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 procedures, which unify 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 example, 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 market 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 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 exceptions [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 [Berkeon (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, \( \hat{p}(x) \), of the response probability. The minimum distance estimator of \( \beta \) is derived by the minimization:

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

where \( m \) is the number of distinct values of \( z \) and \( \omega(z_i) \) is some weighting function. For disaggregated data with continuous explanatory variables, one has to use the method of maximum likelihood [see, e.g., Amemiya (1985)]. 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

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

where \( \omega(z_i) \) 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 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 selection 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 index, 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 framework, we consider estimation of $\theta$ in the following system of equations:

$$l^T(f(z, \theta)|z_6) = l^T(f(z, \theta)|z_6)W(z_6),$$  

(2.1)

where $z$ is a finite dimensional (row) vector of exogenous variables, $z_6$ is a finite dimensional vector consisting of endogenous and/or some exogenous variables in $z$, $f$ is a (column) vector-valued function with known functional form, $W(z_6)$ is an m-dimensional (row) vector of indices, and $\beta$ and $\delta$ can be functions of $\theta$ (a finite dimensional column vector). The latter specification captures possible parameter constraints on $\beta$ and $\delta$. The $E(f(z, \theta)|z_6)$ can be estimated by a nonparametric kernel regression. Suppose that $z$ is a $k$-dimensional vector of continuous random variables and $K$ is a kernel function on $R^k$ with a bandwidth sequence $(b_n)$. Let $(z_i, z_i), i = 1, \ldots, n$, be a random sample of size $n$. Define $A_n(z_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} f(z_i, \theta)K(\frac{z_i - z_6}{b_n})$ and $B_n(z_i) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{z_i - z_6}{b_n})$. The $E(f(z, \theta)|z_6)$ at $z = z_6$ can be estimated by

$$E_n(f(z, \theta)|z_6) = A_n(z_i, \theta)/B_n(z_i).$$  

(2.2)

Similarly, let $J$ be a m-dimensional kernel function with a bandwidth sequence $(b_n)$. Define $A_J(z_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} f(z_i, \theta)J(\frac{z_i - z_6}{b_n})$ and $B_J(z_i) = \frac{1}{n}\sum_{i=1}^{n} J(\frac{z_i - z_6}{b_n})$. The $E(f(z, \theta)|z_6)$ at $z_6 = z_6$ can be estimated by

$$E_n(f(z, \theta)|z_6) = A_n(z_i, \theta)/B_n(z_i, \theta).$$  

(2.3)

To simplify notation, let

$$G_n(z_i, \theta) = E_n(f(z, \theta)|z_6) - E_n(f(z, \theta)|z_6).$$

and $G(z_i, \theta) = E(f(z, \theta)|z_6) - Ef(z, \theta|z_6)$. A semiparametric minimum distance (SMD) estimation method can be defined as

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} G_n(z_i, \theta)W(z_i)G_n(z_i, \theta).$$  

(2.4)
Multi-market disequilibrium models have been specified in Ito (1980) and Gourieroux et al. (1989). Ito's two-market disequilibrium system is specified as

\[ y^d = \tilde{y}^d + \alpha (t - \tilde{P}) \]
\[ y^s = \tilde{y}^s + \alpha (t - \tilde{P}) \]
\[ \tilde{P} = \tilde{P} + \tilde{\beta}(y - \tilde{y}^s) \]
\[ \tilde{P} = \tilde{P} + \tilde{\beta}(y - \tilde{y}^d), \]

where \( y^d \), \( y^s \), and \( \tilde{P} \) are effective demand and supply, \( \tilde{y}^d \) and \( \tilde{y}^s \) are notional demand and supply, and the observed dependent variables are \( y = \min(y^d, y^s) \) and \( t = \min(\tilde{P}, \tilde{P}) \). The notional demand and supply equations are \( y^d = z_0 + u_1 \), \( y^s = z_0 + u_2 \), \( \tilde{P} = z_0 + v_1 \), and \( \tilde{P} = z_0 + v_2 \). 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 \( \tilde{P} < \tilde{P} \), otherwise, \( I_2 = 0 \).

The implied equations for each regime have been derived in Ito (1980). Let \( y^d = \tilde{y}^d + \beta^d P \) and \( P = \tilde{P} - \tilde{P} \).

Regime \((I_1 = I_2 = 1)\): This regime occurs if and only if \( y^d > \tilde{P} \) and \( y^s < \tilde{P} \). The observed equations are \( y = \tilde{y}^d + P \) and \( t = \tilde{P} \).

Regime \((I_1 = 1, I_2 = 0)\): This regime occurs if and only if \( y^d < \tilde{P} \) and \( y^s > \tilde{P} \). The observed equations are \( y = y^d \) and \( t = \tilde{P} \).

Regime \((I_1 = 0, I_2 = 1)\): This regime occurs if and only if \( y^d > \tilde{P} \) and \( y^s < \tilde{P} \). The observed equations are \( y = y^s \) and \( t = \tilde{P} \).

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

From these regime characterizations, the regime probabilities are functions of two indices \( z_0(\alpha - \alpha_1) \) and \( z_0(\beta_2 - \beta_1) \) because the regime inequalities involve only the variables \( y^d \) and \( P \) and some unknown parameters.\(^3\) Combining regimes, since \( I_1 = I_2 = I_1(1 - I_2) \),

\[ y = I_1 \tilde{y}^d + (1 - I_1) y^d = I_1 \tilde{y}^d + I_2 \tilde{y}^s + \alpha_1 I_1 (1 - I_2) (\tilde{P} - P) + \alpha_2 (1 - I_1) (\tilde{P} - P) \]
\[ = I_1 \tilde{y}^d + I_2 \tilde{y}^s + \alpha_1 I_1 (1 - I_2) (\tilde{y}^d - \tilde{y}^s) + \alpha_2 (I_1 - I_2) (\tilde{y}^d - \tilde{y}^s) + \frac{\alpha_2}{1 - \alpha_2} \frac{(\tilde{P} - P)}{(1 - I_1) (1 - I_2) (\tilde{y}^d - \tilde{y}^s)}, \]

where \( I_1(1 - I_2) (\tilde{P} - P) = 0 \) and \( (1 - I_1) (1 - I_2) (\tilde{P} - P) = 0 \). It follows that

\[ E(y|z) = x_0 + E(1|y|x) + \frac{\alpha_1}{1 - \alpha_2} E(I_1(1 - I_2) (\tilde{y}^d - \tilde{y}^s)|x) + \frac{\alpha_2}{1 - \alpha_2} E((I_1 - I_2) (\tilde{y}^d - \tilde{y}^s)|x). \]

Let \( z = (x_0 - \alpha_1, \alpha_1 (1 - I_2) (\tilde{y}^d - \tilde{y}^s)). \) By index restrictions, \( E(1|y|x) = E(1|y_1|x_0) = E(1|y_1|x_0) \) \( E(1|y_2|x_0) E(1|y_1|x_0) = E(1|y_2|x) \) \( E(1|y_1|x) \) \( E(1|y_2|x) \). Eliminating these unknown functions,

\[ E(y|x, \alpha_1, \alpha_2) = E(y|x, \alpha_1, \alpha_2) = E(y|x, \alpha_1, \alpha_2). \]

Similarly,

\[ E(I_1 I_2) = E(I_1 I_2|x, \alpha_1, \alpha_2). \]

The (2.13) and (2.14) can be used for the SMD estimation. The observed \( \alpha_1, \alpha_2, \beta_1, \beta_2 \) 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.

\(^3\) The listing of regimes above demonstrates that the relevant indices in estimation are the two indices \( z_0(\alpha - \alpha_1) \) and \( z_0(\beta_2 - \beta_1) \) instead of the indices \( z_0, z_0, z_0, z_0, \) and \( z_0 \).

\(^4\) There are symmetry properties in (2.13) and (2.14). (The (2.13) implies \( E(y|x) = E(y|x) E(\alpha_1 - \alpha_1, \alpha_1 - \alpha_1) \) and vice versa. The (2.14) is equivalent to \( E(I_1 I_2|x) = E(I_1 I_2|x) E(\alpha_1 - \alpha_1, \alpha_1 - \alpha_1) \).)
and
\[
E(I_1 + I_2x) = E(I_1 + I_2x_6B).
\]  
(2.21)
As $I_1 + I_2 = 1$ if and only if $y_0 > 0$, the above moment equations provide information for the estimation of the structural equation of $y_0$. Similarly, the regimes (i) and (iii) can be combined for the estimation of the structural equation of $y_0$. The corresponding moment equations are
\[
E(y_n - y_0I_1, I_1 + I_2 = 1) = E(y_n - y_0I_2B).
\]  
(2.22)
and
\[
E(I_1 + I_2x) = E(I_1 + I_2x_6B).
\]  
(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 estimation 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 replacing the unknown function $f$ by a nonparametric function in an intermediate step and the final estimation is made 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 $E(y|x) = x + r(x)E(y|x) + r(x)$. The unknown $r$ is then eliminated by taking the difference to have the estimating moment equation $E(y - x|x) = E(y - x|x)$. 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 approaches 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 functions where the dependent variable $y$ cannot 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 $y$. The SNLS method $\min_n \sum y_i - y_i^\lambda / \lambda - \lambda - E_0(y - y_0^\lambda)E(y_i - y_i^\lambda) = \lambda - E_0(y - y_0^\lambda)E(y_i - y_i^\lambda)$, which can be shown to be consistent. As a remark, we note that Powell (1991) has considered estimation of limited dependent variable models 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:
\[
y = x\beta + r(x) + v,
\]  
(2.24)
where $r(\cdot)$ is an unknown function of an index $x$. The semiparametric nonlinear least squares (SNLS) method introduced in Ichimura and Lee (1991) is
\[
\min_n \sum_{i=1}^n (y_i - r(x_i)) - E_n(y - x\theta(x)) + r(x_i))^2,
\]  
where $E_n(y - x\theta(x))$ provides a nonparametric estimate of $r(x)$. This approach is motivated by replacing the unknown function $r$ 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 $E[y|x] = x + r(x)$ and $E[y|x] = E[y|x] + r(x)$. The unknown $r$ is then eliminated by taking the difference to have the estimating moment equation $E[y - x|x] = E[y - x|x]$. 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 functions where the dependent variable $y$ cannot 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 $y$. The SNLS method $\min_n \sum y_i - y_i^\lambda / \lambda - \lambda - E_0(y - y_0^\lambda)E(y_i - y_i^\lambda) = \lambda - E_0(y - y_0^\lambda)E(y_i - y_i^\lambda)$, which can be shown to be consistent. As a remark, we note that Powell (1991) has considered estimation of limited dependent variable models 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.

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 $\theta$ denote the SMD estimate from (2.4). Let $\beta$ be the true parameter vector, and let $\beta_0$ and $\delta_0$ denote, respectively, $\beta$ and $\delta$ evaluated at $\theta = \theta_0$. For any possible value of $\theta_0$, $E[|x|\theta]$ denotes $E[|x|\theta(\theta_0)]$, a conditional expectation conditional on $x|\theta$ for a given value $\theta_0$, for simplicity. Furthermore, $E[|x|\theta]$ stands for $E[|x|\theta]$ evaluated at $x = x_0$, and $E[|x|\theta]$ for $E[|x|\theta]$ evaluated at $x = x_0$. All the expectation operations are taken with respect to the true data generating process at $\theta_0$. The following propositions show that, under proper identification conditions and regularity conditions, $\theta$ is $\sqrt{n}$-consistent. To simplify notation, denote
\[
G_\theta(x, \theta_0) = \frac{\partial E(f(x, \theta_0)|x)}{\partial \theta} - \frac{\partial E(f(x, \theta_0)|x)}{\partial \theta}.
\]
and
\[
E(u) = f(x, \theta_0) - E(f(x, \theta_0)|x)
\]
to this throughout the article. $u$ is the disturbance of the model. As a convention, for any function, say, $f(x, \theta)$ with a subscript $\theta$ will denote its gradient $\frac{\partial f(x, \theta)}{\partial \theta}$ with respect to $\theta$.

Proposition 3.1 Under Assumptions 1-6 and the identification condition that, for any $\theta \neq \theta_0$, $E[|x|\theta] \neq 0$ with positive probability on $X$ where $X = \{x|\theta(\theta_0) \neq 0\}$, $\theta_0$ is a consistent estimator of $\theta_0$ and
\[
\sqrt{n}(\theta_0 - \theta) = -E(G(x, \theta_0)|x)\theta_0^{-1}\frac{1}{\sqrt{n}} \sum_{i=1}^n [C(x_i) - C(x_i)|x]\theta_0^{-1}(1).
\]
(3.1)
where $C(x_i) = G_\theta(x_i, \theta_0) = E[G_\theta(x_i, \theta_0)|x]$. Consequently, $\sqrt{n}(\theta_0 - \theta) \to N(0, \Omega)$, where
\[
\Omega = E[G_\theta(x, \theta_0)|x]G_\theta(x, \theta_0)^{-1}E(G_\theta(x, \theta_0)|x)G_\theta(x, \theta_0)^{-1}
\]
and $\Sigma = E[C(x) - C(x)|x] \Sigma(x) = E[C(x) - C(x)|x] \Sigma(x)$. We note that [Ichimura and Lee (1991), Lemma 4]
\[
G_{\theta}(x, \theta_0) = E(f(x, \theta_0)|x) - E(f(x, \theta_0)|x)\theta_0^{-1}
\]
\[
- \frac{\partial^2 E(f(x, \theta_0)|x)}{\partial \theta^2} - E\left(\frac{\partial^2 E(f(x, \theta_0)|x)}{\partial \theta^2}ight)
\]
(3.3)
where $\nabla E(f(x, \theta_0)|x)$ denotes the gradient of $E(f(x)|x)$ with respect to its argument vector $x$.

It is interesting to point out an implication of Proposition 3.1 for the special case $\theta(\beta, \delta) = \beta - \delta$ for $\delta_0$. For this case, $G_{\theta}(x, \theta_0) = E[y|x] - E(f(x, \theta_0)|x)$, and the model is equivalent to
\[
y - f(x, \theta_0) + f(x, \theta_0) + E(f(x, \theta_0)|x)\theta_0 + u.
\]
(3.4)
An alternative estimation approach for this model can be
\[
\min_{\theta} \sum_{i=1}^{n} \left[ y_i - f(z_i, \theta) - E_{\theta}(f(z_i, \beta)|z_i) - E_{\theta}(f(z_i, \beta)|x_i) \right]^2 W(z_i) \\
\times \left[ y_i - f(z_i, \theta) - E_{\theta}(f(z_i, \beta)|z_i) - E_{\theta}(f(z_i, \beta)|x_i) \right].
\] (3.5)
For the model where the term \( f(z_i, \beta) \) 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 \( \theta \) from (3.5) is consistent and asymptotically normal. For any possible value \( \theta \), denote
\[ r(\theta) = f(z_i, \theta) + E_{\theta}(f(z_i, \beta)|z_i) + E_{\theta}(f(z_i, \beta)|x_i), \]
and \( r_\theta(z_i, \theta) \) be \( r(\theta) \). The \( \hat{\theta}_{NL} \) has the same asymptotic distribution as the SMD estimator \( \hat{\theta}_S \) 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 \( \theta \neq \theta_0 \), \( r(z, \theta) \neq r(z, \theta_0) \) with positive probability on \( X \), the SNLS estimator \( \hat{\theta}_{NL} \) from (3.5) is consistent, and
\[
\sqrt{n}(\hat{\theta}_{NL} - \theta_0) = \left\{ E[r(z, \theta_0)W(z)r(z, \theta_0)] \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [r(z_i, \theta_0)W(z_i)r(z_i, \theta_0)]u_i + o_p(1).
\] (3.6)
The limiting distribution of \( \hat{\theta}_{NL} \) for (3.3) is the same as the limiting distribution of \( \hat{\theta}_S \) in Proposition 3.1 because \( E[f(z_i, \theta)|x_i] = f(z, \theta) \).

### 4. Generalized Semiparametric Estimation

We see from Proposition 3.1 that the asymptotic distribution of SMD estimator depends on the weighting function \( W(z) \). If \( W(z) \) is chosen as functions of \( z \), the asymptotic distribution for the corresponding SMD estimator has simpler expression because \( C_0(z), C_1(z), C_2(z) \) is zero as \( E[C_0(z)|z_i]\) and \( E[C_1(z)|z_i] \) and \( E[C_2(z)|z_i] \) are zero, 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 \( u \) on \( x \) is general function of \( x \) (not a function of \( z \)), such a weighting function does not necessarily improve efficiency because the presence of the additional term \( C_3(z) \), which captures the effect of using nonparametric functions to replace the correct regression functions as an intermediate step in estimation. In many microeconometric models such as the simultaneous equations models introduced before, the conditional variances may depend on \( z \) only through their dependence on \( x \). With such heteroscedasticity variance functions, the optimal weighting function is indeed the inverse of the variance function because the complicated \( C_3(z) \) can be zero. However, if the (limiting) weighting functions depend only on \( z \), \( C_3(z) \) will be zero, that can be regarded as an orthogonality property of the estimation procedure. Thus for models where heteroscedasticity variance functions are functions of \( x, \theta \) more efficient estimator may be derived by some generalized semiparametric procedures.

Let \( \hat{\theta}_S \) be a \( \phi \)-consistant estimator of \( \theta \), for example, \( \hat{\theta}_S \) from the previous SMD estimation. Define a nonparametric variance function:
\[ V(z, \theta) = \frac{\partial E_u[f(z, \theta)]}{\partial \theta} - E_u[f(z, \theta)|x_i]E_u[f(z, \theta)|z_i] \] (4.1)
where \( \hat{\theta} \) and \( \theta \) are \( \phi(\theta) \) and \( \phi(\theta) \) evaluated at \( \hat{\theta} \). This nonparametric variance function is an estimate of the variance function \( V(z, \theta) \) where \( V(z, \theta) = \text{Var}(f(z, \theta)|x_i) \).
\[ V(z, \theta) = \frac{\partial E_u[f(z, \theta)]}{\partial \beta} - E_u[f(z, \theta)|x_i]E_u[f(z, \theta)|z_i] \] (4.2)
is the variance function \( f(z, \theta) \) conditional on \( x_i \). This is so no matter whether \( \text{Var}(f(z, \theta)|x_i) \) equals \( \text{Var}(f(z, \theta)|x_i) \) or not. The feasible weighting matrix \( W_u(z, \theta) \) will be \( W_u(z, \theta) = V^{-1}(z, \theta) \). Denote
\[ C_{3,v}(z, \theta) = \frac{\partial E_u[f(z, \theta)]}{\partial \theta} - \frac{\partial E_u[f(z, \theta)]}{\partial \beta} \]

Our suggested generalized semiparametric minimum distance (GSMĐ) estimator is a two-step estimator based on a generalized semiparametric nonlinear least squares procedure:
\[
\hat{\theta}_G = \left\{ \sum_{i=1}^{n} t_i(x_i, \hat{\theta})G_{3,v}(x_i, \hat{\theta})W_u(x_i, \hat{\theta}) \right\}^{-1} \times \sum_{i=1}^{n} t_i(x_i, \hat{\theta})G_{3,v}(x_i, \hat{\theta})W_u(x_i, \hat{\theta}) \left( G_{3,v}(x_i, \hat{\theta}) + G_{3,v}(x_i, \hat{\theta}) \right),
\] (4.3)
where \( t_i(x, \hat{\theta}) \) is a trimming function which goes to 1 as \( n \) goes to infinity. The function of \( t_i(x, \hat{\theta}) \) 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 GSMĐ estimator is motivated by the GSMĐ method:
\[ \min_{\theta} \sum_{i=1}^{n} t_i(x_i, \theta)G_{3,v}(x_i, \theta)W_u(x_i, \theta) \left( G_{3,v}(x_i, \theta) + G_{3,v}(x_i, \theta) \right). \]
However, for parameter value of \( \theta \) not in the neighborhood of \( \theta \), the trimming by \( t_i(x, \hat{\theta}) \) does not control properly the tails of nonparametric regressions and variance functions. However, the two-step method seems simpler.

**Proposition 4.1** Under Assumptions 1-5 and 7 and the identification condition that, for any \( \theta \neq \theta_0 \), \( E[f(z, \beta)|z_i] = E[f(z, \beta)|x_i] \neq 0 \) with positive probability, \( \hat{\theta}_C \) is a consistent, asymptotically normal estimator with
\[
\sqrt{n}(\hat{\theta}_C - \theta_0) \rightarrow N(0, \Omega_2),
\] (4.4)
where \( \Omega_2 = D^{-1}(\theta_0)C_0D^{-1}(\theta_0) \) with \( D(x_i) = E[C_3(x_i)G^{(-1)}(x_i)G(x_i)], \) and
\[ C_0 = E[C_3(x_i)G^{(-1)}(x_i)\text{Var}(u)^{-1}G(x_i)] \]
Furthermore, if \( \text{Var}(u) \neq 0 \), then \( D(x_i) = D^{-1}(\theta_0) \).
One can compare the limiting variance matrix \( \Omega_2 \) with the variance matrix \( \Omega_1 \) in Proposition 3.1 for the model where \( \text{Var}(u) = 0 \). For such a model
\[ \Omega_1 = E[C_0(x) - C_3(x)\text{Var}(u)C_0(x) - C_3(x)\text{Var}(u)] \]
Since \( E[C_0(x)|x_i] = 0 \) by (3.3),
\[ E[G^2(x, \theta)\text{Var}(u)](x, \theta)) \] (4.5)

Comparing this \( \Omega_1 \) with \( \Omega_2 \) by the generalized Schwartz inequality, we see that \( \Omega_1 \geq \Omega_2 \). Hence \( \hat{\theta}_C \) is asymptotically efficient relative to \( \hat{\theta}_G \) as it uses the optimal weighting.
methods. For the case that \( \text{Var}(f(t, \theta)) = \text{Var}(f(t, \theta)|z_i) \), there is indeed possible efficiency gain in the system GSMD method. To see that, let \( f \) consist of two components such that \( f(t, \beta) = (f_j(t, \beta), f_j(t, \beta)) \). Suppose that \( \theta_i \) is the parameter vector in the first set of moment equations,

\[
E[f(t, \beta)] = E[f(t, \beta)]|z_i, \theta_i],
\]

and \( \phi_i \) is the parameter vector in the remaining moment equations

\[
E[f_i(z, \beta)] = E[f_i(z, \beta)]|z_i, \theta_i].
\]

Consider first the case that \( \theta_1 \) and \( \theta_2 \) are distinct parameters. The \( \theta_i \) can be estimated by using only the moment equations (4.6) of \( f_1 \). Let \( \theta_{1,i} \) be the corresponding GSMD estimator using \( f_1 \) alone. Let \( \theta_{2,i} \) be the corresponding GSMD estimator using both \( f_1 \) and \( f_2 \). Let \( V(z_i) = \begin{pmatrix} V_{11}(z_i) & V_{12}(z_i) \\ V_{21}(z_i) & V_{22}(z_i) \end{pmatrix} \) be the corresponding partitioned variance matrix, and let \( V^{-1}(z_i) = \begin{pmatrix} V_{11}(z_i)^{-1} & V_{12}(z_i)^{-1} \\ V_{21}(z_i)^{-1} & V_{22}(z_i)^{-1} \end{pmatrix} \) be its inverse matrix. Let \( A_i(z) = E[f_i(z, \beta_i)|z_i] - \frac{\partial f_i(z, \beta_i)}{\partial \theta_i} \), and \( A_i(z) = E[f_i(z, \beta_i)|z_i] - \frac{\partial f_i(z, \beta_i)}{\partial \theta_i} \). From Proposition 4.1, the limiting variance matrix of \( \sqrt{n} (\hat{\theta}_i - \theta_i) \) is \( \Omega_{\theta,i} = \left[E[A_i(z)V^{-1}(z_i)A_i(z)^{T}] \right] \), and the limiting variance of \( \sqrt{n} (\hat{\theta}_i - \theta_i) \) is \( \Omega_{\theta,i} = \left[E[A_i(z)V^{-1}(z_i)A_i(z)^{T}] \right] \), where

\[
\mathbf{1}_{L} = \left\{ \mathbf{1}_{L} \right\} \text{ where } L = 1, 2.
\]

This shows that \( \hat{\theta}_{1,i} \) is efficient relative to \( \hat{\theta}_{2,i} \). For the case that \( \theta_1 \) and \( \theta_2 \) contain some common parameters, the system GSMD estimator which incorporates common parameter constraints will be more efficient than the GSMD estimator of \( \theta \) without imposing the constraints. To see that for our semiparametric estimators, let \( \theta_1 \) be the parameter vector of \( \theta \) without imposing constraints. Then for \( \theta \) and \( \phi \) be parameters of \( \Lambda \),

\[
\frac{\partial E[f(t, \beta)]}{\partial \Lambda} = \left( \frac{\partial \theta_1}{\partial \Lambda}, \frac{\partial \phi}{\partial \Lambda} \right) A(z).
\]

where \( A(z) = \begin{pmatrix} A_1(z) \\ A_2(z) \end{pmatrix} \). Let \( \varphi_{\Lambda} \) be the generalized semiparametric estimator of \( \Lambda \), and let \( \varphi_{\Lambda} \) be the corresponding constrained estimator of \( \theta \) and \( \phi \). By Rao's delta method, the asymptotic variance matrix of \( \left( \hat{\theta}_{1,i}, \hat{\phi}_{2,i} \right) \) is

\[
\left( \frac{\partial \theta_1}{\partial \Lambda}, \frac{\partial \phi}{\partial \Lambda} \right) \left( \begin{pmatrix} \frac{\partial \theta_1}{\partial \Lambda} & \frac{\partial \phi}{\partial \Lambda} \end{pmatrix} E[A(z)V^{-1}(z_i)A(z)^{T}] \left( \begin{pmatrix} \frac{\partial \theta_1}{\partial \Lambda} & \frac{\partial \phi}{\partial \Lambda} \end{pmatrix} \right)^{T} \right)^{-1} \left( \begin{pmatrix} \frac{\partial \theta_1}{\partial \Lambda} & \frac{\partial \phi}{\partial \Lambda} \end{pmatrix} \right)^{T} \Omega_{\Lambda} \left( \begin{pmatrix} \frac{\partial \theta_1}{\partial \Lambda} & \frac{\partial \phi}{\partial \Lambda} \end{pmatrix} \right) \right).
\]

The asymptotic variance of the unconstrained estimator \( (\hat{\theta}_{1,i}, \hat{\phi}_{2,i}) \) is \( (E[A(z)V^{-1}(z_i)A(z)^{T}] \right) \). By the Schwartz inequality, the asymptotic variance of \( (\hat{\theta}_{1,i}, \hat{\phi}_{2,i}) \) is apparently larger than the asymptotic variance of \( (\hat{\theta}_{1,i}, \hat{\phi}_{2,i}) \).
where the moment conditions in Assumptions 2, 4, 6, and 7 need, of course, to be modified accordingly to conditional moments conditional on the additional regime occurrence in addition to \( x \) and \( \delta \).

This model is a special index model with \( z_2 = z_2 \). In our estimation framework, \( f(x, \theta(\delta)) = y - x_1 \theta \) and \( G(x, \theta, \delta) = [x_2 - E(x_i|x_2)]^T/\sigma(z_2) \). With \( W(x) \) in Proposition 3.1 being an identity matrix (and trimming function going to unity everywhere) for the homoskedastic variance model, \( \Sigma = \sigma^2 \text{Var}(x_i|x_2)'/\sigma(z_2) \) and

\[
\Omega = \sigma^2 \text{Var}(x_i|x_2)'/\sigma(z_2) \text{Var}(x_i|x_2)',
\]

attains the variance bound of Chamberlain. For the model where \( \sigma^2(z_2) = \sigma^2(z_2) \), as \( V(z_2) = \text{Var}(f(x_i|x_2|x_2)) = \text{Var}(y_i|x_2) \), Proposition 4.1 implies that

\[
\Omega = \left( E \left( \frac{x_i|x_2'}{\sigma(z_2)} \right) - E \left( \frac{x_i|x_2'}{\sigma(z_2)} \right) \right) \text{Var}(x_i|x_2) \text{Var}(x_i|x_2)'^{-1},
\]

which attains the variance bound of Chamberlain for the index variance model. However, for the general case where \( \sigma^2(z_2) \) is a general function of \( z_2 \), neither \( \sigma(z_2) \) nor \( \sigma(z_2) \), attain the efficiency bound. Apparently our estimation methods have utilized only index restrictions for estimation.

This model is a special index model with \( z_2 = z_2 \). In our estimation framework, \( f(x, \theta(\delta)) = y - x_1 \theta \) and \( G(x, \theta, \delta) = [x_2 - E(x_i|x_2)]^T/\sigma(z_2) \). With \( W(x) \) in Proposition 3.1 being an identity matrix (and trimming function going to unity everywhere) for the homoskedastic variance model, \( \Sigma = \sigma^2 \text{Var}(x_i|x_2)'/\sigma(z_2) \) and

\[
\Omega = \sigma^2 \text{Var}(x_i|x_2)'/\sigma(z_2) \text{Var}(x_i|x_2)',
\]

attains the variance bound of Chamberlain. For the model where \( \sigma^2(z_2) = \sigma^2(z_2) \), as \( V(z_2) = \text{Var}(f(x_i|x_2|x_2)) = \text{Var}(y_i|x_2) \), Proposition 4.1 implies that

\[
\Omega = \left( E \left( \frac{x_i|x_2'}{\sigma(z_2)} \right) - E \left( \frac{x_i|x_2'}{\sigma(z_2)} \right) \right) \text{Var}(x_i|x_2) \text{Var}(x_i|x_2)'^{-1},
\]

which attains the variance bound of Chamberlain for the index variance model. However, for the general case where \( \sigma^2(z_2) \) is a general function of \( z_2 \), neither \( \sigma(z_2) \) nor \( \sigma(z_2) \), attain the efficiency bound. Apparently our estimation methods have utilized only index restrictions for estimation.

5. Sample Selection Model with Known Index

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

\[
y^* = z_1 \theta + z_2 \theta + \epsilon,
\]

and

\[
j = \begin{cases} 1, & \text{if } (y_i, z_i) \geq 0, \\ 0, & \text{otherwise,} \end{cases}
\]

where \( n = (x, z) \) and \( p = (y, j) \), where \( y_i = (y_i, j_i) \), are observed; so \( y^* \) is observed only if \( j_i = 1 \). The function \( g_j \) depends on \( z_i \) only via \( z_i \) but is otherwise unrestricted. The disturbances \( \epsilon \) and \( u_i \) satisfy that \( E(\epsilon|x,v) = E(\epsilon|x) = u_i \) are independent of \( z_i \) conditional on \( x_i \). This model is thus a sample selection model with index restriction where the indices are \( z_1 = z_2 \). This model implies that \( E(y_i|x, j = 1) = x_i \theta + \epsilon(z_i) \), where \( \epsilon(z_i) = \epsilon(z_1) + \epsilon(z_2) \).

Chamberlain shows that, for this model, the efficiency bound for \( \theta \) is the same as one of the above \( \rho \) functions. Let \( \sigma^2(z_2) \) denote \( \text{Var}(y_i|x_2) \). The efficiency bound for \( \theta \) is

\[
J = E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) - E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) / E \left( \frac{x_i|x_2}{\sigma(z_2)} \right).
\]

If the model is homoskedastic with \( \sigma^2(z_2) = \sigma^2(z_2) \) constant, then

\[
J = \sigma^2 E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) - E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) / E \left( \frac{x_i|x_2}{\sigma(z_2)} \right).
\]

(5.4)

If the variance satisfies also an index restriction with \( \sigma^2(z_2) = \sigma^2(z_2) \), then

\[
J = E \left[ \sigma^2 E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) - E \left( \frac{x_i|x_2}{\sigma(z_2)} \right) \right] / E \left( \frac{x_i|x_2}{\sigma(z_2)} \right).
\]

(5.5)

The moment conditions in Assumptions 2, 4, 6, and 7 need, of course, to be modified accordingly to conditional moment restrictions in additional to \( \epsilon \).

---

3 The moment conditions in Assumptions 2, 4, 6, and 7 need, of course, to be modified accordingly to conditional moment restrictions in additional to \( \epsilon \).
For the semiparametric binary choice model, which is a generalization of the linear simultaneous equation model, it is unknown what is the efficient bound for the conditional moment equation when it is included in estimation (as in a seemingly unrelated regression framework). This point has been noted in Chamberlain (1982) in his efficiency analysis. The previous section provide an example.

Proposition 4.1 implies that as long as \( \theta'^2 \) does not vanish with probability one, the GSMD estimator of \( \theta \) (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. For the linear simultaneous equation sample selection model, Powell (1987) has derived a two-step semiparametric instrumental variable method for the estimation of the outcome equation conditional on a given \( \gamma \)-consistent estimate of the first stage parameter (the parameter 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 methods. 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 \( g_1(y_t, x_t, \delta_t) = c_t \) and \( g_2(y_t, x_t, \delta_t) = c_t \) be, respectively, the potential outcome equations for regime 1 (\( I = 1 \)) and regime 2 (\( I = 0 \)). The implied conditional moment equations are

\[
E(g_1(y_t, x_t, \delta_t) | I = 1, x_t) = E(g_1(y_t, x_t) | I = 1, x_t)
\]

and

\[
E(g_2(y_t, x_t, \delta_t) | I = 0, x_t) = E(g_2(y_t, x_t) | I = 0, x_t).
\]

Let \( g_1(x_t) = E(g_1(y_t, x_t, \delta_t) | I = 1, x_t) \) and \( g_2(x_t) = E(g_2(y_t, x_t, \delta_t) | I = 0, x_t) \). Let \( \sigma_2(x_t) \) denote the conditional variance of \( g_2(x_t, \delta_t) \) conditional on \( I = 1 \) and \( x_t \), and let \( \bar{\sigma}_2(x_t) \) denote the conditional variance of \( g_2(x_t, \delta_t) \) conditional on \( I = 0 \) and \( x_t \). With (5.14) and (5.15) and \( E(I | x_t) = E(I | 1) \delta_t \), the limiting variance of \( \hat{\theta}_c \) follows from Corollary 4.1 which is

\[
\left[ \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right] \left( E[I(y_t, x_t)] \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right)^{-1}\]

where

\[
\sigma^2[I(y_t, x_t)] = \left( \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right) \left( E[I(y_t, x_t)] \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right)^{-1}.
\]

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.

5. Sample Selection Model with Unknown Index

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

\[
g_t(x_t, \delta_t) = \epsilon_t
\]

is a potential outcome equation before selection. The estimation of \( \theta \) can be based on the equations \( E(g_t(x_t, \delta_t) | I = 1, x_t) = E(g_t(x_t, \delta_t) | I = 1, x_t) \) and \( E(I | x_t) = E(I | x_t) \). Let \( \delta_t(x_t, \delta_t) = E(g_t(x_t, \delta_t) | I = 1, x_t) - E(g_t(x_t, \delta_t) | I = 1, x_t) + \frac{\partial E[I(x_t)]}{\partial \delta_t} \frac{\partial g_t(x_t, \delta_t)}{\partial \delta_t} \), where \( \partial g_t(x_t, \delta_t) \) denotes the gradient of \( g_t(x_t, \delta_t) \) with respect to \( \delta_t(x_t) \). The limiting variance \( \hat{\sigma}_t \) of \( \hat{\sigma}_t(x_t) \) follows from Corollary 4.1:

\[
\left( E[I(y_t, x_t)] \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right) \left( E[I(y_t, x_t)] \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \right)^{-1}\]

For the semiparametric binary choice model, the semiparametric efficiency bound has been derived in Cosslett (1987) and Chamberlain (1986), which is

\[
E[I(y_t, x_t)] \frac{\partial g_1(x_t)}{\partial \delta_t} + \frac{E[I(y_t, x_t)]}{\sigma^2[I(y_t, x_t)]} \frac{\partial g_2(x_t)}{\partial \delta_t} \]

Comparing the inverse of the efficiency bound and the limiting variance of the GSMD estimator, the limiting variance matrix of the GSMD estimator is asymptotically more efficient. Even for the case that \( \theta' = (\theta', \delta') \), as long as \( \bar{\sigma}_2(x_t) \) does not vanish with probability one, the GSMD estimator of \( \theta' \) (with proper normalization) will be efficient relative to semiparametric estimators derived for estimating only the binary choice equation.
matrix of instrumental variables (functions of x). A semiparametric instrumental variable (SIV) estimator taking into account possible variance structure of w can be defined as

\[ \delta_{IV} = \left\{ \sum_{i=1}^{n} t_i(x_i, \theta) w'_i(x_i, \theta) V_i^{-1}(x_i, \theta) w_i(x_i, \theta) \right\}^{-1} \left( \sum_{i=1}^{n} t_i(x_i, \theta) w'_i(x_i, \theta) V_i^{-1}(x_i, \theta) w_i(x_i, \theta) \right) \]

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

\[ \min_{s} \sum_{i=1}^{n} t_i(x_i, \theta) w'_i(x_i, \theta) V_i^{-1}(x_i, \theta) w_i(x_i, \theta, \theta) \]

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 \( A = E(w'V^{-1}(x,z,w)) \) and \( A = E(w'V^{-1}(x,z,w)) \).

It follows from the \( U \)-statistic theory that

\[ \sqrt{n} (\delta_{IV} - \delta_{u}) = - (AA^{-1})^{-1} AA^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_i(x_i, \theta) w'_i(x_i, \theta) V_i^{-1}(x_i, \theta) w_i(x_i, \theta, \theta) + o_p(1). \]

Hence \( \sqrt{n} (\delta_{IV} - \delta_{u}) \rightarrow^d N(0, \Omega_{IV}) \) where

\[ \Omega_{IV} = \left\{ AA^{-1}A' \right\}^{-1} E\left[ \left( w - E(w|x,z) \right)' V^{-1}(x,z,w) \text{Var}(w|x,z) V^{-1}(x,z,w) \right] \left( w - E(w|x,z) \right) A' \left\{ AA^{-1}A' \right\}^{-1}. \]

For the model where \( \text{Var}(w|x) = \text{Var}(w|x,z) \), \( \Omega_{IV} \) becomes

\[ \Omega_{IV} = \left\{ AA^{-1}A' \right\}^{-1} A \left\{ AA^{-1}A' \right\}^{-1}, \]

where \( A = E\left[ w - E(w|x,z) \right]' V^{-1}(x,z,w) \left( w - E(w|x,z) \right) \). From this result, it is interesting to note that better instrumental variables are \( w_{IV} \) instead of \( w \), even if \( E(w|x,z) \neq 0 \). Using \( w = E(w|x) \) as instrumental variables, the limiting variance matrix of the corresponding SIV will be

\[ \Omega_{IV,v} = \left\{ AA^{-1}A' \right\}^{-1}. \]

Because \( A = E\left[ w_{IV}'(x,z,w) V^{-1}(x,z,w) \right] = E\left[ w_{IV}'(x,z,w) V^{-1}(x,z,w) \right] \) by (3.3). By the generalized Schwartz inequality, \( \Omega_{IV,v} \geq \Omega_{IV} \).

It remains to compare the GSMD estimator with any SIV estimator. Since \( w \) is a function of \( x \) and \( \Omega_{IV} = \Omega_{IV,v} \),

\[ E\left[ w_{IV}'(x,z,w) V^{-1}(x,z,w) \right] \left( w - E(w|x,z) \right) = E\left[ E\left[ w_{IV}'(x,z,w) V^{-1}(x,z,w) \right] \right] = E\left[ E\left( w_{IV}'(x,z,w) V^{-1}(x,z,w) \right) \right] \]

In fact, it follows that \( \Omega_{IV,v} \) can be rewritten as

\[ \Omega_{IV,v} = \left\{ A \right\}^{-1} A \left\{ A^{-1}A' \right\}^{-1} \left\{ A \right\}^{-1} A \left\{ A^{-1}A' \right\}^{-1}. \]

Comparing this variance matrix with \( \Omega_{IV} \) in Proposition 4.1 for the model \( \text{Var}(w|x) = \text{Var}(w|x,z) \) by the generalized Cauchy-Schwartz inequality, it is apparent that \( \Omega_{IV,v} \geq \Omega_{IV} \). To show that the GSMD is a SIV estimator, take

\[ w = E(f(x,z)|w) - \nabla E(f(x,z)|w) \left( \frac{\partial^2 (\delta(w|x,z))}{\partial \theta^2} \right) \]
Appendix

A1: Some Useful Propositions

The following propositions will be used repeatedly in subsequent proofs of nonsymmetry 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 summarize 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 \(h(x, z, \omega, \alpha_n)\) takes the form \(h(x, z, \omega, \alpha_n) = \frac{1}{2}h_1(x, z)h_2(\omega, \alpha_n)\), where \(\alpha_n = O(n)\), \(p > 0\), \(d > 0\), \(\theta \in \Theta\), and \(z(x, \theta)\) is a finite dimensional vector-valued function. Suppose that the following conditions are satisfied:

1. \(\Theta\) is a compact subset of a finite dimensional Euclidean space.
2. The function \(h_1(x, \theta)\) is differentiable with respect to \(\theta\). The \(r\)th order moment, where \(r \geq 2\), of \(h_2(\omega, \alpha_n)\) is finite. The first moment of \(h_2(\omega, \alpha_n)\) exists and is finite.
3. \(h_2(\omega, \alpha_n) \leq c\) for some constant \(c\).
4. \(h_2(\omega, \alpha_n)\) is uniformly continuous in \(\theta\) in \(\Theta\), for some \(d\).
5. The functions \(h_1(x, \theta), z(x, \theta), \) and \(z(x, \theta)\) satisfy the bounded Lipschitz condition of order 1 with respect to \(\theta \) and \(x\).

If \(\lim_{n \to \infty} \frac{\sum_{i=1}^{n} h_1(x, z, \omega, \alpha_n) - E[h_1(x, z, \omega, \alpha_n)]}{\sqrt{n}} = 0\), then \(\lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_1(x, z, \omega, \alpha_n) - E[h_1(x, z, \omega, \alpha_n)] = 0\). Further, in addition to the above conditions, if \(E[h_1(x, z, \omega, \alpha_n)]\) converges to a limit function \(h_1(\theta)\) uniformly in \(\Theta\), then \(\lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_1(x, z, \omega, \alpha_n) - h_1(\theta) = 0\).

Proof: This theorem generalizes slightly the uniform law of large numbers 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. In the original 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 \(\sup_{\Theta} \|h(x, z, \theta)\|\) 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.

Proposition A1.2 Let \(K(x)\) be a function on \(\mathbb{R}^d\) with a bounded support \(D\) such that \(\int_{D} K(x)dx > 0\). Let \(\{\xi_i\} \) be a continuous \(d\)-dimensional random vector. Suppose that \(E[h(x, z, \theta)|\xi_i, \theta, g(\theta)]\) is the density function of \(\xi_i, \theta\), uniformly continuous in \(\theta\), uniformly in \(\xi_i\). Then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} K \left( \frac{h(x_i, z_i, \theta) - E[h(x_i, z_i, \theta)\xi_i, \theta, g(\theta)\xi_i, \theta]}{\alpha_n} \right) = 0
\]

Furthermore, if \(K(x)\) is a function with zero moments up to the order \(s\), i.e., \(\int_{D} x_k K(x)dx = 0\), for all \(k = 1, \ldots, d\), then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} K \left( \frac{h(x_i, z_i, \theta) - E[h(x_i, z_i, \theta)\xi_i, \theta, g(\theta)\xi_i, \theta]}{\alpha_n} \right) = 0
\]

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.

Proposition A1.3 Let \(K(x)\) be a function on \(\mathbb{R}^d\) with a bounded support \(D\) such that \(K(x)\) goes to zero at the boundary of \(D\) and its gradient \(\nabla K(x)\) is bounded. Suppose that \(\frac{1}{n} \sum_{i=1}^{n} E[h(x_i, z_i, \theta)|\xi_i, \theta, g(\theta)]\) where \(g(\theta)\) is the density function of \(\xi_i, \theta\), uniformly continuous in \(\theta\), uniformly in \(\xi_i\). Then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left[ E \left[ h(x_i, z_i, \theta) \frac{1}{\alpha_n} \frac{\partial K}{\partial \alpha_n} \left( \frac{h(x_i, z_i, \theta) - E[h(x_i, z_i, \theta)\xi_i, \theta, g(\theta)\xi_i, \theta]}{\alpha_n} \right) \right] \right] = 0
\]

Furthermore, if \(K(x)\) has zero moments up to the order \(s\), \(E[h(x_i, z_i, \theta)|\xi_i, \theta, g(\theta)]\) is differentiable at \(\alpha_n\) everywhere to the order \(s-1\) and these derivatives are uniformly bounded, then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left[ E \left[ h(x_i, z_i, \theta) \frac{1}{\alpha_n} \frac{\partial K}{\partial \alpha_n} \left( \frac{h(x_i, z_i, \theta) - E[h(x_i, z_i, \theta)\xi_i, \theta, g(\theta)\xi_i, \theta]}{\alpha_n} \right) \right] \right] = 0.
\]

Proposition A1.4 Let \(K(x)\) be a twice differentiable function on \(\mathbb{R}^d\) with a bounded support \(D\) such that \(K(x)\) and its gradient \(\nabla K(x)\) go to zero at the boundary of \(D\), and the gradient \(\nabla K(x)\) and its Hessian \(\nabla^2 K(x)\) are bounded. Suppose that \(\frac{1}{n} \sum_{i=1}^{n} E[h(x_i, z_i, \theta)|\xi_i, \theta, g(\theta)]\) are uniformly continuous in \(\theta\), uniformly in \(\xi_i\).

Proposition A1.5 Let \(C_j(z_i)\), \(j = 1, \ldots, k\), be two sequences of measurable functions of an i.i.d. sample \(\{z_i\}\). The \(d_j(d_k)\) is a sequence of measurable functions with the property that either \(E[d_j(d_k)] < \infty\) uniformly in \(\omega\) or \(\sup_{\omega} [d_j(d_k)] = O(1)\).

Suppose that, for each \(j\),

1. \(\sup_{\omega} [E[C_j(z_i)|\xi_i, \theta, g(\theta)] - C_j^0(z_i)] = O(d_j(d_k)) \]
2. \(\sup_{\omega} \text{var}[C_j(z_i)|\xi_i, \theta, g(\theta)] = O(d_j(d_k)) \]

If \(\lim_{n \to \infty} \sqrt{n} d_j = 0\) for \(j = 1, \ldots, 2\), then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} C_j(z_i)|\xi_i, \theta, g(\theta) - C_j^0(z_i) = 0\]

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 \(\{A_j(z_i)\}\) and \(\{C_j(z_i)\}\), \(j = 1, \ldots, k\), be two sequences of measurable functions of an i.i.d. sample \(\{z_i\}\). Suppose that,

1. \(\sum_{j=1}^{k} [A_j(z_i)] = O(1)\)
2. \(\sum_{j=1}^{k} [C_j(z_i)] = O(1)\)
3. \(\text{var} [C_j(z_i)|\xi_i, \theta, g(\theta)] = O(d_j(d_k))\)

If \(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} A_j(z_i)|\xi_i, \theta, g(\theta) - A_j^0(z_i) = 0\), where \(\text{var} [C_j(z_i)|\xi_i, \theta, g(\theta)] = O(d_j(d_k))\), then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} C_j(z_i)|\xi_i, \theta, g(\theta) - C_j^0(z_i) = 0\]

Proof: Since \(\sup_{\omega} \text{var} [C_j(z_i)|\xi_i, \theta, g(\theta)] = O(d_j(d_k))\), it is sufficient to investigate the convergence of \(\{A_j(z_i)\}\) and \(\{C_j(z_i)\}\).
\[ n^\beta \sup_i |C_n(s_i) - K_n(s_i)|^\beta \]. As \( n_{\alpha,n} = O(1/n^{\alpha}) \) in Proposition A.11, \( n = O(1/n^{\alpha}) \). Therefore, \( n^{(1/\beta)/\beta} \sup_i \left| C_n(s_i) - K_n(s_i) \right|^{1/\beta} \) is also uniformly continuous at \( c_i \). If \( \frac{1}{n_{\alpha,n}} \left[ \frac{1}{n_{\alpha,n}} \sum_{i=1}^{n_{\alpha,n}} |\Phi_n(s_i) - \Phi_n(c_i)| \right] \) goes to 0 uniformly in \( i \), then \( \frac{1}{n_{\alpha,n}} \left[ \frac{1}{n_{\alpha,n}} \sum_{i=1}^{n_{\alpha,n}} |\Phi_n(s_i) - \Phi_n(c_i)| \right] \) goes to 0 uniformly in \( i \). This rate requirement is equivalent to that \( \frac{1}{n_{\alpha,n}} \left[ \frac{1}{n_{\alpha,n}} \sum_{i=1}^{n_{\alpha,n}} |\Phi_n(s_i) - \Phi_n(c_i)| \right] \) goes to 0 uniformly in \( i \).

The result follows. Q.E.D.

Proposition A.1.7 Let \( \{ \{ \right \} \) be an i.i.d. sample and \( \Phi_n(x,a_n) \) be a sequence of vector-valued random functions with bandwidth \( \{ a_n \} \). Suppose that

1. There exist square integrable functions \( \psi_j(x) \), \( j = 1, 2 \) such that \( \left| \Phi_n(x,a_n) \right| \leq \psi_j(x) \), \( j = 1, 2 \).
2. \( \lim_n \sup \left| \Phi_n(x,a_n) \right| + \sup \left| \Phi_n(x,a_n) \right| = 0 \).

If \( \psi_j(x) \) and \( \psi_j(x) \) are zero a.e., then \( \lim_n \sup \left| \Phi_n(x,a_n) \right| = 0 \).

On the other hand, if \( \lim_n \sup \left| \Phi_n(x,a_n) \right| = a \) and \( \lim_n \sup \left| \Phi_n(x,a_n) \right| = 0 \), then

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n_{\alpha,n}} \Phi_n(s_i, x, a_n) = N(0,0). 
\]


Proposition A.1.8 Suppose that \( K \) is an \( r \)-dimensional kernel function with a bounded support \( D \) such that \( f(x) K(x,y) \) is uniformly lower bounded on \( D \), and there exists an \( \alpha \) such that \( \|K\|_{\alpha} < \infty \). Let \( \Phi_n(x,a_n) \) be the density function \( \Phi_n(x,a_n) \) and let \( \phi_n(s) \) be the density of \( \Phi_n(x,a_n) \). Denote \( A_{\alpha}(s) = \|K\|_{\alpha} \) and let \( f_{\alpha}(x,a_n) \) be measurable functions such that \( \|f_{\alpha}(x,a_n)\|_{\alpha} \) is square integrable. Suppose that

1. \( E[\phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha}] \) exists uniformly on \( s \).
2. \( E[\phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha}] \) is continuous in \( s \).
3. \( E[\phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha}] \) exists uniformly on \( s \).
4. \( \phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha} \) is uniformly bounded on \( s \).

If \( \lim_n \sup \left| \phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha} \right| = a \) and \( \lim_n \sup \left| \phi_n(s) \|f_{\alpha}(x,a_n)\|_{\alpha} \right| = 0 \), then

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n_{\alpha,n}} \Phi_n(s_i, x, a_n) = N(0,0). 
\]


Assumption 1:

1. \( \Theta \) is a compact convex subset of a finite dimensional Euclidean space, and the true parameter vector \( \theta \) is in the interior of \( \Theta \).
2. The sample observations \( (x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n) \) are i.i.d. variables in \( R^d \).
3. \( x \) is a \( d \)-dimensional vector of continuous random variables with a density \( p(x) \) and a compact support \( S \subset R^d \).
4. \( \theta \) is twice differentiable w.r.t. \( \theta \), and its first two order derivatives are bounded on \( \Theta \).
5. For each \( \theta \in \Theta \), \( z \) is an \( r \)-dimensional vector of continuous variables with a density \( p(z|\theta) \).
6. \( z \) is a vector-valued measurable function with known form which satisfies the relation \( E[f(z,\theta|(\theta))] = E[f(z,\theta,\theta)|\theta] \).

Assumption 2:

1. \( f \) is differentiable w.r.t. \( \theta \) to the third order. The third order moment, where \( r \geq 2 \) of \( \sup_{\theta} \|f(z,\theta)\|_{\infty} \), \( \sup_{\theta} \|f_{\theta}(z,\theta)\|_{\infty} \), \( \sup_{\theta} \|f_{\theta\theta}(z,\theta)\|_{\infty} \), and \( \sup_{\theta} \|f_{\theta\theta\theta}(z,\theta)\|_{\infty} \), where \( \theta \)'s are components of \( \theta \), are finite. The first moment of \( \sup_{\theta} \|f_{\theta}(z,\theta)\|_{\infty} \) is finite for all \( i, j, k, l \), \( j, k, l \) exist.
2. \( \theta \) is twice differentiable w.r.t. \( \theta \), and its first two order derivatives are bounded on \( \Theta \).
3. \( z \) is a vector-valued measurable function with known form which satisfies the relation \( E[f(z,\theta|(\theta))] = E[f(z,\theta,\theta)|\theta] \).

Assumption 3:

1. \( K \) is a continuous kernel function on \( R^d \) with a bounded support.
2. \( K \) is a kernel with zero moments up to the order \( s \), i.e., \( \int_{R^d} K(x)dx = 0 \) for all \( 0 \leq i \), \( l = 1, \ldots, k \), and \( 1 \leq i_1 + \cdots + i_k + i \leq s \).
(3.3) The bandwidth sequence \( \{a_n\} \) is a sequence of positive constants such that \( \lim_{n \to \infty} a_n = 0 \) and 
\[ \lim_{n \to \infty} n^{1/(r+1)} a_n = \infty. \]

Assumption 4:

(4.1) The functions \( p(z|\theta) \), \( E[|z|^r|\theta]|z|^{r+1} \), \( E[|z|^{r+1}|\theta]|z|^{r+1} \), 
\( E \left[ \frac{d}{d\theta} |z|^{r+1} \right] \), and \( E \left[ \frac{d^2}{d\theta^2} |z|^{r+1} \right] \), where \( \theta \) is a component of \( \theta \), are bounded on \( S_\theta \times \Theta \).

(4.2) \( p(z|\theta) \), \( E[|z|^r|\theta]|z|^{r+1} \), and \( E \left[ \frac{d}{d\theta} |z|^{r+1} \right] \), and \( E \left[ \frac{d^2}{d\theta^2} |z|^{r+1} \right] \) are differentiable w.r.t. \( z \) to the order \( s_2 \) and these \( s_2 \) order derivatives are uniformly bounded on \( S_\theta \times \Theta \). \( E[|z|^r|\theta]|z|^{r+1} \) and \( E \left[ \frac{d}{d\theta} |z|^{r+1} \right] \) are differentiable w.r.t. \( z \) to the order \( s_2 + 1 \), and these \( s_2 + 1 \) order derivatives are uniformly bounded on \( S_\theta \times \Theta \).

(4.3) The function \( E \left[ \frac{d^3}{d\theta^3} |z|^{r+1} \right] \) is the first order derivatives of \( E[|z|^r|\theta]|z|^{r+1} \), \( E[|z|^r|\theta]|z|^{r+1} \), and \( E \left[ \frac{d}{d\theta} |z|^{r+1} \right] \) w.r.t. \( z \); and the second order derivatives of \( E[|z|^r|\theta]|z|^{r+1} \) and \( E \left[ \frac{d}{d\theta} |z|^{r+1} \right] \) w.r.t. \( z \) are uniformly continuous in \( z \), uniformly in \( (x, \theta) \) in \( S_\theta \times \Theta \).

Assumption 5:

(5.1) \( J(u) \) is a twice continuously differentiable kernel function on \( \mathbb{R}^m \) with a bounded support. \( \frac{\partial^2}{\partial^2 u} J(u) \) satisfies the bounded Lipschitz condition of order \( 1 \) w.r.t. \( u \).

(5.2) \( V(u) \) is a kernel with zero moments up to the order \( s_3 \) with \( s_3 > 2 \).

(5.3) The bandwidth sequence \( \{b_n\} \) is a sequence of constants such that \( \lim_{n \to \infty} b_n = 0 \) and 
\[ \lim_{n \to \infty} \frac{n^{1/(r+1)} b_n^{1/(r+1)+1}}{\ln n} = \infty. \]

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 \( S_\theta \).

(6.3) \( p(x|\theta) \) is bounded away from zero on \( X \), and \( p(x|\theta) \) is bounded away from zero uniformly on \( X \times \Theta \).

(6.4) \( E[G(x, \theta)W(x)G(x, \theta)] \) is nonsingular.

(6.5) The bandwidths \( a_n \) and \( b_n \) satisfy the rate of convergence that \( \lim_{n \to \infty} na_n = \infty \), \( \lim_{n \to \infty} nb_n = \infty \); and \( \lim_{n \to \infty} a_n^{1/(r+1)} = 0 \) and \( \lim_{n \to \infty} b_n^{1/(r+1)} = 0 \).

Assumption 7:

(7.1) The \( r \)th moment of \( \sup_{\theta \in \Theta} \left| E[f(x, \theta)|\theta]|x|^{r+1} \right| \) exists, where \( r > 2 \).

(7.2) \( E[f(x, \theta)|\theta]|x|^{r+1} \) is uniformly continuous on \( S_\theta \times \Theta \).

(7.3) \( V(x) \) is a positive definite for each \( x \in S \).

(7.4) The matrix \( E[G(x, \theta)G(x, \theta)] = E[G(x, \theta)G(x, \theta)] \) is nonsingular.

(7.5) The bandwidths \( a_n \) and \( b_n \) satisfy the following rates of convergence: 
\[ \lim_{n \to \infty} \frac{n^{1/(r+1)} a_n^{1/(r+1)+1}}{\ln n} = \infty, \]
\[ \lim_{n \to \infty} n a_n^{1/(r+1)} = \infty, \]
\[ \lim_{n \to \infty} n b_n^{1/(r+1)} = \infty, \]
and 
\[ \lim_{n \to \infty} \frac{n^{1/(r+1)} b_n^{1/(r+1)+1}}{\ln n} = \infty. \]
Similar conclusion holds for \( B_{Jn}(x, \theta) \) with its limit bring \( B_J(x, \theta) = p(x, \theta) \). Let \( t_n(x, \theta) \) be a trimming function such that \( t_n(x, \theta) \neq 0 \) only if \( B_{Jn}(x, \theta) \geq \frac{\theta}{n} \). Since

\[
E_n(f(z, \beta)[x, \theta]) - E(f(z, \beta)[x, \theta]) = B_{Jn}(x, \theta) \left[ A_{Jn}(x, \theta) - A_J(x, \theta) \right] B_J(x, \theta) - A_J(x, \theta) \left( B_{J} (x, \theta) - B_{Jn}(x, \theta) \right),
\]

\[
\sup_{x, \theta} \left\{ E_n(f(z, \beta)[x, \theta]) - E(f(z, \beta)[x, \theta]) \right\} \leq \frac{2}{\Delta} \sup_{x, \theta} \left\{ \left[ A_{Jn}(x, \theta) - A_J(x, \theta) \right] B_J(x, \theta) \right\} + \sup \left\{ E(f(z, \beta)[x, \theta]) \right\} \cdot \sup_{x, \theta} \left[ A_{Jn}(x, \theta) - B_{Jn}(x, \theta) \right]
\]

\[
= o_p(1). \tag{4.23}
\]

As \( p(x, \theta) \), \( E[f(z, \beta)[x, \theta]] \), and \( E[f(z, \beta)[x, \theta]] \) are bounded, their variances have order \( O(\Delta^{-1}) \) uniformly on \( S_0 \times \Theta \). By Proposition A1.3 their biases have order \( O(\Delta^2) \) uniformly on \( S_0 \times \Theta \) under Assumptions 4 and 5. Proposition A1.1 implies that as \( \lim_{\Delta \to 0} \Delta^{-1} \sum_{x, \theta} A_{Jn}(x, \theta) - A_J(x, \theta) \| \} = 0 \).

For each component \( \theta \) of \( \theta \), a similar result holds for \( E_n(f(z, \beta)[x, \theta]) \) and \( E(f(z, \beta)[x, \theta]) \) with its limit in

\[
\delta \left( E(f(z, \beta)[x, \theta]) \right) = \frac{1}{B_J(x, \theta)} A_J(x, \theta) - A_J(x, \theta) B_J(x, \theta),
\]

\[
\delta E_n(f(z, \beta)[x, \theta]) = \frac{1}{B_J(x, \theta)} A_J(x, \theta) - A_J(x, \theta) B_J(x, \theta).
\]

\[
= \frac{1}{B_J(x, \theta)} \left[ A_{Jn}(x, \theta) - A_J(x, \theta) \right] B_J(x, \theta) + \left[ E(f(z, \beta)[x, \theta]) B_J(x, \theta) - A_J(x, \theta) \right] B_J(x, \theta) - B_J(x, \theta)
\]

\[
= \left[ E(f(z, \beta)[x, \theta]) B_J(x, \theta) - A_J(x, \theta) \right] B_J(x, \theta) - B_J(x, \theta).
\]

From (4.21), as \( \sup_{x, \theta, \omega} \left\{ \left| B_{Jn}(x, \theta) - B_J(x, \theta) \right| \right\} = o(1) \) when \( \lim_{\Delta \to 0} \Delta^{-1} \sum_{x, \theta} \left| B_{Jn}(x, \theta) - B_J(x, \theta) \right| = \infty \) and \( \lim_{\Delta \to 0} \Delta^{-1} = 0 \), it follows that, with probability arbitrarily close to one for large \( n \), whenever \( B_{Jn}(x, \theta) \geq \Delta \), \( B_J(x, \theta) \geq \frac{\theta}{n} \). Therefore, under the rate \( \lim_{\Delta \to 0} \Delta^{-1} \sum_{x, \theta} \left| B_{Jn}(x, \theta) - B_J(x, \theta) \right| = \infty \) and \( \lim_{\Delta \to 0} \Delta^{-1} = 0 \),

\[
\sup_{x, \theta} \left\{ \left| E_n(f(z, \beta)[x, \theta]) - E(f(z, \beta)[x, \theta]) \right| \right\} = o_p(1). \tag{4.23}
\]

The second order derivatives of the nonparametric regression function is

\[
\frac{\partial^2 E_n(f(z, \beta)[x, \theta])}{\partial \theta \partial \omega} - \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta \partial \omega}
\]

\[
= \frac{1}{B_J(x, \theta)} \left[ \frac{\partial^2 A_J(x, \theta)}{\partial \theta \partial \omega} - \frac{\partial^2 A_J(x, \theta)}{\partial \theta \partial \omega} \right] B_J(x, \theta)
\]

\[
- \frac{\partial E_n(f(z, \beta)[x, \theta])}{\partial \theta} B_J(x, \theta) - B_J(x, \theta) \frac{\partial E_n(f(z, \beta)[x, \theta])}{\partial \theta}
\]

\[
- \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta} B_J(x, \theta) - B_J(x, \theta) \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta}
\]

\[
= \frac{1}{B_J(x, \theta)} \left[ \frac{\partial^2 A_J(x, \theta)}{\partial \theta \partial \omega} - \frac{\partial^2 A_J(x, \theta)}{\partial \theta \partial \omega} \right] B_J(x, \theta)
\]

\[
+ \frac{\partial E_n(f(z, \beta)[x, \theta])}{\partial \theta} B_J(x, \theta) - B_J(x, \theta) \frac{\partial E_n(f(z, \beta)[x, \theta])}{\partial \theta}
\]

\[
- \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta} B_J(x, \theta) - B_J(x, \theta) \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta}
\]

\[
\sup_{x, \theta} \left\{ \left| \frac{\partial E_n(f(z, \beta)[x, \theta])}{\partial \theta} - \frac{\partial E(f(z, \beta)[x, \theta])}{\partial \theta} \right| \right\} = o_p(1). \tag{4.25}
\]

A3: Semiparametric-distance Estimator

Proof of Proposition 3.1: Let

\[
Q_{I1}(z, \theta) = \frac{1}{n} \sum_{i=1}^n C_{I1}(z, \theta) W(z_i) G_{n}(z_i, \theta),
\]

and \( Q_{I1}(z, \theta) = \frac{1}{n} \sum_{i=1}^n C_{I1}(z, \theta) W(z_i) G_{n}(z_i, \theta) \). From Appendix A, \( E[f(z, \beta)[x, \theta]] \) and \( E[f(z, \beta)[x, \theta]] \) converge in probability, respectively, to \( E[f(z, \beta)[x, \theta]] \) and \( E[f(z, \beta)[x, \theta]] \) uniformly on \( \Theta \). From the classical uniform law of large number (e.g., Amemiya [1985], Theorem 4.2.1), it implies that \( Q_{I1}(z, \theta) \) converges in probability to \( Q(z, \theta) \) uniformly on \( \Theta \), where \( Q(z, \theta) = E[G(z, \theta) W(z)] G(z, \theta) \). Under the identification condition, \( Q(z, \theta) = 0 \). In any global minimum at \( z = z_0 \). The consistency of \( \hat{\theta} \) follows.

The SMD estimator \( \hat{\theta} \) satisfies the first order condition:

\[
\sum_{i=1}^n C_{I1}(z_i, \hat{\theta}) W(z_i) G_{n}(z_i, \hat{\theta}) = 0.
\]

Without loss of generality, take \( f \) to be a scalar valued function for simplicity.\footnote{This simplifies only the notation for second order derivatives.} From a mean value theorem,

\[
\sqrt{n}(\hat{\theta} - \theta) = \left( \frac{1}{n} \sum_{i=1}^n C_{I1}(z_i, \theta) W(z_i) G_{n}(z_i, \theta) \right) \cdot \left( \frac{1}{n} \sum_{i=1}^n C_{I1}(z_i, \theta) W(z_i) G_{n}(z_i, \theta) \right)^{-1}
\]

\[
\times \frac{1}{\sqrt{n}} \sum_{i=1}^n C_{I1}(z_i, \theta) W(z_i) G_{n}(z_i, \theta),
\]
where $G_{n,\theta}$ denotes the second order derivative matrix of $G_{n,\theta}$ w.r.t. $\theta$. Since $\theta$ converges in probability to $\theta$, it follows from (A2.2), (A2.5) and (A2.6) that $G_{n,\theta}(x,\theta) = 0$, $G_{n,\theta}(x,\theta) = G_{n,\theta}(x,\theta)$, and $\frac{G_{n,\theta}(x,\theta)}{\theta} \to G_{n,\theta}(x,\theta)$, where $G_{n,\theta}$ is the second order derivative matrix of $G$ w.r.t. $\theta$, uniformly in $x \in X$. It follows that

$$\sqrt{n}(\theta - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( G_{n,\theta}(x_i,\theta) W(x_i,\theta) \right)^{1/2} + o_p(1)$$

Denote $A_n(x) = A_n(x,\theta)$, $A_{n,\theta}(x) = A_{n,\theta}(x,\theta)$ and $B_n(x) = B_n(x,\theta)$ for simplicity. As $n$ goes to infinity, $A_n(x)$ converges in probability to $A(x)$ uniformly in $x$, where $A(x) = E[f(x,\theta)|x]$; $B_n(x)$ converges in probability uniformly to $B(x)$; $A_{n,\theta}(x)$ converges in probability uniformly to $A_{\theta}(x)$; $B_n(x)$ converges in probability uniformly to $B_n(x)$. Since $E[f(x,\theta)|x] = \frac{1}{\theta} \sum_i A_n(x_i,\theta,\theta)$, and $\frac{\partial E[f(x,\theta)|x]}{\partial \theta} = \frac{1}{B_n(x)} A_{n,\theta}(x)$, where

$G_{n}(x,\theta)$ is a function of $A_n(x)$, $B_n(x)$, $A_{n,\theta}(x)$, $B_{n,\theta}(x)$, $A_{n,\theta}(x)$, and $B_{n,\theta}(x)$. By using the expansion of difference, $G_{n}(x,\theta) = G_{n}(x,\theta) + \frac{1}{n} \sum_{i=1}^{n} \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$

and $R_{n,\theta}(x)$

$$R_{n,\theta}(x) = B_n(x) \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

On the other hand, since $G(x,\theta) = 0$, by an expansion up to the second order, $G_{n,\theta}(x) = c_n(x) + R_{n,\theta}(x)$, where

$$c_n(x) = \frac{1}{\theta} \sum_{i=1}^{n} \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

and $R_{n,\theta}(x)$

$$R_{n,\theta}(x) = \frac{1}{\theta} \sum_{i=1}^{n} \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

$$+ \frac{1}{B_n(x)} A_{n,\theta}(x) \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$$

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 $r_n(x,\theta) = f_1(x,\theta) + E_{\theta} f_1(x,\theta) d\theta + E_{\theta} f_1(x,\theta) d\theta$ and $r_n(x,\theta) = \frac{1}{\theta} \sum_{i=1}^{n} \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$. Let $Q_{\theta}(x) = \frac{1}{\theta} \sum_{i=1}^{n} \left( B_n(x_i) - B_n(x) \right)^{1/2} \left( B_n(x_i) - B_n(x) \right)$. By Proposition A1.1 and
uniform laws of large numbers, \( \mathbb{P}_n(\theta) \) converges in probability to \( \mathbb{P}(\theta) \) uniformly in \( \theta \), where \( \mathbb{P}(\theta) = \mathbb{E} p(y - r(z, \theta)W(x)z)W(x)z - r(z, \theta) \). As \( \mathbb{P}(\theta) \) is uniquely minimized at \( \hat{\theta} \), \( \hat{\theta}_N \) is consistent.

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

\[
\sqrt{n}(\hat{\theta}_N - \theta) = \mathbb{E} \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( p_s(x_j, \theta_s) - p_s(x_j, \theta_0) \right) \right] + o_p(1) - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( p_s(x_j, \theta_0) \right) W(x_j)z_{y_j} - r(x_j, \theta_0) \}
\]

Let \( A_N(z) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( p_s(x_j, \theta_0) \right) W(x_j)z_{y_j} - r(x_j, \theta_0) \)

where \( p = \{ p_1, p_2, \ldots \} \) and \( p_0 = \{ p_1, p_2, \ldots \} \) are bounded polynomial functions of \( x \) and \( z \) respectively.

As long as \( p \) is a bounded polynomial function of \( x \) and \( z \), the sample observations at which \( V_s(x, \theta) \) are nearly singular will be deleted. Since

\[
V_s(x, \theta) = V_N(x, \theta)/B^2_N(x, \theta),
\]

where

\[
V_N(x, \theta) = \left[ \frac{1}{n} \sum_{j=1}^{n} f(z_j, \theta) V(z_j) B_s(x_j) \right] - A_N(x, \theta),
\]

and the trimming function \( \tau \) is the product of the above trimming functions, i.e.,

\[
\tau(x, \theta) = \tau^1(B_n(x)/a_n) \cdot \tau^2(B_n(x)/B_n(x)),
\]

A5: Generalized Semi-parametric Estimator

Proof of Proposition 4.1

To simplify notation, let

\[
D_0(x, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left( t_s(x_i, \theta) G(x_i, \theta) W(x_i)z_{y_i} - r(x_i, \theta) \right),
\]

and

\[
U_0(x, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( t_s(x_i, \theta) G(x_i, \theta) W(x_i)z_{y_i} - r(x_i, \theta) \right).
\]

It follows that

\[
\sqrt{n}(\hat{\theta}_N - \theta_0) = \sqrt{n}(\theta - \theta_0) - D_0^2(\theta)/U_0(\theta).
\]

By the uniform convergence of nonparametric regression functions and their derivatives in Appendix A2, \( D_0(x, \theta) - D_0(x, \theta) \) is \( o_p(1) \) where \( D_0(\theta) = \frac{1}{n} \sum_{i=1}^{n} t_s(x_i, \theta) G(x_i, \theta) W(x_i)z_{y_i} - r(x_i, \theta) \). Define \( D_0^2(\theta) = \frac{1}{n} \sum_{i=1}^{n} G(x_i, \theta) W(x_i)z_{y_i} - r(x_i, \theta) \). As \( t_s(x_i, \theta) \) converges in probability to 1 a.e., it follows by the Markov inequality and the dominated convergence theorem that \( D_0^2(\theta) - D_0^2(\theta) \) is \( o_p(1) \). On the other hand, \( D_0^2(\theta) \) converges in probability to \( D(\theta) \), which defines \( D(\theta) = \mathbb{E} G(x, \theta) W(x)z_{y_i} - r(x, \theta) \), uniformly in \( \theta \). Since \( \hat{\theta}_N \) is a consistent estimate of \( \theta \) and \( D(\theta) \) is continuous in \( \theta \), \( D(\theta) \) converges to \( D(\theta) \) in probability. Therefore, \( \hat{\theta}_N \) converges in probability to \( D(\theta) \).

It remains to analyze \( U_0(\theta) \). Without loss of generality, consider that \( f \) is a scalar function for notational simplicity. By a mean value theorem, \( U_0(\theta) = U_0^1(\theta) + U_0^2(\theta) \), where

\[
U_0^1(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( t_s(x_i, \theta) G(x_i, \theta) W(x_i)z_{y_i} - r(x_i, \theta) \right).
\]

The latter term can be decomposed into a sum of two terms: \( U_0^1(\theta) = U_0^1(\theta) + U_0^2(\theta) \) where
Since \( \theta \) lies between \( \bar{\theta} \) and \( \bar{\theta} \), \( G_\delta(x, \theta) \) converges in probability to \( G(x, \theta) = 0 \). By uniform convergence of nonparametric functions, \( U_n^2(\theta) = o_p(1) \). Hence \( U_n^2(\theta) + o_p(1) = D_n^2(\theta) + o_p(1) \). Therefore (A5.1) is reduced to \( v_n(\theta - \theta) = -D_n^2(\theta)U_n(\theta) + o_p(1) \).

Let \( R_{n}(x) = V_n^2(x, \theta) - V_n^1(x, \theta) = V_n^1(x, \theta)[V_n^2(x, \theta) - V_n^1(x, \theta)]V_n^1(x, \theta) \). We have \( W_n(x, \theta) = V_n^1(x, \theta) + R_{n}(x) \). Define

\[
L_n(x, \theta) = \sum_{i=1}^{n} \left[ L_n^1(x, \theta) + L_n^2(x, \theta) \right]
\]

where \( V_j^1(x, \theta) \) is the probability limit of \( V_j(x, \theta) \). With \( u_i(x) \) and the remainders \( R_{n}(x) \) and \( R_{n}(x) \) defined in the proof of Proposition 3.1,

\[
U_n(x, \theta) = \sum_{i=1}^{n} \left[ L_n^1(x, \theta) + (u_i(x, \theta) - L_n^1(x, \theta)) \right] \Delta_n(x, \theta)
\]

where \( R_{n}(x) = R_{n+1}(x) + \cdots + R_{n}(x) \) is the overall remainder term consisting of fifteen terms of high order errors of nonparametric functions, where, for example,

\[
R_{n}^{1}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ L_n^1(x, \theta) + (u_i(x, \theta) - L_n^1(x, \theta)) \right] \Delta_n(x, \theta)
\]

\[
R_{n}^{2}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ L_n^2(x, \theta) + u_i(x, \theta) - L_n^2(x, \theta) \right] \Delta_n(x, \theta)
\]

\[
R_{n}^{3}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ L_n^3(x, \theta) + u_i(x, \theta) - L_n^3(x, \theta) \right] \Delta_n(x, \theta)
\]

and

\[
R_{n}^{4}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ (u_i(x, \theta) - L_n^1(x, \theta)) \Delta_n(x, \theta) \right]
\]

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 A1.1 can be applied to the following term:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ L_n^1(x, \theta) \Delta_n(x, \theta) \right] V_n^1(x, \theta) V_n(x, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ U_n(w_i, w_j) + U_n(w_i, w_j) \right]
\]

where

\[
U_n(w_i, w_j) = L_n^1(x, \theta) \Delta_n(x, \theta) V_n^1(x, \theta) V_n^1(x, \theta) - E[V_n^1(x, \theta) V_n^1(x, \theta)] \frac{1}{h(x)} \Delta_n(x, \theta)
\]

and

\[
U_n(w_i, w_j) = -L_n^1(x, \theta) \Delta_n(x, \theta) V_n^1(x, \theta) V_n^1(x, \theta) - E[V_n^1(x, \theta) V_n^1(x, \theta)] \frac{1}{h(x)} \Delta_n(x, \theta)
\]

By the asymptotic unbiasedness of nonparametric functions, \( E[U_n(w_i, w_j) w] = 0 \). Furthermore, \( \lim_{n \to \infty} E[U_n(w_i, w_j) w] = G(x, \theta) V_n^1(x, \theta) V_n^1(x, \theta) w \), but

\[
\lim_{n \to \infty} E[U_n(w_i, w_j) w] = -E[G(x, \theta) V_n^1(x, \theta) V_n^1(x, \theta) w] = 0.
\]
References
