,\theta)}$. If trimming of the tail distributions of x is fixed as in Assumption 6, then $\frac{1}{B_n(x_i)}$ and $\frac{1}{B_{Jn}(x_i,\theta)}$ will be stochastically bounded and the rate requirements can be weaker. When $r = \infty$, using the fact that, for any real numbers μ and ν with $\nu > 0$, $n^\nu/[\ln n]^\mu$ goes to infinity, the rate requirements in Assumptions 3, 5 and 6 are satisfied if $\lim_{n\to\infty} na_n^{2k} = \infty$, $\lim_{n\to\infty} na_n^{2s_1^*} = 0$ and $\lim_{n\to\infty} nb_n^{2(m+2)} = \infty$ and $\lim_{n\to\infty} nb_n^{2s_2^*} = 0$. The rate requirements in Assumptions 3, 5, 6, and 7 are satisfied if $\lim_{n\to\infty} na_n^{2(k+4)} = \infty$ and $\lim_{n\to\infty} na_n^{2s_1^*} = 0$ for a_n, and if $\lim_{n\to\infty} nb_n^{2(m+4)} = \infty$ and $\lim_{n\to\infty} nb_n^{2s_2^*} = 0$ for b_n when $r = \infty$. Such an a_n exits only if $s_1^* > k + 4$, and b_n exists only if $s_2^* > m + 4$. For $r = 2$, if $\lim_{n\to\infty} \frac{n}{(\ln n)^2} a_n^{4k+9} = \infty$ and $\lim_{n\to\infty} na_n^{2s_1^*} = 0$ for a_n, and $\lim_{n\to\infty} \frac{n}{(\ln n)^2} b_n^{4(m+3)} = \infty$ and $\lim_{n\to\infty} nb_n^{2s_2^*} = 0$ for b_n, then the rate requirements in the above assumptions will be satisfied. Finally we note that the assumption that x consists of all continuous random variables in Assumption 1.3 can be relaxed to allow mixed continuous variables and discrete variables with finite supports. Nonparametric regression functions with discrete regressors with finite supports can be found in Bierans (1987).[13]Our subsequent proofs can be generalized to cover such cases. The continuous assumption greatly simplifies the presentation of proofs. The assumption of compact supports (or finite supports for discrete variables), however, can not be easily relaxed. This is so even for the parametric binary logit model. For models with growing regressors, the logit maximum likelihood estimator can be consistent and asymptotically normal only if the regressors grow with some slow rates [see, for example, Gourieroux and Monfort (1981)].

The following paragraphs provide a brief summary of elementary properties of $E_n(f(z,\beta)|x_i\delta)$ and its first and second order derivatives under Assumptions 1, 4 and 5. Similar properties hold for $E_n(f(z,\beta)|x)$ under Assumptions 1-3. These properties generalize some of the familiar properties in the nonparametric regression and semiparametric econometrics literature [e.g., Rao (1983), Robinson (1988), Powell et al. (1989), and Ichimura and Lee (1991)].

Since $E(f(z,\beta)|x\delta)p(x\delta|\theta)$ and $E(f^2(z,\beta)|x\delta)p(x\delta|\theta)$ are bounded on $S_x \times \Theta$, the variances of $A_{Jn}(x_i,\theta)$ and $B_{Jn}(x_i)$ have the familiar order $O(\frac{1}{nb_n^m})$ uniformly on $S_x \times \Theta$. For any constant $c \geq 0$, the uniform law of large numbers of Proposition A1.1 implies that

$$\sup_{S_x\times\Theta} \left|b_n^{-c}A_{Jn}(x_i,\theta) - E[b_n^{-c}A_{Jn}(x_i,\theta)]\right| = o_P(1),$$

if $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+c)-m} = \infty$. With the kernel J of order s_2^*, under Assumptions 4 and 5, Proposition A1.2 implies that

$$\sup_{S_x\times\Theta} |E(A_{Jn}(x,\theta)|x_i) - A_J(x_i,\theta)| = O(b_n^{s_2^*}),$$

where $A_J(x,\theta) = E(f(z,\beta)|x\delta)p(x\delta|\theta)$. Hence, if $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+c)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-c} = 0$,

$$\sup_{S_x\times\Theta} |b_n^{-c}[A_{Jn}(x,\theta) - A_J(x,\theta)]| = o_P(1). \qquad (A2.1)$$

[13] In the statistics literature, Devroye and Wayner (1980) provides very general conditions for the proof of some consistent properties of nonparametric regression functions which cover the mixed continuous and discrete regressors without restrictions on their supports.

Similar conclusion holds for $B_{Jn}(x,\theta)$ with its limit being $B_J(x,\theta) = p(x\delta|\theta)$. Let $t_n(x,\theta)$ be a trimming function such that $t_n(x,\theta) \neq 0$ only if $B_{Jn}(x,\theta) \geq \frac{\Delta}{2} b_n$. Since

$$
\begin{aligned}
& E_n(f(z,\beta)|x_i\delta) - E(f(z,\beta)|x_i\delta) \\
&= \frac{1}{B_{Jn}(x_i,\theta)B_J(x_i,\theta)}[(A_{Jn}(x_i,\theta) - A_J(x_i,\theta))B_J(x_i,\theta) - A_J(x_i,\theta)(B_{Jn}(x_i,\theta) - B_J(x_i,\theta))], \\
& \sup_{S_x} |t_n(x,\theta)\,[E_n(f(z,\beta)|x_i\delta) - E(f(z,\beta)|x_i\delta)]| \\
&\leq \frac{2}{\Delta}\{ \sup_{S_x\times\Theta} |b_n^{-1}[A_{Jn}(x,\theta) - A_J(x,\theta)]| + \sup_{S_x\times\Theta} |E(f(z,\beta)|x\delta)| \cdot \sup_{S_x\times\Theta} |b_n^{-1}[B_{Jn}(x,\theta) - B_J(x,\theta)]|\} \qquad (A2.2) \\
&= o_P(1),
\end{aligned}
$$

when $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+1)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-1} = 0$. As $\text{plim}_{n\to\infty} t_n(x,\theta) = 1$ at each x and θ, the above result indicates that by proper trimming of the tails of the nonparametric density estimate of $x\delta$ and relaxing it slowly as sample size increases, $E_n(f(z,\beta)|x_i\delta)$ can converge in probability uniformly to $E(f(z,\beta)|x_i\delta)$ in some desirable fashion.

Let $\nabla J(u) = \partial J(u)/\partial u$. The first order derivatives of $A_{Jn}(x_i,\theta)$ and $B_{Jn}(x_i,\theta)$ with respect to θ are

$$
\begin{aligned}
A_{Jn,\theta'}(x_i,\theta) &= \frac{1}{(n-1)b_n^m} \sum_{j\neq i}^{n} f_{\theta'}(z_j,\beta) J\left(\frac{x_i\delta - x_j\delta}{b_n}\right) \\
&+ \frac{1}{(n-1)b_n^{m+1}} \sum_{j\neq i}^{n} f(z_j,\beta)\left(\frac{\partial x_i\delta}{\partial\theta'} - \frac{\partial x_j\delta}{\partial\theta'}\right) \nabla J\left(\frac{x_i\delta - x_j\delta}{b_n}\right),
\end{aligned}
$$

and

$$
B_{Jn,\theta'}(x_i,\theta) = \frac{1}{(n-1)b_n^{m+1}} \sum_{j\neq i}^{n} \left(\frac{\partial x_i\delta}{\partial\theta'} - \frac{\partial x_j\delta}{\partial\theta'}\right) \nabla J\left(\frac{x_i\delta - x_j\delta}{b_n}\right).
$$

As $p(x\delta|\theta)$, $E(\|f_{\theta'}(z,\beta)\|^2|x\delta)$ and $E(\|f(z,\beta)(x_i - x)\|^2|x\delta, x_i)$ are bounded, their variances have order $O(\frac{1}{nb_n^{m+2}})$ uniformly on $S_x \times \Theta$. By Proposition A1.3 their biases have order $O(b_n^{s_2^*})$ uniformly on $S_x \times \Theta$ under Assumptions 4 and 5. Proposition A1.1 implies that as $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+1+c)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-c} = 0$,

$$
\sup_{S_x\times\Theta} \|b_n^{-c}\,[A_{Jn,\theta'}(x_i,\theta) - A_{J,\theta'}(x_i,\theta)]\| = o_P(1). \qquad (A2.3)
$$

Similar result holds for $B_{Jn,\theta'}(x_i,\theta)$. For the second order derivatives, it follows from Propositions A1.1 and A1.4 that as $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+2+c)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-c} = 0$,

$$
\sup_{S_x\times\Theta} \left\| b_n^{-c} \left[\frac{\partial^2 A_{Jn}(x_i,\theta)}{\partial\theta_l\partial\theta'} - \frac{\partial^2 A_J(x_i,\theta)}{\partial\theta_l\partial\theta'}\right]\right\| = o_P(1), \qquad (A2.4)
$$

for each component θ_l of θ. Similar result holds for $\frac{\partial^2 B_{Jn}(x_i,\theta)}{\partial\theta_l\partial\theta'}$.

Since $\frac{\partial E_n(f(z,\beta)|x\delta)}{\partial\theta'} = \frac{1}{B_{Jn}(x,\theta)} A_{Jn,\theta'}(x,\theta) - \frac{A_{Jn}(x,\theta)}{B_{Jn}^2(x,\theta)} B_{Jn,\theta'}(x,\theta)$ and its limit is

$$
\frac{\partial E(f(z,\beta)|x\delta)}{\partial\theta'} = \frac{1}{B_J(x,\theta)} A_{J\theta'}(x,\theta) - \frac{A_J(x,\theta)}{B_J^2(x,\theta)} B_{J\theta'}(x,\theta),
$$

$$
\begin{aligned}
& \frac{\partial E_n(f(z,\beta)|x\delta)}{\partial\theta'} - \frac{\partial E(f(z,\beta)|x\delta)}{\partial\theta'} \\
&= \frac{1}{B_{Jn}(x,\theta)}[A_{Jn,\theta'}(x,\theta) - A_{J\theta'}(x,\theta)] - \frac{E_{Jn}(f(z,\beta)|x\delta)}{B_{Jn}(x,\theta)}[B_{Jn,\theta'}(x,\theta) - B_{J\theta'}(x,\theta)] \\
&+ [E_{Jn}(f(z,\beta)|x\delta)B_{J\theta'}(x,\theta) - A_{J\theta'}(x,\theta)]\frac{1}{B_{Jn}(x,\theta)B_J(x,\theta)}[B_{Jn}(x,\theta) - B_J(x,\theta)] \\
&- [E_{Jn}(f(z,\beta)|x\delta) - E_J(f(z,\beta)|x\delta)]B_{J\theta'}(x,\theta)\frac{1}{B_J(x,\theta)}.
\end{aligned}
$$

From (A2.1), as $\sup_{S_x\times\Theta} |b_n^{-1}[B_{Jn}(x,\theta) - B_J(x,\theta)]| = o_P(1)$ when $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+1)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-1} = 0$, it follows that, with probability arbitrarily close to one for large n, whenever $B_{Jn}(x,\theta) \geq \Delta b_n$, $B_J(x,\theta) \geq \frac{\Delta}{2} b_n$. Therefore, under the rate $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+2)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-2} = 0$,

$$
\sup_{S_x} \left| t_n(x,\theta)\left(\frac{\partial E_n(f(z,\beta)|x\delta)}{\partial\theta'} - \frac{\partial E(f(z,\beta)|x\delta)}{\partial\theta'}\right)\right| = o_P(1). \qquad (A2.5)
$$

The second order derivatives of the nonparametric regression function is

$$
\begin{aligned}
\frac{\partial^2 E_n(f(z,\beta)|x\delta)}{\partial\theta_l\partial\theta'} = \frac{1}{B_{Jn}(x,\theta)}\Bigg(& \frac{\partial^2 A_{Jn}(x,\theta)}{\partial_l\theta\partial\theta'} - E_{Jn}(f(z,\beta)|x\delta)\frac{\partial^2 B_{Jn}(x,\theta)}{\partial\theta_l\partial\theta'} \\
&- \frac{\partial B_{Jn}(x,\theta)}{\partial\theta_l}\frac{\partial E_{Jn}(f(z,\beta)|x\delta)}{\partial\theta'} - \frac{\partial E_{Jn}(f(z,\beta)|x\delta)}{\partial\theta_l}\frac{\partial B_{Jn}(x,\theta)}{\partial\theta'}\Bigg).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
& \frac{\partial^2 E_n(f(z,\beta)|x\delta)}{\partial\theta_l\partial\theta'} - \frac{\partial^2 E(f(z,\beta)|x\delta)}{\partial\theta_l\partial\theta'} \\
&= \frac{1}{B_{Jn}(x,\theta)}\Bigg\{ \left[\frac{\partial^2 A_{Jn}(x,\theta)}{\partial\theta_l\partial\theta'} - \frac{\partial^2 A_J(x,\theta)}{\partial\theta_l\partial\theta'}\right] - \left[E_{Jn}(f(z,\beta)|x\delta)\frac{\partial^2 B_{Jn}(x,\theta)}{\partial\theta_l\partial\theta'} - E_J(f(z,\beta)|x\delta)\frac{\partial^2 B_J(x,\theta)}{\partial\theta_l\partial\theta'}\right] \\
&- \left[B_{Jn,\theta_l}(x,\theta)\frac{\partial E_{Jn}(f(z,\beta)|x\delta)}{\partial\theta'} - B_{J\theta_l}(x,\theta)\frac{\partial E_J(f(z,\beta)|x\delta)}{\partial\theta'}\right] \\
&- \left[\frac{\partial E_{Jn}(f(z,\beta)|x\delta)}{\partial\theta_l}B_{Jn,\theta'}(x,\theta) - \frac{\partial E_J(f(z,\beta)|x\delta)}{\partial\theta_l}B_{J\theta'}(x,\theta)\right]\Bigg\} + \Bigg(\frac{\partial^2 A_J(x,\theta)}{\partial\theta_l\partial\theta'} \\
&- E_J(f(z,\beta)|x\delta)\frac{\partial^2 B_J(x,\theta)}{\partial\theta_l\partial\theta'} - B_{J\theta_l}(x,\theta)\frac{\partial E_J(f(z,\beta)|x\delta)}{\partial\theta'} - \frac{\partial E_J(f(z,\beta)|x\delta)}{\partial\theta_l}B_{J\theta'}(x,\theta)\Bigg) \\
&\times \frac{1}{B_J(x,\theta)B_{Jn}(x,\theta)}(B_J(x,\theta) - B_{Jn}(x,\theta)).
\end{aligned}
$$

Under the rates that $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{2(1+1/r)(m+3)-m} = \infty$ and $\lim_{n\to\infty} b_n^{s_2^*-2} = 0$,

$$
\sup_{S_x\times\Theta} \left\| t_n(x,\theta)\frac{\partial^2 E_n(f(z,\beta)|x\delta)}{\partial\theta_l\partial\theta'} - \frac{\partial^2 E(f(z,\beta)|x\delta)}{\partial\theta_l\partial\theta'}\right\| = o_P(1). \qquad (A2.6)
$$

A3: Semiparametric-distance Estimator

Proof of Proposition 3.1: Let

$$
Q_{I,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} G_n'(x_i,\theta)W(x_i)G_n(x_i,\theta),
$$

and $Q_{I,n}^*(\theta) = \frac{1}{n}\sum_{i=1}^n G'(x_i,\theta)W(x_i)G(x_i,\theta)$. From Appendix A1, $E_n(f(z,\beta)|x_i)$ and $E_n(f(z,\beta)|x_i\delta)$ converge in probability, respectively, to $E(f(z,\beta)|x_i)$ and $E(f(z,\beta)|x_i\delta)$ uniformly in $(x_i,\theta) \in X \times \Theta$. As $\frac{1}{n}\sum_{i=1}^n \|W(x_i)\| = O_p(1)$, $Q_{I,n}(\theta) - Q_{I,n}^*(\theta)$ converges in probability to zero uniformly on Θ. The classical uniform law of large number (e.g., Amemiya [1985], Theorem 4.2.1) implies that $Q_{I,n}^*(\theta)$ converges in probability to $Q_I^*(\theta)$ uniformly on Θ, where $Q_I^*(\theta) = E(G'(x,\theta)W(x)G(x,\theta))$. Under the identification condition, $Q_I^*(\theta)$ has a unique global minimum at $\theta = \theta_o$. The consistency of $\hat\theta_I$ follows.

The SMD estimator $\hat\theta_I$ satisfies the first order condition: $\sum_{i=1}^n G_{n,\theta}(x_i,\hat\theta_I)W(x_i)G_n(x_i,\hat\theta_I) = 0$. Without loss of generality, take f to be a scalar valued function for simplicity.[14] By a mean value theorem,

$$
\begin{aligned}
\sqrt{n}(\hat\theta_I - \theta_o) = -\Big\{&\frac{1}{n}\sum_{i=1}^{n} G_{n,\theta}(x_i,\bar\theta)W(x_i)G_{n,\theta}'(x_i,\bar\theta) + \frac{1}{n}\sum_{i=1}^{n} G_{n,\theta\theta'}(x_i,\bar\theta)W(x_i)G_n(x_i,\bar\theta)\Big\}^{-1} \\
&\times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{G_{n,\theta}(x_i,\theta_o)W(x_i)G_n(x_i,\theta_o)\},
\end{aligned}
$$

[14] This simplifies only the notation for sencond order derivatives.

where $G_{n,\theta\theta'}$ denotes the second order derivative matrix of G_n w.r.t. θ. Since $\bar\theta$ converges in probability to θ_o, it follows from (A2.2), (A2.5) and (A2.6) that $G_n(x_i,\bar\theta) \xrightarrow{p} G(x_i,\theta_o) = 0$, $G_{n,\theta'}(x_i,\bar\theta) \xrightarrow{p} G_{\theta'}(x_i,\theta_o)$, and $G_{n,\theta\theta'}(x_i,\bar\theta) \xrightarrow{p} G_{\theta\theta'}(x_i,\theta_o)$, where $G_{\theta\theta'}$ is the second order derivative matrix of G w.r.t. θ, uniformly in $x_i \in X$. It follows that

$$\sqrt{n}(\hat\theta_I - \theta_o) = -\left\{E\left(G'_\theta(x,\theta_o)W(x)G_{\theta'}(x,\theta_o)\right) + o_p(1)\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{G'_{n,\theta}(x_i,\theta_o)W(x_i)G_n(x_i,\theta_o)\}.$$

Denote $A_n(x_i) = A_n(x_i,\theta_o)$, $A_{Jn}(x_i) = A_{Jn}(x_i,\theta_o)$ and $B_{Jn}(x_i) = B_{Jn}(x_i,\theta_o)$ for simplicity. As n goes to infinity, $A_n(x_i)$ converges in probability to $A(x_i)$ uniformly in $x_i \in X$, where $A(x_i) = E(f(z,\beta_o)|x_i)h(x_i)$; $B_n(x_i)$ converges in probability uniformly to $B(x_i)(= h(x_i))$; $A_{J,n}(x_i)$ converges in probability uniformly to $A_J(x)\left(= E(f(z,\beta_o)|x\delta_o)p(x\delta_o)\right)$; and $B_{J,n}(x_i)$ converges in probability uniformly to $B_J(x_i)(= p(x_i\delta_o|\theta_o))$. Since $\frac{\partial E_n(f(z,\beta_o)|x_i)}{\partial\theta} = \frac{1}{B_n(x_i)}A_{n,\theta}(x_i)$, and

$$\frac{\partial E_n(f(z,\beta_o)|x_i\delta_o)}{\partial\theta} = \frac{1}{B_{Jn}(x_i)}A_{Jn,\theta}(x_i) - \frac{A_{Jn}(x_i)}{B_{Jn}^2(x_i)}B_{Jn,\theta}(x_i),$$

$G_{n,\theta}(x_i,\theta_o)$ is a function of $A_n(x_i)$, $B_n(x_i)$, $A_{Jn}(x_i)$, $B_{Jn}(x_i)$, $A_{n,\theta}(x_i)$, $A_{Jn,\theta}(x_i)$, and $B_{Jn,\theta}(x_i)$. By using the expansion of difference, $G'_{n,\theta}(x_i,\theta_o) = G'_\theta(x_i,\theta_o) + R'_{n,1}(x_i)$ where $R_{n,1}(x_i) = R_{n,1,1}(x_i) + R_{n,1,2}(x_i)$,

$$\begin{aligned}
&R_{n,1,1}(x_i)\\
&= \frac{1}{B(x_i)}[A_{n,\theta}(x_i) - A_\theta(x_i)] - \frac{1}{B^2(x_i)}A_\theta(x_i)[B_n(x_i) - B(x_i)]\\
&- \frac{1}{B_J(x_i)}[A_{Jn,\theta}(x_i) - A_{J\theta}(x_i)] + \frac{E_J(x_i)}{B_J(x_i)}[B_{Jn,\theta}(x_i) - B_{J\theta}(x_i)]\\
&+ \frac{1}{B_J^2(x_i)}B_{J\theta}(x_i)[A_{Jn}(x_i) - A_J(x_i)] + \{A_{J\theta}(x_i) - 2E_J(x_i)B_{J\theta}(x_i)\}\frac{1}{B_J^2(x_i)}[B_{Jn}(x_i) - B_J(x_i)],
\end{aligned}$$

and

$$\begin{aligned}
&R_{n,1,2}(x_i)\\
&= \frac{1}{B_n(x_i)B^2(x_i)}A_\theta(x_i)[B_n(x_i) - B(x_i)]^2 - \frac{1}{B_n(x_i)B(x_i)}[B_n(x_i) - B(x_i)][A_{n,\theta}(x_i) - A_\theta(x_i)]\\
&+ \frac{1}{B_{Jn}(x_i)B_J(x_i)}[B_{Jn}(x_i) - B_J(x_i)][A_{Jn,\theta}(x_i) - A_{J\theta}(x_i)]\\
&+ \frac{1}{B_{Jn}^2(x_i)}[A_{Jn}(x_i) - A_J(x_i)][B_{Jn,\theta}(x_i) - B_{J\theta}(x_i)]\\
&+ \left\{\frac{E_J(x_i)}{B_{Jn}(x_i)B_J(x_i)}B_{J\theta}(x_i)\left(\frac{1}{B_{Jn}(x_i)} + \frac{2}{B_J(x_i)}\right) - \frac{1}{B_{Jn}(x_i)B_J^2(x_i)}A_{J\theta}(x_i)\right\}[B_{Jn}(x_i) - B_J(x_i)]^2\\
&- \frac{E_J(x_i)}{B_{Jn}(x_i)}\left(\frac{1}{B_{Jn}(x_i)} + \frac{1}{B_J(x_i)}\right)[B_{Jn}(x_i) - B_J(x_i)][B_{Jn,\theta}(x_i) - B_{J\theta}(x_i)]\\
&- \frac{1}{B_{Jn}(x_i)B_J(x_i)}\left(\frac{1}{B_{Jn}(x_i)} + \frac{1}{B_J(x_i)}\right)B_{J\theta}(x_i)[B_{Jn}(x_i) - B_J(x_i)][A_{Jn}(x_i) - A_J(x_i)].
\end{aligned}$$

On the other hand, since $G(x_i,\theta_o) = 0$, by an expansion up to the second order, $G_n(x_i,\theta_o) = \epsilon_n(x_i) + R_{n,2}(x_i)$, where

$$\begin{aligned}
\epsilon_n(x_i) &= \left[(A_n(x_i) - A(x_i)) - E(f(z,\beta_o)|x_i)(B_n(x_i) - B(x_i))\right]\frac{1}{h(x_i)}\\
&- \left[(A_{Jn}(x_i) - A_J(x_i)) - E(f(z,\beta_o)|x_i\delta_o)(B_{Jn}(x_i) - B_J(x_i))\right]\frac{1}{p(x_i\delta_o|\theta_o)},
\end{aligned}$$

and

$$\begin{aligned}
R_{n,2}(x_i) &= -\frac{1}{B^2(x_i)}\{[A_n(x_i) - A(x_i)][B_n(x_i) - B(x_i)] - E(f(z,\beta_o)|x_i)[B_n(x_i) - B(x_i)]^2\}\\
&+ \frac{1}{B_J^2(x_i)}\{[A_{Jn}(x_i) - A_J(x_i)][B_{Jn}(x_i) - B_J(x_i)] - E(f(z,\beta_o)|x_i\delta_o)[B_{Jn}(x_i) - B_J(x_i)]^2\}\\
&+ \frac{1}{B^2(x_i)B_n(x_i)}\{[A_n(x_i) - A(x_i)][B_n(x_i) - B(x_i)]^2 - E(f(z,\beta_o)|x_i)[B_n(x_i) - B(x_i)]^3\}\\
&- \frac{1}{B_J^2(x_i)B_{Jn}(x_i)}\{[A_{Jn}(x_i) - A_J(x_i)][B_{Jn}(x_i) - B_J(x_i)]^2 - E(f(z,\beta_o)|x_i\delta_o)[B_{Jn}(x_i) - B_J(x_i)]^3\}.
\end{aligned}$$

It follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G'_{n,\theta}(x_i,\theta_o)W(x_i)G_n(x_i,\theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}G'_\theta(x_i,\theta_o)W(x_i)\epsilon_n(x_i) + R_n,$$

where $R_n = R_n^{(1)} + R_n^{(2)} + R_n^{(3)}$ with

$$R_n^{(1)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}G'_\theta(x_i,\theta_o)W(x_i)R_{n,2}(x_i),$$

$$R_n^{(2)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(R_{n,1,1}(x_i) + R_{n,1,2}(x_i))'W(x_i)\epsilon_n(x_i),$$

and $R_n^{(3)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(R_{n,1,1}(x_i) + R_{n,1,2}(x_i))'W(x_i)R_{n,2}(x_i)$. Since each term within the summations of $R_n^{(1)}$, $R_n^{(2)}$, and $R_n^{(3)}$ contains errors of higher orders: i.e., $(A_n(x_i) - A(x_i))^2$, etc, Proposition A1.5 implies, under the assumed bandwidth rates in Assumptions 3, 5 and 6, that $R_n = o_P(1)$.

The remaining term can be analyzed as a U-statistic since

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G'_\theta(x_i,\theta_o)W(x_i)\epsilon_n(x_i) = \frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\{U_{1n}(w_i,w_j) + U_{2n}(w_i,w_j)\},$$

where $w_i = (x_i,z_i)$,

$$U_{1n}(w_i,w_j) = G'_\theta(x_i,\theta_o)W(x_i)[f(z_j,\beta_o) - E(f(z,\beta_o)|x_i)]\frac{1}{h(x_i)}\frac{1}{a_n^k}K\left(\frac{x_i - x_j}{a_n}\right),$$

and

$$U_{2n}(w_i,w_j) = -G'_\theta(x_i,\theta_o)W(x_i)[f(z_j,\beta_o) - E(f(z,\beta_o)|x_i\delta_o)]\frac{1}{p(x_i\delta_o|\theta_o)}\frac{1}{b_n^m}J\left(\frac{x_i\delta_o - x_j\delta_o}{b_n}\right).$$

$E[U_{ln}(w_i,w_j)|w_i]$ converges to zero for $l = 1,2$, but

$$\lim_{n\to\infty}E[U_{1n}(w_j,w_i)|w_i] = G'_\theta(x_i,\theta_o)W(x_i)u_i,$$

and

$$\lim_{n\to\infty}E[U_{2n}(w_j,w_i)|w_i] = -E\left\{G'_\theta(x,\theta_o)W(x)|x_i\delta_o\right\}u_i.$$

The asymptotic distribution follows from the central limit theorem for U-statistics in Proposition A1.7. Q.E.D.

Proof of Proposition 3.2 : Denote $r_n(x_i,\theta) = f_1(x_i,\beta) + E_n(f_2(z,\beta)|x_i) + E_n(f(z,\beta)|x_i\delta)$ and $r'_{n,\theta}(x_i,\theta) = \frac{\partial r'_n(x_i,\theta)}{\partial\theta}$. Let $Q_{F,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - r_n(x_i,\theta)]'W(x_i)[y_i - r_n(x_i,\theta)]$. By Proposition A1.1 and uniform laws of large numbers, $Q_{F,n}(\theta)$ converges in probability to $Q_F^*(\theta)$ uniformly in θ, where $Q_F^*(\theta) = E([y - r(x,\theta)]'W(x)[y - r(x,\theta)])$. As $Q_F^*(\theta)$ is uniquely minimized at θ_o, $\hat{\theta}_{NL}$ is consistent.

By a Taylor series expansion and the uniform convergence of the nonparametric regression functions and their derivatives,

$$\sqrt{n}(\hat{\theta}_{NL} - \theta_o) = \{E[r'_\theta(x,\theta_o)W(x)r_{\theta'}(x,\theta_o)] + o_p(1)\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{r'_{n,\theta}(x_i,\theta_o)W(x_i)(y_i - r_n(x_i,\theta_o))\}.$$

Let $A_{2,n}(x_i) = \frac{1}{(n-1)a_n^k}\sum_{j\neq i}^{n} f_2(z_j,\beta_o)K\left(\frac{x_i - x_j}{a_n}\right)$ and $\epsilon_i^* = y_i - r(x_i,\theta_o)$. By similar arguments in the proof of Proposition 3.2,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{r'_{n,\theta}(x_i,\theta_o)W(x_i)(y_i - r_n(x_i,\theta_o))\}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r'_\theta(x_i,\theta_o)W(x_i)\Big\{\epsilon_i^* - \Big([A_{2,n}(x_i) - E(f_2(z,\beta_o)|x_i)B_n(x_i)]\frac{1}{h(x_i)}\\
&\qquad - [A_{J,n}(x_i) - E(f(z,\beta_o)|x_i\delta_o)B_{J,n}(x_i)]\frac{1}{p(x_i\delta_o|\theta_o)}\Big)\Big\}\\
&= \frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} U(w_i,w_j,b_n),
\end{aligned}$$

where $w_i = (x_i, z_i)$ and

$$\begin{aligned}
U(w_i,w_j,b_n) = r'_\theta(x_i,\theta_o)W(x_i)\Big\{\epsilon_i^* - \Big([f_2(z_i,\beta_o) - E(f_2(z,\beta_o)|x_i)]\frac{1}{h(x_i)}\frac{1}{a_n^k}K\left(\frac{x_i - x_j}{a_n}\right)\\
- [f(z_i,\beta_o) - E(f(z,\beta_o)|x_i\delta_o)]\frac{1}{p(x_i\delta_o|\theta_o)}\frac{1}{b_n^m}J\left(\frac{x_i\delta_o - x_j\delta_o}{b_n}\right)\Big)\Big\}.
\end{aligned}$$

As $\lim_{n\to\infty} E[U(w_i,w_j,b_n)|w_i] = r'_\theta(x_i,\theta_o)W(x_i)\epsilon_i^*$, and

$$\lim_{n\to\infty} E[U(w_j,w_i,b_n)|w_i] = -r'_\theta(x_i,\theta_o)W(x_i)[f_2(z_i,\beta_o) - E(f_2(z,\beta_o)|x_i)] + E\{r'_\theta(x,\theta_o)W(x)|x_i\delta_o\}u_i,$$

the asymptotic distribution of $\hat{\theta}_{NL}$ follows from the U-statistic central limit theorem in Proposition A1.7.
Q.E.D.

A4: Trimming

The function of trimming is to control for the erratic behavior of the nonparametric regression estimates so that uniform convergence of the nonparametric regression functions and their derivatives is feasible on the relevant range. Let p be a distribution function with support on the interval $[\frac{\Delta}{2}, \Delta]$, where $\Delta > 0$ is a specified constant, such that both the first and second order derivates of p at the end points $\frac{\Delta}{2}$ and Δ are zero. A simple p with these smooth property is the following function

$$p(c) = \int_{-1}^{\frac{4}{\Delta}c-3}\frac{15}{16}(1 - w^2)^2 dw, \qquad \frac{\Delta}{2} \leq c \leq \Delta,$$

where $\frac{15}{16}(1 - w^2)^2$ is a density function with support $[-1, 1]$. This p is apparently a bounded polynomial function on $[\frac{\Delta}{2}, \Delta]$. There are three sets of relevant nonparametric regression functions in our method, namely, $E_n(f(z,\hat{\beta})|x_i)$, $E_n(f(z,\hat{\beta})|x_i\hat{\delta})$ and $V_n(x_i\hat{\delta})$. Define the following function

$$t^*(c) = \begin{cases} 1 & \Delta < c \\ p(c) & \frac{\Delta}{2} < c \leq \Delta \\ 0 & c \leq \frac{\Delta}{2} \end{cases} \tag{A4.1}$$

The denominator in $E_n(f(z,\hat{\beta})|x_i)$ is $B_n(x_i)$ and the corresponding trimming function will be $t^*(B_n(x_i)/a_n)$. The sample x_i will be deleted in the estimation procedure if $B_n(x_i)$ is less than $\frac{\Delta}{2}a_n$ and is down-weighted if it is greater than $\frac{\Delta}{2}a_n$ but less than Δa_n. Since a_n goes to zero as n goes to infinity and $B_n(x_i)$ converges to the density $h(x_i)$ of x at x_i, $t^*(B_n(x_i)/a_n)$ goes to 1 in probability as n goes to infinity. A similar trimming has been used in Robinson (1988). For $E_n(f(z,\hat{\beta})|x_i\hat{\delta})$, the denominator is $B_{Jn}(x_i,\hat{\theta})$, and the trimming function will be $t^*(B_{Jn}(x_i,\hat{\theta})/b_n)$. For the nonparametric variance matrix, the sample observations at which $V_n(x_i\hat{\delta})$ are nearly singular will be deleted. Since

$$V_n(x_i\hat{\delta}) = V_{Jn}(x_i,\hat{\theta})/B_{Jn}^2(x_i,\hat{\theta}), \tag{A4.2}$$

where

$$V_{Jn}(x_i,\hat{\theta}) = B_{Jn}(x_i,\hat{\theta})\frac{1}{(n-1)b_n^m}\sum_{j\neq i}^{n} f(z_j,\hat{\beta})f'(z_j,\hat{\beta})J\left(\frac{x_i\hat{\delta} - x_j\hat{\delta}}{b_n}\right) - A_{Jn}(x_i,\hat{\theta})A'_{Jn}(x_i,\hat{\theta}), \tag{A4.3}$$

from (4.1). Thus $W_n(x_i\hat{\delta}) = B_{Jn}^2(x_i,\hat{\theta})V_{Jn}^*(x_i,\hat{\theta})/\det[V_{Jn}^*(x_i,\hat{\theta})]$ where V_{Jn}^* is the adjoint of V_{Jn}. The determinant $\det[V_{Jn}^*(x_i,\hat{\theta})]$ is a finite order polynomial of nonparametric estimators of unknown functions. The trimming function for $W_n(x_i\hat{\delta})$ can be $t^*(\det([V_{Jn}^*(x_i,\hat{\theta})])/b_n)$. The overall trimming function $t_n(x_i,\hat{\theta})$ is the product of the above trimming functions, i.e.,

$$t_n(x_i,\theta) = t^*(B_n(x_i)/a_n)\cdot t^*(B_{Jn}(x_i,\theta)/b_n)\cdot t^*(\det([V_{Jn}^*(x_i,\hat{\theta})])/b_n). \tag{A4.4}$$

A5: Generalized Semiparametric Estimator

Proof of Proposition 4.1

To simplify notation, let

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} t_n(x_i,\theta)G'_{n,\theta}(x_i,\theta)W_n(x_i\delta)G_{n,\theta'}(x_i,\theta),$$

and

$$U_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t_n(x_i,\theta)G'_{n,\theta}(x_i,\theta)W_n(x_i\delta)G_n(x_i,\theta).$$

It follows that

$$\sqrt{n}(\hat{\theta}_G - \theta_o) = \sqrt{n}(\hat{\theta} - \theta_o) - D_n^{-1}(\hat{\theta})U_n(\hat{\theta}). \tag{A5.1}$$

By the uniform convergence of nonparametric regression functions and their derivatives in Appendix A2, $D_n(\hat{\theta}) - D_n^*(\hat{\theta}) = o_P(1)$ where $D_n^*(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} t_n(x_i,\hat{\theta})G'_\theta(x_i,\hat{\theta})V^{-1}(x_i\hat{\delta})G_{\theta'}(x_i,\hat{\theta})$. Define $D_n^o(\theta) = \frac{1}{n}\sum_{i=1}^{n} G'_\theta(x_i,\theta)V^{-1}(x_i\delta)G_{\theta'}(x_i,\theta)$. As $t_n(x_i,\hat{\theta})$ converges in probability to 1 a.e., it follows by the Markov inequality and the dominated convergence theorem that $D_n^*(\hat{\theta}) - D_n^o(\hat{\theta}) = o_P(1)$. On the other hand, $D_n^o(\theta)$ converges in probability to $D(\theta)$, where $D(\theta) = E(G'_\theta(x,\theta)V^{-1}(x\delta)G_{\theta'}(x,\theta))$, uniformly in θ. Since $\hat{\theta}$ is a consistent estimate of θ and $D(\theta)$ is continuous in θ, $D(\hat{\theta})$ converges to $D(\theta_o)$ in probability. Therefore, $D_n(\hat{\theta})$ converges in probability to $D(\theta_o)$.

It remains to analyze $U_n(\hat{\theta})$. Without loss of generalization, consider that f is a scalar function for notational simplicity. By a mean value theorem, $U_n(\hat{\theta}) = U_n(\theta_o) + U_n^d(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta_o)$, where

$$U_n^d(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}\{t_n(x_i,\theta)G_{n,\theta}(x_i,\theta)W_n(x_i\delta)G_n(x_i,\theta)\}.$$

The latter term can be decomposed into a sum of two terms: $U_n^d(\theta) = U_n^{d1}(\theta) + D_n(\theta)$ where

$$U_n^{d1}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}\{t_n(x_i,\theta)G_{n,\theta}(x_i,\theta)W_n(x_i\delta)\}G_n(x_i,\theta).$$

Since $\bar{\theta}$ lies between $\hat{\theta}_I$ and θ_o, $G_n(x_i,\bar{\theta})$ converges in probability to $G(x_i,\theta_o)=0$. By uniform convergence of nonparametric functions, $U_n^{d1}(\bar{\theta})=o_P(1)$. Hence $U_n^d(\bar{\theta})=D_n(\bar{\theta})+o_P(1)=D^o(\theta_o)+o_P(1)$. Therefore (A5.1) is reduced to $\sqrt{n}(\hat{\theta}_G-\theta_o)=-D_n^{-1}(\bar{\theta})U_n(\theta_o)+o_p(1)$.

Let $R_{n3}(x_i)=V_n^{-1}(x_i\delta_o)-V^{-1}(x_i\delta_o)=V_n^{-1}(x_i\delta_o)[V(x_i\delta_o)-V_n(x_i\delta_o)]V^{-1}(x_i\delta_o)$. We have $W_n(x_i\delta_o)=V^{-1}(x_i\delta_o)+R_{n3}(x_i)$. Define

$$t_n^{\infty}(x,\theta_o)=t^*\left(\frac{B(x)}{a_n}\right)t^*\left(\frac{B_J(x,\theta_o)}{b_n}\right)t^*(\det([V_J^*(x,\theta_o)])/b_n),$$

where $V_J^*(x,\theta_o)$ is the probability limit of $V_{J,n}^*(x,\theta_o)$. With $\epsilon_n(x_i)$ and the remainders $R_{n1}(x_i)$ and $R_{n2}(x_i)$ defined in the proof of Proposition 3.1,

$$\begin{aligned}U_n(\theta_o)&=\frac{1}{\sqrt{n}}\sum_{i=1}^n[t_n^{\infty}(x_i,\theta_o)+(t_n(x_i,\theta_o)-t_n^{\infty}(x_i,\theta_o))][G_\theta'(x_i,\theta_o)+R_{n1}(x_i)]\\&\quad\times[V^{-1}(x_i,\theta_o)+R_{n3}(x_i)][\epsilon_n(x_i)+R_{n2}(x_i)]\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n^{\infty}(x_i,\theta_o)G_\theta'(x_i,\theta_o)V^{-1}(x_i,\theta_o)\epsilon_n(x_i)+R_n,\end{aligned}$$

where $R_n=R_n^{(1)}+R_n^{(2)}+\cdots+R_n^{(15)}$ is the overall remainder term consisting of fifteen terms of high order errors of nonparametric functions, where, for example

$$R_n^{(1)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n^{\infty}(x_i,\theta_o)G_{n,\theta}(x_i,\theta_o)W_n(x_i\delta_o)R_{n2}(x_i),$$

$$R_n^{(2)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n^{\infty}(x_i,\theta_o)R_{n1}(x_i)W_n(x_i\delta_o)\epsilon_i(x_i),$$

$$R_n^{(3)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n^{\infty}(x_i,\theta_o)G_\theta'(x_i,\theta_o)R_{n3}(x_i)\epsilon_n(x_i),$$

and

$$R_n^{(4)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n[t_n(x_i,\theta_o)-t_n^{\infty}(x_i,\theta_o)]G_\theta'(x_i,\theta_o)V^{-1}(x_i,\theta_o)\epsilon_n(x_i),$$

etc. The applications of Propositions A1.5 and A1.6 imply that $R_n=o_P(1)$. The U-statistic central limit theorem in Proposition A1.7 can be applied to the following term:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n^{\infty}(x_i,\theta_o)G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)\epsilon_n(x_i)=\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^n\sum_{j=1}^n\{U_{1n}(w_i,w_j)+U_{2n}(w_i,w_j)\},$$

where

$$U_{1n}(w_i,w_j)=t_n^{\infty}(x_i,\theta_o)G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)[f(z_j,\beta_o)-E(f(z,\beta_o)|x_i)]\frac{1}{h(x_i)}\frac{1}{a_n^k}K\left(\frac{x_i-x_j}{a_n}\right),$$

and

$$U_{2n}(w_i,w_j)=-t_n^{\infty}(x_i,\theta_o)G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)[f(z_j,\beta_o)-E(f(z,\beta_o)|x_i\delta_o)]\frac{1}{p(x_i\delta_o|\theta_o)}\frac{1}{b_n^m}J\left(\frac{x_i\delta_o-x_j\delta_o}{b_n}\right).$$

By the asymptotic unbiasedness of nonparametric functions, $E[U_{ln}(w_i,w_j)|w_i]$ converges to zero for $l=1,2$. Furthermore, $\lim_{n\to\infty}E[U_{1n}(w_j,w_i)|w_i]=G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)u_i$, but

$$\lim_{n\to\infty}E[U_{2n}(w_j,w_i)|w_i]=-E\left\{G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)\middle|x_i\delta_o\right\}u_i=0.$$

In conclusion,

$$\sqrt{n}(\hat{\theta}_G-\theta_o)=-D^{-1}(\theta_o)\frac{1}{\sqrt{n}}\sum_{i=1}^n G_\theta'(x_i,\theta_o)V^{-1}(x_i\delta_o)u_i+o_P(1).$$

Q.E.D.

A6: Semiparametric Instrumental Variables

By the uniform law of large numbers in Proposition A1.1, $\frac{1}{n}\sum_{i=1}^n t_n(x_i,\hat{\theta})u_{n,\theta}'(z_i,x_i,\hat{\theta})V_n^{-1}(x_i\hat{\delta})w_i\xrightarrow{P}A$, and $\frac{1}{n}\sum_{i=1}^n t_n(x_i,\hat{\theta})w_i'V_n^{-1}(x_i\hat{\delta})w_i\xrightarrow{P}\Lambda$. Denote

$$U_{n,v}(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n(x_i,\theta)w_iV_n^{-1}(x_i\delta)u_n(z_i,x_i,\theta). \tag{A6.1}$$

By the mean value theorem,

$$U_{n,v}(\hat{\theta})=U_{n,v}(\theta_o)+U_{n,v}^d(\bar{\theta})\cdot\sqrt{n}(\hat{\theta}-\theta_o),$$

where $U_{n,v}^d(\bar{\theta})=\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\theta'}\{t_n(x_i,\bar{\theta})w_iV_n^{-1}(x_i\bar{\delta})u_n(z_i,x_i,\bar{\theta})\}$, which coverges in probability to A'. As shown below that $U_{n,v}(\theta_o)$ is stochastically bounded, hence

$$\sqrt{n}(\hat{\theta}_{IV}-\theta_o)=-\{A\Lambda^{-1}A'\}^{-1}A\Lambda^{-1}U_{n,v}(\theta_o)+o_p(1).$$

Since

$$\begin{aligned}&u_n(z_i,x_i,\theta_o)\\&=u_i-\{E_n(f(z,\beta_o)|x_i\delta_o)-E(f(z,\beta_o)|x_i\delta_o)\}\\&=u_i-\frac{1}{B_J(x_i)}[A_{Jn}(x_i)-E(f(z,\beta_o)|x_i\delta_o)B_{Jn}(x_i)]\\&\quad+\frac{1}{B_{Jn}(x_i)B_J(x_i)}[B_{Jn}(x_i)-B_J(x_i)]\{[A_{Jn}(x_i)-A_J(x_i)]-E(f(z,\beta_o)|x_i\delta_o)[B_{Jn}(x_i)-B_J(x_i)]\},\end{aligned}$$

$$U_{n,v}(\theta_o)=\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n(x_i,\theta_o)w_i'V^{-1}(x_i\delta_o)\{u_i-\frac{1}{B_J(x_i)}[A_{Jn}(x_i)-E(f(z,\beta_o)|x_i\delta_o)B_{Jn}(x_i)]\}+R_n, \tag{A6.2}$$

where R_n is the remainder. As the remainder contains high order error terms, $R_n=o_p(1)$. By Proposition A1.8 in Appendix A1,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n(x_i,\theta_o)w_i'V^{-1}(x_i\delta_o)\frac{1}{B_J(x_i)}[A_{Jn}(x_i)-A_J(x_i)]\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\{E(w'|x_i\delta_o)V^{-1}(x_i\delta_o)f(z_i,\beta_o)-E[E(w'|x\delta_o)V^{-1}(x\delta_o)E(f(z,\beta_o)|x\delta_o)]\}+o_p(1),\end{aligned} \tag{A6.3}$$

and

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n t_n(x_i,\theta_o)w_i'V^{-1}(x_i\delta_o)\frac{E(f(z,\beta_o)|x_i\delta_o)}{B_J(x_i)}[B_{Jn}(x_i)-B_J(x_i)]\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\{E(w'|x_i\delta_o)V^{-1}(x_i\delta_o)E(f(z_i,\beta_o)|x_i\delta_o)-E[E(w'|x\delta_o)V^{-1}(x\delta_o)E(f(z,\beta_o)|x\delta_o)]\}+o_p(1).\end{aligned} \tag{A6.4}$$

It follows from $(A6.2)-(A6.4)$ that

$$U_{n,v}(\theta_o)=\frac{1}{\sqrt{n}}\sum_{i=1}^n(w_i-E(w|x_i\delta_o))'V^{-1}(x_i\delta_o)u_i+o_p(1). \tag{A6.5}$$

References

1. Amemiya, T. (1974), "Multivariate Regression and Simultaneous Equation Models When the Dependent Variables Are Truncated Normal," *Econometrica* 42, 999-1012.
2. Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature* 19, 1483-1536.
3. Amemiya, T. [1985], *Advanced Econometrics*, Harvard University Press, Cambridge, Mass.
4. Berkson, J. (1944), "Application of the Logistic Function to Bio-Assay," *Journal of the American Statistical Association* 39, 357-65.
5. Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics* 18, 5-46.
6. Chamberlain, G. (1986), "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics* 32: 189-218.
7. Chamberlain, G. (1992), "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60, 567-596.
8. Cosslett, S. (1987), "Efficiency Bounds for Distribution-free Estimators of the Censored Regression Models," *Econometrica* 55, 559-585.
9. Devroye, L.P. and T.J. Wagner (1980), "Distribution-free Consistency Results in Nonparametric Discrimination and Regression Function Estimation," *The Annals of Statistics* 8, 231-239.
10. Engle, R., C. Granger, J. Rice, and A. Weiss (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association* 81, 310-320.
11. Gourieroux, C., J.J. Laffont, and A. Monfort, (1980), "Disequilibrium Econometrics in a Simultaneous Equations Systems," *Econometrica* 48, 75-96.
12. Gourieroux, C., and A. Monfort, (1981), "Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models," *Journal of Econometrics* 17, 83-97.
13. Heckman, J.J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica* 42, 679-693.
14. Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics* 19, 293-325.
15. Ichimura, H. (1987), "Estimation of Single Index Models," Ph.D. thesis, Department of Economics, MIT, Cambridge, MA.
16. Ichimura, H. and L.F. Lee, (1991), "Semiparametric Estimation of Multiple Index Models: Single Equation Estimation," Ch. 1 in W.A. Barnett, J. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge U. Press, New York, NY.
17. Ito, T. (1980), "Methods of Estimation for Multi-Market Disequilibrium Models," *Econometrica* 48, 97-126.
18. Lee, L.F. (1986), "The Specification of Multi-Market Disequilibrium Econometric Models," *Journal of Econometrics* 32, 297-332.
19. Lee, L.F. (1991), "Semiparametric Instrumental Variable Estimation of Simultaneous Equation Sample Selection Models," Discussion Paper No. 263, Center for Economic Research, University of Minnesota, Minneapolis.
20. Lee, L.F. (1992), "Semiparametric Nonlinear Least-squares Estimation of Truncated Regression Models," *Econometric Theory* 8, 52-94.
21. Lee, L.F., G.S. Maddala, and R.P. Trost, (1980), "Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equations Models with Selectivity," *Econometrica* 48, 491-503.
22. Newey, W.K. (1985), "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Annales de L'INSEE* 59/60, 219-236.
23. Powell, J.L. (1987), "Semiparametric Estimation of Bivariate Latent Variable Models," Discussion Paper No. 8704, Social Systems Research Institute, U. of Wisconsin.
24. Powell, J.L. (1991), "Estimation of Monotonic Regression Models under Quantile Restrictions," Ch. 14 in W.A. Barnett, J. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge U. Press, New York, NY.
25. Powell, J.L., J.H. Stock, and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.
26. Rao, B.L.S.P. (1983), *Nonparametric Functional Estimation*. New York: Academic Press.

27. Robinson, P.M. (1988), "Root-n-Consistent Semiparametric Regression," *Econometrica* 56, 931-54.
28. Rothenberg, T. (1973), *Efficient Estimation with a Prior Information*, Yale University Press, New Haven.
29. Taylor, W.F. (1953), "Distance Functions and Regular Best Asymptotically Normal Estimates." *Annals of Mathematical Statistics* 24, 85-92.