Semiparametric Minimum-distance Estimation

Lung-fei Lee

November, 1991

Number 92–08
Semiparametric Minimum-distance Estimation

by
Lung-fei Lee
Department of Economics
The University of Michigan
Ann Arbor, MI 48109

November, 1991

Abstract
Semiparametric minimum-distance estimation methods are introduced for the estimation of parametric or semiparametric econometric models. The semiparametric minimum-distance estimation methods share some familiar properties of the classical minimum-distance estimation method. However, they can be applied to the estimation of models with disaggregated data. Asymptotic properties of the estimators are analyzed. Some goodness-of-fit test statistics are introduced. For the estimation of some econometric models, weighted minimum-distance estimators can be asymptotically efficient. The minimum-distance estimators are asymptotically invariant with respect to some transformations.

JEL classification number: 211

Keywords:
Semiparametric estimation, minimum-distance, minimum-chi-square, semiparametric goodness-of-fit test, nonparametric kernel regression, index restriction, quantal response, limited dependent variable, simultaneous equation models.

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

by Lung-fei Lee*

1. Introduction

The use of minimum-distance (MD) estimation methods has a long history in statistics and econometrics; e.g., Berkson [1944], Neyman [1949], Taylor [1953], Ferguson [1958], Rothenberg [1973], and Chamberlain [1982], among others. The MD estimation techniques are methods for imposing restrictions in estimation. In the context of the qualitative response (QR) model, MD methods were first proposed by Berkson [1944] for instances when there are many observations of responses for each value of the dependent variable. These minimum-distance estimation methods share some common features. There are deterministic relations or restrictions on the parameters of interest. The estimation procedure tries to solve for the unknown parameters of interest while the unrestricted components in the relations can be estimated consistently. As there are more equations of restrictions than unknown parameters, minimization with respect to some distance measure is used to resolve the over-identification problem.

As pointed out in many textbooks, the Berkson method for the QR model can be used only when there are many observations on the dependent variable (the response variable) for each value of the independent variable. This is so because the frequencies of responses are used to create the unrestricted estimates of the response probabilities. The Berkson method is applicable to situations with grouped data or discrete explanatory variables. For disaggregated data with continuous explanatory variables, one has to use the method of maximum likelihood (see, e.g., Amemiya [1985]). However, from the recent development of nonparametric methods, we see that 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 or sample mean. One may wonder whether MD type estimation methods can be generalized to the semiparametric setting. This article provides a positive answer to this question.

In this article we consider the generalization of MD estimation methods to the estimation of models with disaggregated data and conditional expectation structures. The conditional expectation structures provide model restrictions for identification and estimation. We consider models with conditional expectation structures of which the conditioning arguments are either exogenous variables or functions of exogenous variables in index forms involving unknown parameters. The MD estimation methods can be applied to the estimation of parametric models or semiparametric models.

This article is organized as follows. In Section 2, we introduce a semiparametric MD estimation method. The semiparametric MD method can be applied to the estimation of parametric and semiparametric models. Section 3 considers the semiparametric MD estimation method with as many restrictions as the number of the sample size. Asymptotic properties of the estimator are analyzed. Section 4 considers a weighted MD estimation procedure. Asymptotic efficiency issues are discussed for some models. The semiparametric MD estimation procedures possess some transformation-invariance properties as shown in Section 5. In the same section, a local goodness-of-fit test statistic is developed. Applications of the semiparametric MD procedure to the estimation of some semiparametric simultaneous equation models with qualitative and/or limited dependent variables and endogenous switching regression models are illustrated in Section 6. Section 7 points out possible generalizations of the estimation method to semiparametric models with model restrictions not covered in the previous sections. Section 8 provides a summary of this article. Four appendices are provided. Appendix 1 collects some propositions which are useful for the proofs of our results. Appendix 2 summarizes the assumed regularity conditions. Relevant asymptotic properties of nonparametric kernel regression functions and their derivatives are summarized in this appendix for reference. Appendix 3 collects the proofs of the main results of this article. Appendix 3 considers semiparametric MD methods with a finite number of restrictions. Asymptotic properties and their relations with minimum chi-square methods are derived.

* I am grateful for support from the NSF under grant no. SES-9010516 for my research.
2. Conditional Expectation and Semiparametric Minimum-distance Estimation

In the context of the QR model, MD methods were proposed by Berkson [1944] for instances when there are many observations of responses for each value of the independent variable. Consider a binary response case. Let \( y \) be the dichotomous indicator, and \( z \) a vector of explanatory variables. In a parametric QR model, one specifies a parametric response probability \( F(z, \alpha) \) explained by \( z \), i.e., \( \text{Prob}(y = 1|z) = F(z, \alpha) \).

The logistic or normal probability models with \( z \) and \( \alpha \) appearing as an index \( za \) in \( F(za) \) are the popular specifications. With many observations of responses for each \( z \), \( \text{Prob}(y = 1|z) \) can be estimated by a frequency estimate, say \( \hat{p}(z) \). The Berkson MD estimators of \( \alpha \) are derived by minimization:

\[
\min_{\alpha} \sum_{i=1}^{m} \omega(z_i)(\hat{p}(z_i) - F(z_i, \alpha))^2, \tag{2.1}
\]

where \( m \) is the number of distinct values of \( z \), and \( \omega(z_i) \) is some weighting function. An optimal weighting function for this method is the inverse of the variance function of \( \hat{p}(z) \). For models with \( za \) as an index, a computationally simpler Berkson MD estimation procedure is possible. As \( F \) is a strictly increasing function, \( F^{-1}[\text{Prob}(y = 1|z)] = za \). The parameter \( \alpha \) can be estimated by a least-squares or weighted least-squares procedure to the equation:

\[
F^{-1}[\hat{p}(z_i)] = z_i\alpha + \epsilon_i, \quad i = 1, \ldots, m, \tag{2.2}
\]

where \( \epsilon_i = F^{-1}[\hat{p}(z_i)] - F^{-1}[\text{Prob}(y = 1|z_i)] \). This modified MD method is computationally simple, and the optimal weighted least-squares estimator is asymptotically efficient (see, Amemiya [1981]). Such a computationally simple MD procedure can be generalized to the estimation of many complicated QR models (Amemiya [1981]). For the case that \( z \) is a continuous random vector (with an absolutely continuous distribution), many observations of \( y \) for each \( z \) will not be possible. In such a circumstance, the recommended estimation method is usually the method of maximum likelihood.

The MD estimation methods for the QR model rely on the availability of a consistent estimator of \( \text{Prob}(y = 1|z) \) for each \( z \). When \( z \) is a continuous random vector, a frequency estimate of \( \text{Prob}(y = 1|z) \) does not exist, but nonparametric estimates of \( E(y|z) \) exist at each value of \( z \). As \( E(y|z) = \text{Prob}(y = 1|z) \) in the binary QR model, this motivates a nonparametric or semiparametric MD estimation procedure. For a random sample of size \( n \), suppose that \( E_n(y|z) \) denotes a nonparametric regression estimate of \( \text{Prob}(y = 1|z) \) at \( z \). Then a semiparametric version of (2.1) is

\[
\min_{\alpha} \sum_{i=1}^{n} \omega(z_i)(E_n(y|z_i) - F(z_i, \alpha))^2, \tag{2.3}
\]

and a corresponding version of (2.2) is

\[
F^{-1}[E_n(y|z_i)] = z_i\alpha + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.4}
\]

The QR model motivates our estimation approach. However, this approach can be applied to the estimation of many parametric and semiparametric models. Consider the estimation of \( \theta \) in the following equation of a parametric model:

\[
G[E(f(z_1, \beta(\theta))|z), z, \alpha(\theta)] = 0, \tag{2.5}
\]

where \( G \) and \( f \) are functions with known functional forms and \( \alpha \) and \( \beta \) are functions of a deep parameter vector \( \theta \). To simplify notation, \( \theta \) in \( \alpha(\theta) \), \( \beta(\theta) \), and \( \gamma(\theta) \) will be suppressed in subsequent presentations when there is no danger of confusion.

For the binary QR model, \( G[E(y|z), z, \alpha] = E(y|z) - F(z, \alpha) \) or \( G[E(y|z), z, \alpha] = F^{-1}[E(y|z)] - x\alpha \). The model also includes regression models and simultaneous equation models. A regression model,

\[
y = g(x, \alpha) + \epsilon, \quad E(\epsilon|x) = 0, \tag{2.6}
\]

implies \( E(y|x) = g(x, \alpha) \). A simultaneous equation model,

\[
f(z_1, \beta) = u, \quad E(u|x) = 0, \tag{2.7}
\]
where $z_i$ is a vector consisting of endogenous and exogenous variables and $x$ is the vector of all exogenous variables in the system, implies $E(f(z_i, \beta)|x) = 0$. Many other econometric models may imply some relations as in (2.5).

Semiparametric MD methods can be applied to the estimation of (2.5). The vector of conditional expectations $E(f(z_i, \beta)|x)$ can be estimated by nonparametric kernel regressions. 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 $n$. $E(f(z_i, \beta)|x_i)$ can be estimated by

$$E_n(f(z_i, \beta)|x_i) = \frac{1}{n-1} \sum_{j \neq i} f(z_{ij}, \beta) K \left( \frac{z_i - z_{ij}}{a_n} \right).$$

A semiparametric minimum-distance (SMD) estimation procedure can be defined as

$$\min_{\theta \in \Theta} \sum_{i=1}^{m_n} I_X(x_i) G^2 \left[ E_n(f(z_i, \beta(\theta))|x_i), x_i, \alpha(\theta) \right],$$

where $m_n \leq n$, $\Theta$ is the parameter space of $\theta$, and $I_X$ is an indicator of some set $X$ which is a compact subset contained in the interior of the support of $z$. $X$ is constructed to trim the tails of the distribution of $x$ because the nonparametric regression estimates at the boundary of the support of $z$ might not be well-behaved.

The semiparametric MD estimation approach can be generalized to the estimation of semiparametric models. We consider the estimation of semiparametric models with index restrictions (Stoker [1986] and Ichimura [1987]), which have broad applications in microeconometrics. We consider estimation of $\theta$ in the following equation:

$$G(E(f(z_1, \beta(\theta))|x), E(g(z_2, \gamma(\theta))|x_\delta(\theta)), x, \alpha(\theta)) = 0,$$

where $f$ and $g$ are vector-valued functions, and $f$, $g$, and $G$ are functions with known functional forms, and $x_\delta$ is an $m$-dimensional vector of indices. Equation (2.10) differs from (2.5) in the conditioning arguments. As the $\delta$'s can be arbitrary functions, it also generalizes (2.5) to allow subsets of variables in $x$ to be the conditioning variables. $E(f(z_1, \beta)|x_\delta)$ can be estimated by (2.8), and $E(g(z_2, \gamma)|x_\delta)$ can be estimated by

$$E_n(g(z_2, \gamma)|x_\delta) = \frac{1}{n-1} \sum_{j \neq i} g(z_{2j}, \gamma) J \left( \frac{z_{ij} - z_{ij}}{b_n} \right),$$

where $J$ is a $m$-dimensional kernel function with a bandwidth $\{b_n\}$. An SMD estimation procedure is

$$\min_{\theta \in \Theta} \sum_{i=1}^{m_n} I_X(x_i) G^2 \left[ E_n(f(z_1, \beta(\theta))|x_i), E_n(g(z_2, \gamma(\theta))|x_\delta(\theta)), x_i, \alpha(\theta) \right].$$

Some specific applications of this method to the estimation of simultaneous equation models with qualitative and limited dependent variables will be provided in a subsequent section.

A special feature of (2.5) and (2.10) is that there are no explicit disturbances in the structural equations. Stochastic elements are introduced only through the nonparametric regression estimates of $E(f(z_1, \beta)|x)$ and $E(g(z_2, \gamma)|x_\delta)$ in the estimation procedure. As the function $G$ may be nonlinear, the sampling errors $E_n(f(z_1, \beta)|x) - E(f(z_1, \beta)|x)$ and $E_n(g(z_2, \gamma)|x_\delta) - E(g(z_2, \gamma)|x_\delta)$ appear implicitly in the estimated equation. As in classical MD estimation, consistency of the estimator of $\theta$ is possible because the sampling errors converge in probability to zero as the sample size $n$ increases to infinity. Consistency of the SMD estimator of $\theta$ requires a large (infinite) sample so that $E_n(f(z_1, \beta)|x)$ can converge in probability to $E(f(z_1, \beta)|x)$, and $E_n(g(z_2, \gamma)|x_\delta)$ is a consistent estimate of $E(g(z_2, \gamma)|x_\delta)$. The number $m_n$ of equations used for estimation is an interesting issue. In classical MD estimation, only a finite number of restrictions is available. For our model, there are implicitly infinitely many restrictions. There must be some benefit of using all the
restrictions for estimation. Indeed, that is the case. As shown in Appendix 4, SMD estimators derived from a finite and fixed number of restrictions (not depending on sample size) can be consistent and asymptotically normal. However, the rate of convergence of the parameter estimates is as slow as the rate of convergence of the nonparametric regression function estimates. With a finite and fixed number of restrictions, a properly weighted SMD procedure is also a minimum chi-square procedure. Its minimized distance function is proportional to a chi-square random variable with the number of degrees of freedom being the number of overidentification restrictions. With a finite and fixed number of restrictions, the SMD method shares some of these familiar properties of the classical MD methods. The details are referred to Appendix 4. In the subsequent sections, we investigate the semiparametric MD method with as many restrictions as the number of sample points. The number of restrictions increases as the sample size increases. With infinitely many restrictions, the semiparametric MD estimators can converge at the $\sqrt{n}$ parametric rate.
3. Semiparametric Minimum-distance Estimation

With $m_n = n$, the SMD estimation is

$$
\min_{\theta \in \Theta} \sum_{i=1}^{n} I_X (x_i) G^2 [E_n (f(z_1, \beta(\theta)) | x_i), E_n (g(z_2, \gamma(\theta)) | x_i \delta(\theta)), x_i, \alpha(\theta)].
$$

(3.1)

Let $\hat{\theta}_I$ denote the SMD estimate from (3.1) and $\theta_0$ the true parameter vector, and let $\alpha_0$, $\beta_0$, $\gamma_0$ and $\delta_0$ denote respectively the functions $\alpha$, $\beta$, $\gamma$, and $\delta$ evaluated at $\theta_0$. For any possible value $\theta$ of $\theta_0$ in $\Theta$, $E(\cdot | x \delta, \theta)$ refers to a conditional expectation conditional on $x \delta$ for a given value $\theta$. At $\theta_0$, $E(\cdot | x \delta_0)$ represents $E(\cdot | x \delta_0, \theta_0)$ for simplicity. 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, $\hat{\theta}_I$ can be consistent and its rate of convergence is $O(\frac{1}{\sqrt{n}})$, the parametric rate of convergence. Detailed proofs of all the propositions in this article can be found in Appendix 3. Appendix 2 summaries the underlying regularity conditions and assumptions for the model.

Proposition 3.1 Under Assumptions 1-5 and the identification condition that, for any $\theta \neq \theta_0$,

$$
G[E(f(z_1, \beta(\theta)) | x), E(g(z_2, \gamma(\theta)) | x \delta(\theta), \theta), x, \alpha(\theta)] \neq 0
$$

with positive probability on $X$, $\hat{\theta}_I$ is a consistent estimator of $\theta_0$ and

$$
\sqrt{n}(\hat{\theta}_I - \theta_0) \xrightarrow{D} N(0, \Omega_I),
$$

where

$$
\Omega_I = \{ E (I_X (x) G_\theta [E(f(z_1, \beta_0) | x), E(g(z_2, \gamma_0) | x \delta_0), x, \alpha_0] G_\theta' [E(f(z_1, \beta_0) | x), E(g(z_2, \gamma_0) | x \delta_0), x, \alpha_0]) \}^{-1}
$$

and

$$
\Sigma_I = E \left[ \left( C_1(x) (f(z_1, \beta_0) - E(f(z_1, \beta_0) | x)) + C_2(x) (g(z_2, \gamma_0) - E(g(z_2, \gamma_0) | x \delta_0)) \right)^2 \right].
$$

The explicit expression for $G_\theta$ is derived in Appendices 2 and 3:
where $\nabla G[\cdot]$ denote the gradient vector of $G$ w.r.t. $E(f(z_1, \beta)|x)$, $E(g(z_2, \gamma)|x\delta, \theta)$, and $\alpha$. Here,

$$\frac{\partial E(g(z_2, \gamma)|x\delta, \theta)}{\partial \theta} = \frac{1}{B_j(x, \theta)} \frac{\partial A_j(x, \theta)}{\partial \theta} - \frac{A_j(x, \theta)}{B_j^2(x, \theta)} \frac{\partial B_j(x, \theta)}{\partial \theta},$$

with

$$A_j(x, \theta) = E(g(z_2, \gamma)|x\delta, \theta)p(x\delta|\theta),$$

$$B_j(x, \theta) = p(x\delta|\theta),$$

$$\frac{\partial B_j(x_i, \theta)}{\partial \theta_i} = (x_i - E(x|z_i, \delta, \theta)) \frac{\partial \delta}{\partial \theta_i} \frac{\partial p(x_i \delta|\theta)}{\partial \delta} + \text{tr} \left\{ \frac{\partial E(x|z_i \delta, \theta)}{\partial \theta_i} \frac{\partial \delta}{\partial \theta_i} \right\} p(x_i \delta|\theta),$$

and

$$\frac{\partial A_j(x_i, \theta)}{\partial \theta_i} = E \left( \frac{\partial g(z_2, \gamma)}{\partial \theta_i} | x_i \delta, \theta \right) p(x_i \delta|\theta) + E(g(z_2, \gamma)(x_i - z_i)|x_i \delta, \theta) \frac{\partial \delta}{\partial \theta_i} \frac{\partial p(x_i \delta|\theta)}{\partial \delta}$$

$$+ \text{tr} \left\{ \frac{\partial E(g(z_2, \gamma)(x_i - z)|x_i \delta, \theta)}{\partial \theta_i} \frac{\partial \delta}{\partial \theta_i} \right\} p(x_i \delta|\theta),$$

where $\theta_i$ denotes the $i$th component of $\theta$, and $p(x\delta|\theta)$ is the density function of $x\delta$ given $\theta$.

The SMD estimation of (2.5) can be regarded as a special case of (3.1). Let $\hat{\theta}_S$ denote the semiparametric estimator of the true parameter vector $\theta_0$ from (2.9) with $m_n = n$. Consistency and the asymptotic distribution of $\hat{\theta}_S$ follow from Proposition 3.1. Because (2.5) has simpler conditioning arguments, the asymptotic covariance matrix is simpler.

**Corollary 3.1** Under Assumptions 1-3 and the identification condition that, for any $\theta \neq \theta_0$,

$$G[E(f(z_1, \beta(\theta))|x), x, \alpha(\theta)] \neq 0$$

with positive probability on $X$, $\hat{\theta}_S$ is a consistent estimator of $\theta_0$ and

$$\sqrt{n}(\hat{\theta}_S - \theta_0) \stackrel{D}{\rightarrow} N(0, \Omega),$$

where

$$\Omega = \left\{ E(I_X(x)G_\theta[E(f(z_1, \beta_0)|x), x, \alpha_0] \cdot G_\psi[E(f(z_1, \beta_0)|x), x, \alpha_0]) \right\}^{-1}$$

$$\times \sum \left\{ E(I_X(x)G_\theta[E(f(z_1, \beta_0)|x), x, \alpha_0] \cdot G_\psi[E(f(z_1, \beta_0)|x), x, \alpha_0]) \right\}^{-1},$$

and

$$\Sigma = E \left\{ I_X(x)G_\theta[E(f(z_1, \beta_0)|x), x, \alpha_0] \cdot \nabla_1 G[E(f(z_1, \beta_0)|x), x, \alpha_0] \right\}$$

$$\times E \left\{ (f(z_1, \beta_0) - E(f(z_1, \beta_0)|x))(f(z_1, \beta_0) - E(f(z_1, \beta_0)|x))^T \right\}$$

$$\times \nabla_1 G[E(f(z_1, \beta_0)|x), x, \alpha_0] \cdot G_\psi[E(f(z_1, \beta_0)|x), x, \alpha_0].$$

It is interesting to point out an implication of Proposition 3.1 for the special case:

$$G[E(f(z_1, \beta_0)|x), E(g(z_2, \gamma_0)|x\delta_0), x, \alpha_0] = E(y|x) - F[E(f_1(z_1, \beta_0)|x), E(g(z_2, \gamma_0)|x\delta_0), x, \alpha_0],$$

where $f(z_1, \beta) = (y, f_1(z_1, \beta))$. For this case, the model is equivalent to

$$y = F[E(f_1(z_1, \beta_0)|x), E(g(z_2, \gamma_0)|x\delta_0), x, \alpha_0] + \epsilon, \quad E(\epsilon|x) = 0,$$  \hspace{1cm} (3.6)
which motivates an alternative estimation approach as

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} I_X(x_i) (y_i - F_\theta(E_n(f_1(z_i, \beta)|x_i), E_n(g(x_z, \gamma)|x_i, \alpha))^2.$$  \hfill (3.7)

By arguments similar to those in the proof of Proposition 3.1, the estimator $\hat{\theta}_A$ of $\theta$ from (3.7) is consistent and asymptotically normal. In addition, $\hat{\theta}_A$ has the same asymptotic distribution as the SMD estimator $\hat{\theta}_I$ of this model.

**Proposition 3.2** Under Assumptions 1-5 and the identification condition that, for any $\theta \neq \theta_0$,

$$F[E(f_1(z_1, \beta)|x), E(g(x_z, \gamma)|x, \alpha) \neq F[E(f_1(z_1, \beta_0)|x), E(g(x_z, \gamma_0)|x, \alpha_0)]$$

with positive probability on $X$, the semiparametric estimator $\hat{\theta}_A$ from (3.7) is consistent, and

$$\sqrt{n} (\hat{\theta}_A - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) F_\theta[E(f_1(z_i, \beta_0)|x_i), E(g(x_z, \gamma_0)|x_i, \alpha_0)] \right\}^{-1}$$

$$\times \left( \epsilon_i - \nabla \nabla_i F[E(f_1(z_i, \beta_0)|x_i), E(g(x_z, \gamma_0)|x_i, \alpha_0)] [f_1(z_i, \beta_0) - E(f_1(z_i, \beta_0)|x_i)] \right)$$

$$- E\left\{ I_X(x) F_\theta[E(f_1(z_1, \beta_0)|x), E(g(x_z, \gamma_0)|x_0, \alpha_0)] \mid x, \alpha_0 \right\} \left( [g(x_z, \gamma_0) - E(g(x_z, \gamma_0)|x_0, \alpha_0)] \right) + o_p(1).$$

Similarly, the model $E(y|x) = F[E(f_1(z_1, \beta_0)|x), x, \alpha_0]$ is equivalent to

$$y = F(E(f_1(z_1, \beta_0)|x), x, \alpha_0) + \epsilon, \quad E(\epsilon|x) = 0,$$  \hfill (3.8)

which can be estimated by

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} I_X(x_i) (y_i - F_\theta(E_n(f_1(z_i, \beta)|x_i), x_i, \alpha))^2.$$  \hfill (3.9)

The asymptotic distribution of the estimator $\hat{\theta}_{A,S}$ from (3.9) follows immediately from Proposition 3.2.

**Corollary 3.2** Under Assumptions 1-3 and the identification condition that, for any $\theta \neq \theta_0$,

$$F[E(f_1(z_1, \beta)|x), x, \alpha) \neq F[E(f_1(z_1, \beta_0)|x), x, \alpha_0]$$

with positive probability on $X$, the semiparametric estimator $\hat{\theta}_{A,S}$ from (3.9) is consistent, and

$$\sqrt{n} (\hat{\theta}_{A,S} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) F_\theta[E(f_1(z_i, \beta_0)|x_i), x_i, \alpha_0] \right\}^{-1}$$

$$\times \left( \epsilon_i - \nabla \nabla_i F[f_1(z_1, \beta_0)|x_i], x_i, \alpha_0] [f_1(z_i, \beta_0) - E(f_1(z_i, \beta_0)|x_i)] \right) + o_p(1).$$

For the estimation of a nonlinear structural equation in a simultaneous equation model $f(z_i, \beta) = u_i$, $i = 1, \cdots, n$, where $z$ contains both endogenous and exogenous variables, and $u$ is independent from all

7
exogenous variables \( z \) in the model and has a constant variance \( \sigma^2 \), a well-known estimation method is a nonlinear two-stage least-squares (NL2S) method (Amemiya [1974]):

\[
\min_\beta f'(Z, \beta) W (W'W)^{-1} W' f(Z, \beta),
\]

where \( f(Z, \beta) \) denotes the \( n \)-dimensional vector consisting of \( f(z_i, \beta) \), and \( W \) is some matrix of instrumental variables. Amemiya [1974] showed that the NL2S estimator is consistent and asymptotically normal, and, in general, the asymptotic covariance of the estimator depends on the choice of \( W \). Amemiya [1985] pointed out that the best nonlinear two-stage least-squares (BNL2S) estimator with a minimum asymptotic covariance matrix corresponds to \( W = E \left( \frac{\partial f(Z, \beta)}{\partial \beta} \right) | x \). The asymptotic covariance matrix of \( \sqrt{n} \) times the BNL2S estimator is

\[
\sigma^2 \left[ E \left( \frac{\partial f(Z, \beta)}{\partial \beta} \frac{\partial f(Z, \beta)}{\partial \beta} \right) \right]^{-1}.
\]

This model implies that \( E(f(z, \beta)|x) = 0 \), and \( G[E(f(z, \beta)|x, x, \alpha] = E(f(z, \beta)|x) \) in our format with \( \theta = \beta \). The model can be estimated by the SMD method. From Corollary 3.1, we see that the SMD estimator is asymptotically equivalent to a BNL2S estimator, since \( \nabla_1 G[f(z, \beta_0)|x, x, \alpha_o] = 1 \) and \( E[(f(z, \beta_0) - E(f(z, \beta_0)|x))^2|x] = \sigma^2 \) for this case. The SMD estimation is one of the feasible best nonlinear two-stage least-squares procedures. There are various suggestions in the econometric literature on the construction of feasible BNLS estimators, e.g., Amemiya [1985], Newey [1987, 1990], Andrews [1989], and Robinson [1989]. All these approaches suggested ways to construct some optimal instruments for \( W \). For this simultaneous equation model, the SMD estimation happens to be the same as an approach described briefly in Newey [1991].
4. Weighted Semiparametric Minimum-distance Estimation

Since the model in (2.5) is a special case of (2.10), the asymptotic distribution of \( \hat{\beta}_S \) is simpler than the asymptotic distribution of \( \hat{\beta}_I \) for the general model (2.10). The asymptotic distribution of \( \hat{\beta}_I \) might be simplified for some special cases. If \( G_s[\cdot] \) and \( \nabla_2 G[\cdot] \) were functions with \( x \) appearing only in the index form \( x \delta \) and trimming of the regressors could be done such that \( I_X(\cdot) \) were a function of \( x \delta \), then

\[
C_2(z) = E \left\{ I_X(z \delta_0)G_s[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0] \nabla_2 G[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0] \right\} = I_X(z \delta_0)G_s[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0] \nabla_2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0],
\]

and the asymptotic distribution of \( \hat{\beta}_I \) would be simplified as

\[
\sqrt{n}(\hat{\beta}_I - \theta_0) = - \frac{1}{n} \sum_{i=1}^{n} \left\{ I_X(z \delta_0)G_s[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0]G_s[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0] \right\}^{-1} \left( \nabla_1 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0] \right) \cdot (f(z_1, \beta_0) - E(f(z_1, \beta_0)|z_1), g(z_2, \gamma_0) - E(g(z_2, \gamma_0)|z_1 \delta_0))' + o_p(1).
\]

For some special models, when \( I_X(\cdot) \) is a function of \( x \delta \), it might be possible that \( C_2(z_i) = 0 \). A good example of this case is the Ichimura binary choice index model (Ichimura [1987]). Under such circumstances,

\[
\sqrt{n}(\hat{\beta}_I - \theta_0) = - \frac{1}{n} \sum_{i=1}^{n} \left\{ I_X(z \delta_0)G_s[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0]G_s[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|z \delta_0), x, \alpha_0] \right\}^{-1} \left( \nabla_1 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0] \right) \cdot (f(z_1, \beta_0) - E(f(z_1, \beta_0)|z_1), g(z_2, \gamma_0) - E(g(z_2, \gamma_0)|z_1 \delta_0))' + o_p(1).
\]

For these special models, a weighted semiparametric MD estimation procedure may be possible. The trimming of the indices is, however, a tedious and technically complicated problem. While there are suggestions (see, e.g., Manski [1985], Robinson [1988], Klein and Spady [1987], and Lee [1991]) for different semiparametric models, the technical difficulty remains. So far, the technical complication of trimming can be overcome only in the adaptive estimation approach (see Manski [1985]) and in certain two step estimation methods. For the estimation of the general model (2.10), a properly weighted semiparametric MD approach does not seem to be possible. For these reasons, we limit our attention in this section to the estimation of the equation (2.5), for which a weighted semiparametric MD estimator can be derived.

From Corollary 3.1, we see that \( G[E_n(f(z_1, \beta_0)|z_1), x_i, \alpha_0] \) might be treated asymptotically as if it were a random variable \( u_i \), where \( u_i = \nabla_1 G[E(f(z_1, \beta_0)|z_1), x_i, \alpha_0][f(z_1, \beta_0) - E(f(z_1, \beta_0)|z_1)] \). Conditional on \( x_i \), \( u_i \) is a heteroskedastic disturbance with a variance \( v(x_i) \), where

\[
v(x_i) = \nabla_1 G[E(f(z_1, \beta_0)|z_1), x_i, \alpha_0] \sigma^2(x_i) \nabla_1 G[E(f(z_1, \beta_0)|z_1), x_i, \alpha_0],
\]

with \( \sigma^2(x_i) = E[f(z_1, \beta_0)f'(z_1, \beta_0)|z_i] - E(f(z_1, \beta_0)|z_i)E(f'(z_1, \beta_0)|z_i) \). A relatively efficient estimation procedure shall take into account this heteroskedasticity. With a \( \sqrt{n} \)-consistent estimate \( \tilde{\theta} \), \( \sigma^2(x_i) \) can be estimated by

\[
\tilde{\sigma}_n^2(x_i, \tilde{\theta}) = E_n[f(z_1, \tilde{\theta})f'(z_1, \tilde{\theta})|z_i] - E_n(f(z_1, \tilde{\theta})|z_i)E_n(f'(z_1, \tilde{\theta})|z_i),
\]

and \( v(x_i) \) can be estimated by

\[
v_n(x_i, \tilde{\theta}) = \nabla_1 G[E_n(f(z_1, \tilde{\theta})|z_i), x_i, \tilde{\alpha}] \tilde{\sigma}_n^2(x_i) \nabla_1 G[E_n(f(z_1, \tilde{\theta})|z_i), x_i, \tilde{\alpha}],
\]

for some special models.
where \( \hat{\alpha} = \alpha(\hat{\theta}) \) and \( \hat{\beta} = \beta(\hat{\theta}) \). A feasible weighted semiparametric minimum-distance (WSMD) estimation procedure is

\[
\min_{\theta \in \Theta} \sum_{i=1}^{n} I(x_i) \frac{1}{v_n(x_i, \theta)} G^2[E_n(f(z_1, \beta(\theta)) | x_i), x_i, \alpha(\theta)].
\]  
(4.1)

Proposition 4.1 Under Assumptions 1-3 and \( \delta \) and the identification condition in Corollary 3.1, the WSMD estimator \( \hat{\theta}_w \) from (4.1) is consistent, and

\[
\sqrt{n}(\hat{\theta}_w - \theta_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{1}{v(x_i)} G_\theta [E(f(z_1, \beta_0) | x_i), x_i, \alpha_0] G_\theta [E(f(z_1, \beta_0) | x_i), x_i, \alpha_0] \right)^{-1} \times \left( (f(z_1, \beta_0) - E(f(z_1, \beta_0) | x_i)) + o_p(1) \right).
\]

Consequently, \( \sqrt{n}(\hat{\theta}_w - \theta_0) \to D N(0, \Omega) \), where

\[
\Omega = \left\{ E \left( \frac{1}{v(x)} G_\theta [E(f(z_1, \beta_0) | x), x, \alpha_0] G_\theta [E(f(z_1, \beta_0) | x), x, \alpha_0] \right)^{-1} G_\theta [E(f(z_1, \beta_0) | x), x, \alpha_0] \right\}.
\]

The WSMD procedure can be easily generalized to the estimation of a system of equations with \( G \) being a finite dimensional vector-valued function. In this case, \( v_n(x_i, \theta) \) becomes a matrix and the WSMD procedure will be generalized to

\[
\min_{\theta \in \Theta} \sum_{i=1}^{n} I(x_i) G^2[E_n(f(z_1, \beta(\theta)) | x_i), x_i, \alpha(\theta)][v_n(x_i, \theta)]^{-1} G[E_n(f(z_1, \beta(\theta)) | x_i), x_i, \alpha(\theta)].
\]

Let \( \hat{\theta}_{M,w} \) be the systematic WSMD estimator. The limiting distribution of \( \sqrt{n}(\hat{\theta}_{M,w} - \theta_0) \) will be \( N(0, \Omega_M) \), where

\[
\Omega_M = \left\{ E \left( \frac{1}{v(x)} G_\theta [E(f(z_1, \beta_0) | x), x, \alpha_0] [v(x)]^{-1} G_\theta [E(f(z_1, \beta_0) | x), x, \alpha_0] \right)^{-1} \right\},
\]

and \( v(x) = \nabla^2 G[E(f(z_1, \beta_0) | x), x, \alpha_0] \sigma^2(x) \nabla^2 G[E(f(z_1, \beta_0) | x), x, \alpha_0] \).

Comparing the asymptotic distribution of the SMD estimator \( \hat{\theta}_s \) with the asymptotic distribution of the WSMD estimator \( \hat{\theta}_w \), we see that \( \hat{\theta}_w \) is asymptotically efficient relative to \( \hat{\theta}_s \). Asymptotic efficiency issues in the estimation of models with conditional moment restrictions have been studied in Chamberlain [1987]. Chamberlain considers the model with conditional moment restrictions:

\[
E[f(z_1, \theta_0) | x] = 0,
\]
(4.2)

where \( f \) has a known functional form; and he finds that for any \( \sqrt{n} \)-consistent estimator \( \hat{\theta} \) of \( \theta_0 \), the random variable \( \sqrt{n}(\hat{\theta} - \theta_0) \) can be no more concentrated about zero than a random variable distributed \( N(0, \Omega) \), where

\[
\Omega = \left\{ E \left( \frac{\partial f(x_1, \theta_0)}{\partial \theta} \right) \left[ E(f(x_1, \theta_0) f(x_1, \theta_0) | x) \right]^{-1} E \left( \frac{\partial f(x_1, \theta_0)}{\partial \theta} | x \right) \right\}^{-1}.
\]

For this case, we see that the WSMD estimator \( \hat{\theta}_{M,w} \) would attain the efficiency bound found by Chamberlain if trimming of \( x \) were absent. The model (4.2) is essentially a nonlinear simultaneous equation model with unknown heteroskedasticity. It includes the linear regression model with unknown heteroskedasticity as a special case. For the linear regression model, Robinson [1987] provides an estimator which attains Chamberlain's efficiency bound. Newey [1987, 1990] reports an estimator that attains the bound for the conditional moment restriction model (4.2). Newey's approach suggests procedures to construct an optimal
instrumental variables matrix for estimation. Chamberlain's efficiency bound is constructed for the model in (4.2). For the general equation model (2.10), the efficiency issue remains to be investigated.

As an illustration, it is interesting to demonstrate a WSMD estimation procedure for a QR model. Consider a binary choice model

\[ y = P(z \alpha + \epsilon), \quad (4.4) \]

where \( y \) is a dichotomous response and \( P \) is the conditional choice probability. For this model, \( E(\epsilon|x) = 0 \) and \( \text{var}(\epsilon|x) = P(z \alpha)[1 - P(z \alpha)] \). As a popular parameter probability model, \( P \) might be a logit or probit probability function. If \( P \) is assumed to be invertible, then (4.4) implies that \( P^{-1}[E(y|x)] = z \alpha \). With this relation, \( G[E(y|x), x, \alpha] = P^{-1}[E(y|x)] - z \alpha \). It follows that \( G_{\alpha}[E(y|x), x, \alpha] = -x' \) and \( \nabla_{1} G[E(y|x), x, \alpha] = \frac{1}{p[P^{-1}[E(y|x)]]} \), where \( p \) is the probability density corresponding to \( P \). The WSMD estimator of \( \alpha \) is

\[ \hat{\alpha}_w = \left( \sum_{i=1}^{n} \frac{1}{v_n(x_i)} I_X(x) x'_i \right)^{-1} \sum_{i=1}^{n} \frac{1}{v_n(x_i)} x'_i P^{-1}[E_n(y|x_i)], \]

where the weight \( v_n(x_i) = v_n(x_i) = (E_n(y^2|x_i) - E_n^2(y|x_i)) \frac{1}{p[P^{-1}[E_n(y|x_i)]]} \). The asymptotic covariance matrix of \( \sqrt{n}(\hat{\alpha}_w - \alpha_0) \) is \( E(I_X(x) x'_i \text{var}(\epsilon|x)) \frac{1}{p^2[P^{-1}[E(y|x)]]^2} x'_i \text{var}(\epsilon|x)^{-1} \). This WSMD estimator would be asymptotically efficient if there were no trimming, as is the maximum likelihood estimator. An interesting point is that this asymptotically efficient estimator has a closed form expression, and its computation does not require an iterative algorithm.
5. Some Related Properties

5.1 Invariance Properties

The SMD estimation procedures possess some invariant properties. Suppose that \( \phi \) is a differentiable origin-preserving transformation, i.e., \( \phi(x) = 0 \) if and only if \( x = 0 \), and \( \frac{d\phi(z)}{dz} \neq 0 \). The model (2.10) will be equivalent to

\[
\Phi[E(f(z_1, \beta_0)|x), E(g(z_2, \gamma_0)|x, \alpha_0)] = 0,
\]

where \( \Phi = \phi \circ G \) is a composite function of \( \phi \) and \( G \). Let \( \tilde{\theta}_\phi \) be the SMD estimator from

\[
\min_{\theta \in \Theta} \sum_{i=1}^{n} I_X(x_i) \Phi^2[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_1, \beta), z_i, \alpha].
\]

Let \( v(x, \theta_o) = (E(f(z_1, \beta_0)|x), E(g(z_2, \gamma_0)|x, \alpha_0) \) denote the vector of arguments in the above functions. Since \( \Phi[v(x, \theta_o)] = \nabla \phi(0) \cdot G \Phi[v(x, \theta_o)] \), \( \nabla_1 \Phi[v(x, \theta_o)] = \nabla \phi(0) \cdot \nabla_1 G[v(x, \theta_o)] \), and \( \nabla_2 \Phi[v(x, \theta_o)] = \nabla \phi(0) \cdot \nabla_2 G[v(x, \theta_o)] \), it follows from Proposition 3.1 that

\[
\sqrt{n}(\tilde{\theta}_\phi - \theta_o) = -\{E(I_X(x) \Phi[v(x, \theta_o)] \Phi^2[v(x, \theta_o)])\}^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) \Phi[v(x_i, \theta_o)] \cdot \nabla_1^t \Phi[v(x_i, \theta_o)] \right\}
\times (f(z_1, \beta_0) - E(f(z_1, \beta_0)|x_i)) + E(I_X(x_i) \Phi[v(x_i, \theta_o)] \nabla_2^t \Phi[v(x_i, \theta_o)] | x_i, \theta_o)
\times (g(z_2, \gamma_0) - E(g(z_2, \gamma_0)|x_i, \theta_o)) + o_p(1)
\]

which has the same asymptotic distribution as \( \sqrt{n}(\tilde{\theta}_\phi - \theta_o) \) in Proposition 3.1. Hence the SMD estimator is asymptotically invariant with respect to the origin-preserving transformation \( \phi \). A similar conclusion holds for the WSMD estimator.

Consider a special model with \( G \) as a sum of two functions:

\[
G[E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x, \alpha)] = G_1[E(f_1(z_1, \beta)|x), E(g_1(z_2, \gamma)|x, \alpha)] - G_2[E(f_2(z_1, \beta)|x), E(g_2(z_2, \gamma)|x, \alpha)].
\]

The structural model is

\[
G_1[E(f_1(z_1, \beta_0)|x), E(g_1(z_2, \gamma_0)|x, \alpha_0)] = G_2[E(f_2(z_1, \beta_0)|x), E(g_2(z_2, \gamma_0)|x, \alpha_0)].
\]

Let \( \psi \) be a strictly monotonic function. This model is equivalent to

\[
\psi \circ G_1[E(f_1(z_1, \beta_0)|x), E(g_1(z_2, \gamma_0)|x, \alpha_0)] = \psi \circ G_2[E(f_2(z_1, \beta_0)|x), E(g_2(z_2, \gamma_0)|x, \alpha_0)].
\]

Denote \( \Psi \) by

\[
\Psi[E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x, \alpha)] = \psi \circ G_1[E(f_1(z_1, \beta)|x), E(g_1(z_2, \gamma)|x, \alpha)] - \psi \circ G_2[E(f_2(z_1, \beta)|x), E(g_2(z_2, \gamma)|x, \alpha)].
\]

12
The SMD estimation method can be applied to the function $\Psi$ instead of $G$. To simplify notation for this case, let $v(x, \theta) = \langle E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x, \alpha), v_1(x, \theta) = \langle E(f_1(z_1, \beta)|x), E(g_1(z_2, \gamma)|x, \alpha), \rangle$ and $v_2(x, \theta) = \langle E(f_2(z_1, \beta)|x), E(g_2(z_2, \gamma)|x, \alpha)\rangle$. Since $v_1(x, \theta_0) = v_2(x, \theta_0)$,

$$\Psi_{\theta}[v(x, \theta_0)] = \nabla \psi(G_1[v_1(x, \theta_0)]\{G_{1,\theta}[v_1(x, \theta_0)] - G_{2,\theta}[v_2(x, \theta_0)]\} = \nabla \psi(G_1[v_1(x, \theta_0)]\{G_{2,\theta}[v_2(x, \theta_0)]\},$$

and $\nabla \Psi_{\theta}[v(x, \theta_0)] = \nabla \psi(G_1[v_1(x, \theta_0)])\nabla G[v(x, \theta_0)]$. From these relations, we see that the WSMD estimation procedure is asymptotically invariant with respect to any strictly monotonic transformation because the factor $\nabla \psi(G_1[v_1(x, \theta_0)])$ has been cancelled out by the weighting scheme in the asymptotic covariance matrix. However, the unweighted SMD procedure may not be asymptotically invariant with respect to a monotonic transformation. The following example provides an illustration. Consider a nonlinear regression model: $y = f(x\alpha) + \epsilon$, $\epsilon|x = 0$, where $x\alpha$ appears as an index and $f$ is invertible. This model implies two possible formulations for $G$, say, $G^{(1)}$ and $G^{(2)}$: $G^{(1)}[y|x, \alpha] = E(y|x) - f(x\alpha)$, and $G^{(2)}[y|x, \alpha] = f^{-1}(E(y|x)) - x\alpha$. Let $\hat{\alpha}_1$ be the SMD estimator derived with $G^{(1)}$, and let $\hat{\alpha}_2$ be the SMD estimator derived with $G^{(2)}$. Corollary 3.1 implies that

$$\sqrt{n}(\hat{\alpha}_1 - \alpha_0) = \left\{E\left(I_x(x) \frac{\partial f(x\alpha_0)}{\partial \alpha} \frac{\partial f(x\alpha_0)}{\partial \alpha'}\right)\right\}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_x(x_i) \frac{\partial f(x_i\alpha_0)}{\partial \alpha} \epsilon_i + o_p(1),$$

which would be asymptotically equivalent to the standard nonlinear regression estimator if there were no trimming. The SMD estimator $\hat{\alpha}_2$ is computationally simpler as it does not involve an iterative algorithm. It has the following asymptotic distribution from Corollary 3.1:

$$\sqrt{n}(\hat{\alpha}_2 - \alpha_0) = \left\{E(I_x(x)\nabla f(x\alpha_0)^2)\right\}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_x(x_i) \nabla f(x_i\alpha_0) x_i' \epsilon_i + o_p(1),$$

where $\nabla f(\cdot)$ denotes the derivative of $f$. The asymptotic distributions of the two estimators are apparently different. On the other hand, the WSMD estimators have the same asymptotic distributions. Suppose that the disturbance $\epsilon$ is homoskedastic. In this case, weighting of the SMD estimation with $G$. To simplify notation for this case, let $G_1(x, \alpha)$ be some specified values of $x$ and let $G_2(x, \alpha)$ be the SMD estimator derived with $G^{(1)}$. Corollary 4.1 implies that

$$\sqrt{n}(\hat{\alpha}_{2, w} - \alpha_0) = \left\{E(I_x(x)\nabla f(x\alpha_0)^2)\right\}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_x(x_i) \nabla f(x_i\alpha_0) x_i' \epsilon_i + o_p(1),$$

which is asymptotically equivalent to $\hat{\alpha}_1$ because $\frac{\partial f(x\alpha_0)}{\partial \alpha} = \nabla f(x\alpha_0) \cdot x'$.

### 5.2 A Local Goodness-of-Fit Test Statistic

In Appendix 4, we demonstrate the similarity of the SMD estimation procedure to the minimum chi-square procedure when some finite restrictions are used for estimation. The minimized weighted-distance function multiplied by a proper rate of convergence is asymptotically distributed chi-square. This can be used as a goodness-of-fit test statistic. Unfortunately, when there are infinitely many restrictions, the corresponding minimized-distance function might not be proportional to a chi-square random variable with a finite degree of freedom. Indeed, it is not clear what its limiting distribution is in general. A straightforward generalization of the statistic of Appendix 4 is, however, possible for a local goodness-of-fit test.

Let $\tilde{\theta}$ be the SMD estimator from Section 3. Let $x_l, l = 1, \ldots, L$, be some specified values of $x$ lying inside the interior of the support of $x$. By a Taylor expansion of $G$ at $\theta_0$,

$$G[\mathbb{E}_n(f(z_1, \beta)|x_1), \mathbb{E}_n(g(z_2, \gamma)|x_1, \alpha)]$$

$$= G[\mathbb{E}_n(f(z_1, \beta_0)|x_1), \mathbb{E}_n(g(z_2, \gamma_0)|x_1, \alpha_0)] + \mathbb{E}_n[G[\mathbb{E}_n(f(z_1, \beta)|x_1), \mathbb{E}_n(g(z_2, \gamma)|x_1, \alpha)](\tilde{\theta} - \theta_0)$$

$$= \nabla f[\mathbb{E}_n(f(z_1, \beta_0)|x_1), \mathbb{E}_n(g(z_2, \gamma_0)|x_1, \alpha_0)]\times$$

$$\times [E_n(f(z_1, \beta)|x_1) - E(f(z_1, \beta_0)|x_1), E_n(g(z_2, \gamma)|x_1, \alpha_0) - E(g(z_2, \gamma_0)|x_1, \alpha_0)]'$$

$$+ G[\mathbb{E}_n(f(z_1, \beta)|x_1), \mathbb{E}_n(g(z_2, \gamma)|x_1, \alpha)](\tilde{\theta} - \theta_0).$$

13
As $\hat{\theta}$ is $\sqrt{n}$-consistent,

$$
\sqrt{n}a_n^k G[E_n(f(z_1, \beta(\hat{\theta}))[x_i]), E_n(g(z_2, \gamma(\hat{\theta}))[x_i\delta(\hat{\theta})]), x_i(\alpha(\hat{\theta}))]
$$

$$
= \nabla_1^2 G[E_n(f(z_1, \beta_0)[x_i]), E_n(g(z_2, \gamma_0)[x_i\delta_0]), x_i, \alpha_0]
$$

$$
\times \left[ \sqrt{n}a_n^k (E_n(f(z_1, \beta_0)[x_i]) - E(f(z_1, \beta_0)[x_i])), \sqrt{n}b_n^m (E_n(f(z_1, \beta_0)[x_i]) - E(f(z_1, \beta_0)[x_i])) \right] \\
+ G[E_n(f(z_1, \beta_0)[x_i]), E_n(g(z_2, \gamma_0)[x_i\delta_0]), x_i, \alpha_0] \sqrt{n} \cdot \sqrt{n}(\hat{\theta} - \theta_0)
$$

$$
= \nabla_1^2 G[E(f(z_1, \beta_0)[x_i]), E(g(z_2, \gamma_0)[x_i\delta_0]), x_i, \alpha_0] \sqrt{n}a_n^k (E_n(f(z_1, \beta_0)[x_i]) - E(f(z_1, \beta_0)[x_i])) + o_p(1),
$$

under the assumption that $\lim_{n \to \infty} \frac{a_n^k}{b_n^m} = 0$ (which is in general the case in practice as $m < k$). Define

$$
v_n(z_i) = \nabla_1^2 G[E_n(f(z_1, \beta)[x_i]), E_n(g(z_2, \gamma)[x_i\delta]), x_i, \hat{\alpha}] \frac{1}{B_n(z_i)} \int K^2(v)dv
$$

$$
\times \left[ E_n(f^2(z_1, \beta)[x_i]) - E^2(f(z_1, \beta)[x_i]) \right] \nabla_1^2 G[E_n(f(z_1, \beta)[x_i]), E_n(g(z_2, \gamma')[x_i\delta]), x_i, \hat{\alpha}].
$$

Since $\sqrt{n}a_n^k (E_n(f(z_1, \beta_0)[x_i]) - E(f(z_1, \beta_0)[x_i]))$ converges in distribution to $N(0, \Sigma)$, where

$$
\Sigma = \frac{1}{h(z_i)} \left[ E(f^2(z_1, \beta_0)[x_i]) - E^2(f(z_1, \beta_0)[x_i]) \right] \cdot \int K^2(u)du,
$$

it follows that $\sqrt{n}a_n^k \sum_{i=1}^{L} v_n(z_i) G[E_n(f(z_1, \beta(\hat{\theta}))[x_i]), E_n(g(z_2, \gamma(\hat{\theta}))[x_i\delta(\hat{\theta})]), x_i, \alpha(\hat{\theta})] \xrightarrow{D} N(0, 1)$. A local goodness-of-fit test statistic can be constructed as

$$
na_n^k \sum_{i=1}^{L} v_n(z_i) G^2[E_n(f(z_1, \beta(\hat{\theta}))[x_i]), E_n(g(z_2, \gamma(\hat{\theta}))[x_i\delta(\hat{\theta})]), x_i, \alpha(\hat{\theta})],
$$

which is asymptotically distributed chi-square with $L$ degrees of freedom. This statistic has $L$ degrees of freedom because $\hat{\theta}$ converges at a faster rate than the nonparametric regression functions and the slow convergent random variables dominate the asymptotic distribution. This test statistic is useful for detecting the goodness-of-fit of the model at some specified points, such as the sample mean or median, which are regarded as important points by an empirical investigator. However, if it were regarded as an over-all goodness-of-fit statistic, it would suffer from the arbitrary number of restrictions for the test. Subsequent research should consider how to utilize all the information to derive more powerful test statistics. Much research needs to be done on the development of goodness-of-fit statistics in the semiparametric framework.

---

1 It is possible to derive test statistics with a faster rate of convergence. For example, at a point $z_i$, take a sample average of $G$ over a neighborhood of $z_i$ with fixed diameter. A $\sqrt{n}$ multiplication of such an average will converge in distribution to a normal variable. An asymptotic chi-square statistic can be constructed with this average value. However, the computation of such a statistic will be much complicated. We leave the possible approach for future investigation.
6. Semiparametric Estimation of Simultaneous Equation Microeconometric Models

In this section, we point out briefly possible applications of the SMD approach to the estimation of some semiparametric simultaneous equation microeconometric models, namely, endogenous switching regression models and simultaneous equation limited dependent variables models based on index restrictions. Many other models may also be estimated with the SMD estimation framework. We will demonstrate briefly how some conditional moment equations can be derived from a specific semiparametric model with index restrictions. Detailed analysis on any specific model and its identification conditions will not be reported here due to space limitation.

6.1 SEMIPARAMETRIC ENDOGENOUS SWITCHING REGRESSION MODELS

Consider a model with two regimes:

\[ y_1 = x_1 \alpha_1 + \epsilon_1, \quad (6.1.1) \]

and

\[ y_2 = x_2 \alpha_2 + \epsilon_2, \quad (6.1.2) \]

where \( y_1 \) can be observed only when \( x_3 \delta > u \) and \( y_2 \) can be observed only when \( x_3 \delta < u \). In this framework, \( y \) is the observed value of the dependent variable, and \( x_1, x_2, \) and \( x_3 \) are subvectors of \( x \). Let \( I \) be a regime indicator with \( I = 1 \) for the first regime, and \( I = 0 \) for the second regime. There are two possible situations depending on whether \( I \) is observable or not. If \( I \) is observable, the model is a switching regression model with sample separation information. If \( I \) is not observable, it is a model without sample separation information. Assume either that the disturbances \( \epsilon_1, \epsilon_2, \) and \( u \) are independent of \( x \), or their distributions conditional on \( x \) are only a function of the index \( \alpha_3 \).

This model implies that

\[ y = Ix_1 \alpha_1 + (1 - I)x_2 \alpha_2 + \epsilon_1 + (1 - I) \epsilon_2. \quad (6.1.3) \]

For the model with sample separation information, (6.1.3) implies

\[ E(y|x) = E(I|x)x_1 \alpha_1 + E((1 - I)|x)x_2 \alpha_2 + E(I \epsilon_1 + (1 - I) \epsilon_2|x), \quad (6.1.4) \]

and

\[ E(y|x_3 \delta) = E(Ix_1|x_3 \delta) \alpha_1 + E((1 - I)x_2|x_3 \delta) \alpha_2 + E(I \epsilon_1 + (1 - I) \epsilon_2|x_3 \delta). \quad (6.1.5) \]

By the index property of the model, \( E(I \epsilon_1 + (1 - I) \epsilon_2|x) \) can only be a function of \( x_3 \delta \); hence, \( E(I \epsilon_1 + (1 - I) \epsilon_2|x) = E(I \epsilon_1 + (1 - I) \epsilon_2|x_3 \delta). \) Therefore,

\[ E(y|x) - E(y|x_3 \delta) = \{E(I|x)x_1 - E(Ix_1|x_3 \delta)\} \alpha_1 + \{E((1 - I)|x)x_2 - E((1 - I)x_2|x_3 \delta)\} \alpha_2. \quad (6.1.6) \]

The sample separation indicator implies also that

\[ E(I|x) = E(I|x_3 \delta). \quad (6.1.7) \]

The parameters in the model can be estimated by the SMD procedure applied to (6.1.6) and (6.1.7).

For the model without sample separation information, (6.1.3) implies

\[ E(y|x) = x_2 \alpha_2 + E(I|x)(x_1 \alpha_1 - x_2 \alpha_2) + E(\epsilon_2|x) + E(I(\epsilon_1 - \epsilon_2)|x), \quad (6.1.8) \]

and

\[ E(y|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2) \]

\[ = E(x_2|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2) \alpha_2 + E(I|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2)(x_1 \alpha_1 - x_2 \alpha_2) \]

\[ + E(\epsilon_2 + I(\epsilon_1 - \epsilon_2)|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2). \quad (6.1.9) \]

Since \( E(I|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2) = E(I|x) \), and \( E(\epsilon_2 + I(\epsilon_1 - \epsilon_2)|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2) = E(\epsilon_2 + I(\epsilon_1 - \epsilon_2)|x) \), the difference of (6.1.8) and (6.1.9) is

\[ E(y|x) - E(y|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2) = [x_2 - E(x_2|x_3 \delta, x_1 \alpha_1 - x_2 \alpha_2)] \alpha_2. \quad (6.1.10) \]
The SMD method can be applied to (6.1.10).

Disequilibrium market models can be regarded as special cases of the above models. The simplest disequilibrium market model consists of a demand equation and a supply equation where the price is fixed and the traded quantity is determined by the short side, i.e., \( y = \min\{y_1, y_2\} \). This model implies that

\[
E(y|x) - E(y|x_1\alpha_{1,0} - x_2\alpha_{2,0}) = [x_2 - E(x_2|x_1\alpha_{1,0} - x_2\alpha_{2,0})]\alpha_{2,0}, \tag{6.1.11}
\]

which can be used for the SMD estimation.

The estimation procedure can be easily generalized to the estimation of models with a Box-Cox transformation. For example, instead of (6.1.1) and (6.1.2), the equations are

\[
\frac{y^\lambda - 1}{\lambda} = [E(I|x)x_1 - E(I|x_1|x_3\delta_0)]\alpha_{1,0} + [E((1-I)|x)x_2 - E((1-I)x_2|x_3\delta_0)]\alpha_{2,0},
\]

which can be estimated within the SMD framework.

### 6.2 Semiparametric Simultaneous Equation Models with Limited Dependent Variables

There are various cases of such models. Consider the following model with mixed discrete and limited dependent variables:

\[
y^*_1 = x_1\delta_{1,0} + \epsilon_1, \quad y^*_2 = I_1\alpha_0 + x_2\delta_{2,0} + \epsilon_2,
\]

where \( y^*_1 \) is a latent variable with dichotomous indicator \( I_1 \), \( y^*_2 \) is a limited dependent variable determined by the latent variable \( y^*_1 \) as \( y_2 = \max\{y^*_1, 0\} \). Either the disturbance \( \epsilon_1 \) is independent of \( x \), or its distribution depends only on \( x_1\delta_{1,0} \); either \( \epsilon_2 \) is independent of \( x \), or its distribution depends on \( x_1\delta_{1,0} \) and/or \( x_2\delta_{2,0} \). Let \( I_2 \) be a dichotomous indicator: \( I_2 = 1 \) if and only if \( y^*_2 > 0 \). The first equation implies that

\[
E(I_1|x) = E(I_1|x_1\delta_{1,0}). \tag{6.2.1}
\]

The second equation implies that \( I_2y_2 = I_2I_1\alpha_0 + I_2x_2\delta_{2,0} + I_2\epsilon_2 \). Since \( E(I_2\epsilon_2|x) \) can only be a function of \( x_1\delta_{1,0} \) and \( x_2\delta_{2,0} \), it follows that

\[
E(I_2y_2|x) - E(I_2y_2|x_1\delta_{1,0}, x_2\delta_{2,0}) = [E(I_2I_1|x) - E(I_2I_1|x_1\delta_{1,0}, x_2\delta_{2,0})]\alpha_0 + [E(I_2|x)x_2 - E(I_2x_2|x_1\delta_{1,0}, x_2\delta_{2,0})]\delta_{2,0}. \tag{6.2.2}
\]

The SMD procedure can be applied to the estimation of (6.2.1) and (6.2.2).

Simultaneous equations with limited dependent variables can be estimated within this framework. Consider a two equation model:

\[
y^*_1 = x_1\delta_{1,0} + \epsilon_1, \quad y^*_2 = y_1\alpha_0 + x_2\delta_{2,0} + \epsilon_2,
\]

where \( y^*_l \), \( l = 1, 2 \) are latent variables and the observed dependent variables are \( y_l = \max\{y^*_l, 0\}, l = 1, 2 \). The distribution of \( \epsilon_1 \) conditional on \( x \) is assumed to depend only on \( x_1\delta_{1,0} \) and the distribution of \( \epsilon_2 \) depends on \( x_1\delta_{1,0} \) and/or \( x_2\delta_{2,0} \). Let \( I_1 \) denote the dichotomous indicator of \( y^*_1 \): \( I_1 = 1 \) if and only if \( y^*_1 > 0 \). Similarly, let \( I_2 \) be the dichotomous indicator for \( y^*_2 \). The first equation implies that

\[
E(y_1|x) - E(y_1|x_1\delta_{1,0}) = [E(I_1|x)x_1 - E(I_1x_1|x_1\delta_{1,0})]\delta_{1,0}, \tag{6.2.3}
\]

and the second equation implies that

\[
E(y_2|x) - E(y_2|x_1\delta_{1,0}, x_2\delta_{2,0}) = [E(I_2y_1|x) - E(I_2y_1|x_1\delta_{1,0}, x_2\delta_{2,0})]\alpha_0 + [E(I_2|x)x_2 - E(I_2x_2|x_1\delta_{1,0}, x_2\delta_{2,0})]\delta_{2,0}. \tag{6.2.4}
\]

These equations (6.2.3) and (6.2.4) can be used for the SMD estimation.
7. Possible Generalizations

In the previous sections, we considered the SMD estimation of equations involving nonparametric regression functions with or without index restrictions. The SMD methods can in principle be generalized to the estimation of equations involving other nonparametric functions. As an illustration, we discuss a case of the estimation of sample selection models subject to Tobit selection rule described in Lee [1990]. This example is interesting because the SMD method provides a different view of looking at the estimation and identification of such a model.

The model in Lee [1990] is a model with two equations:

\[ y_i^* = \alpha_0 + u, \quad (7.1) \]

and

\[ y_2^* = \beta_0 + v, \quad (7.2) \]

where \( y_i^* \) is a Tobit regression model with observation \( y_1 = \max\{0, y_i^*\} \), and \( y_2^* \) is a latent outcome equation with \( y_2 = y_2^* \) being observed only when \( y_i^* > 0 \). Conditional on \( y_2 \) being observed, we have

\[ E(y_2|y_1 > 0, x) = \beta_0 + E(v|u > -\alpha_0). \quad (7.3) \]

The estimation method in Lee [1990] is based on the assumption that \( u \) and \( v \) in (7.1) and (7.2) are independent of the regressor \( x \). Under this independence assumption, Lee [1990] suggested a two-stage method for the estimation of (7.2). Let \( f(v, u) \) be the joint density of \( (v, u) \), and \( f_u(-) \) be the marginal density of \( u \). It follows by definition that

\[
E(v|u > -\alpha_0) = \frac{\int_{-\infty}^{\infty} \int_{-\alpha_0}^{\infty} v f(v, u) dudv}{\int_{-\alpha_0}^{\infty} f_u(t) dt},
\]

where \( f(v, u) \) is the density of \( x\alpha_0 \). Given a consistent estimate \( \hat{\alpha} \) from some semiparametric estimation of (7.1), Lee’s method is a semiparametric least-squares procedure:

\[
\min_{\beta} \sum_{i=1}^{n} I_X(x_i) \left( y_{2i} - x_i \beta - \frac{A_n(z_{i1}, \beta)}{B_n(x_i)} \right), \quad (7.4)
\]

where

\[
A_n(z_{i1}, \beta) = \int_{-x_i \hat{\alpha}}^{\infty} \int_{-x_i \hat{\alpha}}^{\infty} \frac{1}{(n-1)a^2_n} \sum_{j \neq i} y_j(\beta) K \left[ \frac{u - u_j(\hat{\alpha})}{a_n}, \frac{z - x_j \hat{\alpha}}{a_n} \right] dudz,
\]

and

\[
B_n(x_i) = \int_{-x_i \hat{\alpha}}^{\infty} \int_{-x_i \hat{\alpha}}^{\infty} \frac{1}{(n-1)a^2_n} \sum_{j \neq i} K \left[ \frac{u - u_j(\hat{\alpha})}{a_n}, \frac{z - x_j \hat{\alpha}}{a_n} \right] dudz,
\]

with \( u_j(\alpha) = y_{1j} - x_j \alpha \) and \( v_j(\beta) = y_{2j} - x_j \beta \). This method differs from Ichimura’s semiparametric least-squares method in the construction of a nonparametric estimate of the conditional expectation \( E(v|u > -x\alpha_0) \). Ichimura’s method used only the index restriction in the model. The estimates \( A_n \) and \( B_n \) above indirectly estimate the density functions of the model by using the independence restriction in addition to the index restriction. An especially interesting feature of this estimation method is that even though the regressors \( x \) appearing in both (7.1) and (7.2) are the same, the parameter \( \beta \) is identifiable from the above estimation procedure. Lee [1990] provided a proof based on the identification of a least-squares procedure.

Let

\[
C_n(x_i) = \int_{-x_i \hat{\alpha}}^{\infty} \int_{-x_i \hat{\alpha}}^{\infty} \frac{1}{(n-1)a^2_n} \sum_{j \neq i} y_{2j} K \left[ \frac{u - u_j(\hat{\alpha})}{a_n}, \frac{z - x_j \hat{\alpha}}{a_n} \right] dudz,
\]
and

\[ D_n(z_i) = \int_{z_i \hat{\alpha}}^{\infty} \int_{z_i \hat{\alpha}}^{\infty} \frac{1}{(n-1)a_n^2} \sum_{j \neq i} x_{2i} K \left[ \frac{u - u_j(\hat{\alpha})}{a_n}, \frac{z - z_j(\hat{\alpha})}{a_n} \right] du dz. \]

This two-stage estimator of \( \beta \) has a closed form expression. The two-stage estimator \( \hat{\beta} \) is

\[
\hat{\beta} = \left\{ \sum_{i=1}^{n} I(x_i) \left[ x_{2i} - \frac{D_n(x_i)}{B_n(x_i)} \right] \left[ y_{2i} - \frac{D_n(x_i)}{B_n(x_i)} \right] \right\}^{-1} \left\{ \sum_{i=1}^{n} I(x_i) \left[ x_{2i} - \frac{D_n(x_i)}{B_n(x_i)} \right] \left( y_{2i} - \frac{C_n(x_i)}{B_n(x_i)} \right) \right\}.
\]

\[ (7.5) \]

\( D_n(x_i) \) provides a consistent estimate of \( E(z|x_0 > x_i\alpha) \). \( \hat{\beta} \) is a consistent estimator of \( \beta_0 \) under the identification condition that the components of \( x_i - E(x|x_0 > x_i\alpha) \) where \( x_i \in X \) are not linearly dependent. The curiosity about the identification of \( \beta \) is that there are no conditions to rule out the special case that \( E(v|u) \) is a linear function of \( u \) and \( E(u|u > -x_0) \) is a linear function of \( x_0 \) in (7.3).

When \( u \) is either a uniform variate or an exponential variate, \( E(u|u > -x_0) \) can be a linear function of \( x_0 \). Under such circumstances, \( \beta_0 \) can not be identified from the bias-adjusted regression equation (7.3). The estimation method in (7.4) seems to utilize a semiparametric bias-adjusted regression equation for estimation. The consistency of the estimation procedure (7.4) seems questionable from this point of view. However, the estimation procedure can be understood from a different angle. The model with (7.1) and (7.2) implies (7.3). It also implies the following equation at each point \( x_i \):

\[ E(y_2|u > -x_i\alpha_0, x_0 > x_i\alpha_0) = E(x|x_0 > x_i\alpha_0)\beta_0 + E(v|u > -x_i\alpha_0, x_0 > x_i\alpha_0). \]

(7.6)

Since \( u > -x_i\alpha_0 \) and \( x_0 > x_i\alpha_0 \) imply \( y_1 > 0 \), the conditional expectation functions \( E(x|x_0 > x_i\alpha_0) \) and \( E(y_2|u > -x_i\alpha_0, x_0 > x_i\alpha_0) \) can be estimated with observed sample.

Since \( E(v|u > -x_i\alpha_0, x_0 > x_i\alpha_0) = E(v|u > -x_i\alpha_0) \), under the independence property, the difference of (7.3) and (7.6) is

\[ E(y_2|y_1 > 0, x_i) - E(y_2|u > -x_i\alpha_0, x_0 > x_i\alpha_0) = [x_i - E(x|x_0 > x_i\alpha_0)]\beta_0. \]

(7.7)

Equation (7.6) provides a source for the identification of \( \beta_0 \) because, even if \( E(v|u > -x_i\alpha_0) \) were linear in \( x_i\alpha_0 \), \( E(x|x_0 > x_i\alpha_0) \) would in general not be linear in \( x_i\alpha_0 \). Equation (7.7) can be used for estimation by some SMD procedure. The estimator in (7.5) will be asymptotically equivalent to such an estimator.

The above example illustrates that there are semiparametric models with other conditional expectation functions different from the ones with index restrictions. The SMD method can be generalized to the estimation of such models. Indeed, it may be possible to generalize the SMD procedures to estimate models involving many other nonparametric functions, such as derivatives of nonparametric regression functions. The details may be best left for a case by case study, since the detailed asymptotic analysis might be different with different nonparametric functional estimates.

\[ \text{I am grateful for Professor Peter Schmidt for pointing out the exponential variate case.} \]

18
8. Conclusions

The SMD estimation methods introduced in this article are useful for the estimation of parametric and semiparametric models. For the estimation of parametric QR models, the WSMD estimation method provides an asymptotically efficient alternative to the maximum likelihood method even with disaggregated data. For some of the QR models, the SMD estimators have closed form expressions while the classical maximum likelihood method requires numerical iterative algorithms for computation. The SMD methods can be applied to the estimation of linear or nonlinear simultaneous equation models. There is no need to consider how to construct (optimal) instrumental variables for estimation. All that are required in the model are a list of proper conditioning variables and conditional expectation functions, which are structural. The SMD methods provide a unified framework for estimation. Once the conditional expectation functions or related functions are determined and the model is identifiable, the execution of the estimation procedure makes no distinction between the estimation of regression equations or simultaneous equations. With only conditional moment restrictions as in Chamberlain [1987], the WSMD estimator attains the Chamberlain's semiparametric efficiency bound.

The SMD methods can be applied to the estimation of many semiparametric models with index restrictions or certain moment restrictions. In the MD framework, the distinction between the estimation of parametric models and semiparametric models is also minimal. The important ingredients for semiparametric models are the identification of some implied structural equations which involve conditional expectations or related functions, which can be estimated by familiar nonparametric functions.

In this article, some local goodness-of-fit test statistics were also derived for the testing of model restrictions. However, many issues, such as asymptotically efficient estimation and testing of semiparametric models, have not been completely discussed here. Future research shall address some of such issues.
Appendix 1 : Some Useful Propositions

The following propositions are useful for the proof of our results. These propositions and their proofs can be found in Ichimura and Lee [1991] and Lee [1991]. The first four propositions are useful for establishing uniform convergence in probability of nonparametric regression functions with index restrictions and their first and second order derivatives (see Appendix 2). The last two propositions will be used in the proofs in Appendix 3. They are provided here for convenient reference.

Proposition A1.1 (A Uniform Law of Large Numbers) Let \( \{y_i\} \) be a sequence of i.i.d. random vectors. The measurable function \( h(y, a_n, \theta) \) takes the form, \( h(y, a_n, \theta) = \sum_{i=1}^{d} h_1(y_i, \theta) h_2(y, \theta, \frac{z(y, \theta)}{a_n}) \), where \( a_n = O(\frac{1}{n^d}) \), \( p > 0, d \geq 0, \theta \in \Theta \), and \( z(y, \theta) \) 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 \( h_1(y, \theta) \) is uniformly bounded by a \( \delta \)-order polynomial of \( y \):
   \[ \sup_{\theta \in \Theta} |h_1(y, \theta)| \leq c_j \|y\|^2, \]
   where \( c_j, j = 0, \ldots, \delta, \) are constants.
3. The first \( \delta \cdot r \) moments of \( y \) exist, where \( r \geq 2 \).
4. \( |h_2| \leq c \) for some constant \( c \).
5. \( E(h_1^2) = o(n^d) \) uniformly in \( \theta \in \Theta \).
6. The functions \( h_2(y, \theta, u) \), and \( s(y, \theta) \) satisfy the bounded Lipschitz condition of order 1 with respect to \( \theta \) and \( u \).

If \( \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(y_i, \theta, a_n) - E(h(y, \theta, a_n)) = 0 \), then \( \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} h(y_i, \theta, a_n) - h_\infty(\theta) = o(1) \) uniformly in \( \theta \in \Theta \).

Furthermore, if \( \Theta \) is a compact subset of a finite dimensional Euclidean space.

Proposition A1.2 Let \( K(v) \) be a function on \( \mathbb{R}^m \) 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 \( t(z, \theta) \), are uniformly continuous in \( t \), uniformly in \( (z_i, \theta) \). Then

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

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

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

Proposition A1.3 Let \( K(v) \) be a function on \( \mathbb{R}^m \) 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 \( (z_i, \theta) \). Then

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

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

\[ \sup_{z_i, \theta} \left| E \left[ c(z, z_i, \theta) \frac{1}{a_n+1} \frac{\partial K}{\partial v} \left( \frac{t(z_i, \theta) - \hat{t}(z_i, \theta)}{a_n} \right) \right] \right| = O(a_n^{s^*}). \]
Proposition A1.4 Let $K(v)$ be a twice differentiable function on $\mathbb{R}^m$ 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 are bounded. Suppose that $\frac{\partial^2}{\partial v \partial v'}[E(c(z, z_1, \theta)|t, z_1, \theta)g(t|\theta)]$ are uniformly continuous in $t$, uniformly in $(z_1, \theta)$. Then
\[
\lim_{n \to \infty} \sup_{z_1, \theta} \left\| E\left[ \left( z, z_1, \theta \right) - \left( z, z_1, \theta \right) \right] - \frac{1}{a_n + 2} \frac{\partial^2 K}{\partial v \partial v'} \left( \left( z, z_1, \theta \right) - \left( z, z_1, \theta \right) \right) \right\| = 0.
\]

Proposition A1.5 Let $C_{j,n}(z_1)$, $j = 1, 2$ be two sequences of measurable functions of an i.i.d. sample $\{z_i\}$. Suppose that, for each $j$,
\begin{enumerate}
\item $\sup_z |E(C_{j,n}(z_1)|z_1) - C_j(z_1)| = O\left(\frac{1}{n^{\frac{1}{2}}}\right)$, for some measurable function $C_j(z_1)$, and
\item $\sup_z \text{var}(C_{j,n}(z_1)|z_1) = O\left(\frac{1}{n^{\frac{1}{2}}}\right)$, $j = 1, 2$.
\end{enumerate}
If $\lim_{n \to \infty} a_{j,n}^2 = \infty$, $\lim_{n \to \infty} \frac{a_{j,n}^2}{a_{1,n}^2} = 0$, $\lim_{n \to \infty} \frac{a_{j,n}^2}{a_{1,n}^2} = 0$, and $\lim_{n \to \infty} n a_{1,n}^2 a_{2,n}^2 = 0$, then
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{z_i}(z_1)n C_{1,n}(z_1) - C_1(z_1)|C_{2,n}(z_1) - C_2(z_1)| \xrightarrow{p} 0.
\]

Proposition A1.6 Let $\{z_i\}$ be an i.i.d. sample and $\Phi_n(z_1, z_2, a_n)$ be a sequence of vector-valued random functions with bandwidth $\{a_n\}$. Suppose that
\begin{enumerate}
\item there exist square integrable functions $h_j(z)$, $j = 1, 2$ such that $|E(\Phi_n(z_1, z_2, a_n)|z_j)| \leq h_j(z_j)$, $j = 1, 2$,
\item $E(\Phi_n(z_1, z_2, a_n)) = O\left(\frac{1}{a_n^2}\right)$ and $\text{var}(\Phi_n(z_1, z_2, a_n)) = O\left(\frac{1}{a_n^2}\right)$,
\item $\lim_{n \to \infty} E(\Phi_n(z_1, z_2, a_n)|z_j) = \psi_j(z_j)$, a.e., for some measurable functions $\psi_j$, $j = 1, 2$, and
\item $\lim_{n \to \infty} \sqrt{n}a_n^2 = 0$ and $\lim_{n \to \infty} n a_n^2 = \infty$.
\end{enumerate}
If $\psi_1(z)$ and $\psi_2(z)$ are zero a.e., then $\frac{1}{\sqrt{n(n-1)}} \sum_{i=1}^{n} \sum_{j \neq i} \Phi_n(z_i, z_j, a_n) \xrightarrow{p} 0$.
On the other hand, if $\lim_{n \to \infty} \{[\psi_1(z) + \psi_2(z)][(\psi_1(z) + \psi_2(z))^\prime]\} = \Sigma$, then
\[
\frac{1}{\sqrt{n(n-1)}} \sum_{i=1}^{n} \sum_{j \neq i} \Phi_n(z_i, z_j, a_n) \xrightarrow{D} N(0, \Sigma).
\]
Appendix 2: 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 regression functions with index restrictions and related functions. Assumption 1 below contains the basic regularity conditions of our model. Assumptions 2 and 3 are conditions for the nonparametric regression function of \( f(z_1, \beta) \) on \( x \) and its first and second order derivatives. Assumptions 4 and 5 are conditions for the nonparametric regression function of \( g(z_2, \gamma) \) on \( z \delta \) and related functions. Assumption 6 is an additional assumption which is required only for the weighted estimation method.

Assumption 1:

1. \( \Theta \) is a compact convex subset of a finite dimensional Euclidean space, and the true parameter vector \( \theta_0 \) is in the interior of \( \Theta \).
2. The sample observations \((x_i, z_{1i}, z_{2i}), i = 1, \ldots, n\) are i.i.d.\(^3\)
3. \( \alpha(\theta) \) is a twice continuously differentiable vector-valued function of \( \theta \).
4. \( G[u, v, z, \alpha] \) is twice differentiable in \((u, v, \alpha)\) for each \( x \) and is a measurable function of \( x \) for each \((u, v, \alpha)\).
5. \( \sup_{x} \|G[E(f(z_1, \beta)|z), E(g(z_2, \gamma)|z_0), x, \alpha_0]G'[E(f(z_1, \beta)|z), E(g(z_2, \gamma)|z_0), x, \alpha_0]\) is nonsingular.

Assumption 2:

1. \( x \) is a \( k \)-dimensional vector of continuous random variables with a density \( h(x) \) on \( \mathbb{R}^k \).\(^4\)
2. \( x \) is chosen to be a compact subset contained in the interior of the support \( S_x \) of \( x \).
3. \( f \) is a vector-valued function which is twice differentiable w.r.t. \( \theta \), \( \sup_{x} \|f(z_1, \beta)\|, \sup_{x} \|\frac{\partial f(z_1, \beta)}{\partial \theta_j}\| \), and \( \sup_{x} \|\frac{\partial^2 f(z_1, \beta)}{\partial \theta_{j_1} \partial \theta_{j_2}}\| \), where \( j_1, j_2 = 1, \ldots, \dim \theta \), are bounded by a \( t_1 \)-order polynomial of \( z_1 \). The first \( t_1 \cdot r_1 \) moments with \( r_1 \geq 2 \) exist.
4. \( h(x), E(|f(z_1, \beta)|^2|x), E\left(\left\|\frac{\partial f(z_1, \beta)}{\partial \theta_{j_1}}\right\|^2|x\right), \) and \( E\left(\left\|\frac{\partial^2 f(z_1, \beta)}{\partial \theta_{j_1} \partial \theta_{j_2}}\right\|^2|x\right) \), where \( j_1, j_2 = 1, \ldots, \dim \theta \), are bounded on \( S_x \times \Theta \).
5. \( h(x), E(|f(z_1, \beta)|x), E\left(\left\|\frac{\partial f(z_1, \beta)}{\partial \theta_{j_1}}\right\|\right), \) and \( E\left(\left\|\frac{\partial^2 f(z_1, \beta)}{\partial \theta_{j_1} \partial \theta_{j_2}}\right\|\right), \) where \( j_1, j_2 = 1, \ldots, \dim \theta \), 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:

1. \( K(v) \) is a continuous kernel function on \( \mathbb{R}^k \) with a bounded support.
2. \( K(v) \) is a kernel with zero moments up to the order \( s_2^* \), i.e., \( \int v_1^{i_1} \cdots v_k^{i_k} K(v) dv = 0 \) for all \( 0 \leq i_l, l = 1, \ldots, k, \) and \( 1 \leq i_1 + \cdots + i_k < s_2^* \).
3. The bandwidth sequence \( \{a_n\} \) is chosen with a rate of convergence such that
   \[ \lim_{n \to \infty} \frac{n}{\ln n} a_n^{1+2i/r_1} = \infty, \quad \lim_{n \to \infty} n a_n^{2k} = \infty, \quad \text{but} \lim_{n \to \infty} n a_n^{2s_1^*} = 0. \]

Assumption 4:

1. \( \delta(\theta) \) is twice differentiable w.r.t. \( \theta \), and its first two order derivatives are bounded on \( \Theta \).
2. For each \( \theta \in \Theta, z \delta \) is an \( m \)-dimensional vector of continuous variables with a density \( p(z \delta|\theta) \).
3. \( p(z \delta|\theta) \) is bounded away from zero uniformly on \( X \times \Theta \).
4. \( g \) is a vector-valued function which is twice differentiable w.r.t. \( \theta \), \( \sup_{\theta} \|g(z_2, \gamma)\|, \sup_{\theta} \|\frac{\partial g(z_2, \gamma)}{\partial \theta_j}\|, \) and \( \sup_{\theta} \|\frac{\partial^2 g(z_2, \gamma)}{\partial \theta_{j_1} \partial \theta_{j_2}}\| \), where \( j_1, j_2 = 1, \ldots, \dim \theta \), are bounded by a \( t_2 \)-order polynomial of \( z_2 \). The first \( (t_2 + 2) \cdot r_2 \) moments with \( r_2 \geq 2 \) of \( (z_2, x) \) exist.

---

\(^3\) \( x, z_1, \) and \( z_2 \) may contain common variables.

\(^4\) This assumption can be generalized to allow some of the variables in \( x \) to be discrete. Nonparametric regression functions with discrete regressors can be found in Bierans [1987]. This assumption simplifies greatly the presentation of the proofs.
(4.5) The functions $p(x_6|\theta), E(||x||^4|x), E(||g(z_2, \gamma)||^2|x, \theta), E(||x||^4g(z_2, \gamma)||^2|x, \theta), E \left( \left| g_{\theta \theta_1} \right|^2 \right), E \left( \left| g_{\theta \theta_2} \right|^2 \right)$, and $E \left( \left| g_{\theta \theta_3} \right|^2 \right)$, where $j_1, j_2 = 1, \ldots, \dim \theta$, are bounded on $S_\delta \times \Theta$.

(4.6) $p(x_6|\theta), E(g(z_2, \gamma)|x, \theta), E \left( \left| g(z_2, \gamma) \right|^2 \right), \text{ and } E \left( \left| g(z_2, \gamma) \right|^2 \right)$, where $j_1, j_2 = 1, \ldots, \dim \theta$, are differentiable w.r.t. $x_6$ to the order $s_2$, and these $s_2$ order derivatives are uniformly bounded on $S_\delta \times \Theta$.

(4.7) For any $\theta_1, \theta_2, \theta_3$, and $j_1, j_2 = 1, \ldots, \dim \theta$, $E(a_{\theta_1} \mid x_6) \text{ and } E(a_{\theta_2} \mid x_6 \times \Theta)$.

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 $s_2^*$. 

(5.3) The bandwidth sequence $\{b_n\}$ is chosen with a rate of convergence such that

$$\lim_{n \to \infty} \frac{1}{\ln n} b_n^{(m+1)} = \infty, \quad \lim_{n \to \infty} n b_n^{m+2} = \infty, \quad \lim_{n \to \infty} n b_n^{m+2} = \infty$$

Assumption 6:

(6.1) $\sup ||f(z_1, \beta_1)f'(z_1, \beta_1)||$ is bounded by a $t_1$-order polynomial of $z_1$.

(6.2) $E(||f(z_1, \beta_1)f'(z_1, \beta_1)||^2|x)$ is bounded on $S_\delta \times \Theta$.

(6.3) $E(f(z_1, \beta_1)f'(z_1, \beta_1)|x)$ is uniformly continuous on $S_\delta \times \Theta$.

(6.4) $v(x)$ is strictly positive and bounded away from zero on $X$.

(6.5) The matrix $E(f(x, \beta_0)|x, x, \alpha_0)G_{\theta}[E(f(z_1, \beta_0)|x, x, \alpha_0)]$ is nonsingular.

A relatively stronger but simpler rate of convergence for $\{b_n\}$ can be used to replace the rate in Assumption 5.3. Given the rate of $\{a_n\}$ in Assumption 3.3, a rate for $\{b_n\}$ such that

$$\lim_{n \to \infty} \frac{1}{\ln n} b_n^{(m+4)+2(t_2+2)(m+2)/r_2} = \infty, \quad \lim_{n \to \infty} n b_n^{2(m+2)} = \infty, \quad \lim_{n \to \infty} n b_n^{2m+2} = 0,$$

implies Assumption 5.3. Under these stronger rates, $s_1^*$ and $s_2^*$ can be chosen such that

$$s_1^* > \max \left\{ \frac{1}{2} \left( k \left( 1 + \frac{2t_1}{r_1} \right) \right) \right\},$$

and

$$s_2^* > \max \left\{ m + 2, \frac{1}{2} (m + 4) + 2(t_2 + 2)(m + 2)/r_2 \right\}.$$

Suppose that $a_n = O \left( \frac{1}{n p_1} \right)$ and $b_n = O \left( \frac{1}{n p_2} \right)$, where $p_1$ and $p_2$ are chosen to satisfy the inequalities:

$$\frac{1}{2s_1^*} < p_1 < \min \left\{ \frac{1}{k(1 + 2t_1 \{ r_1 \}}, \frac{1}{2k} \right\},$$

and

$$\frac{1}{2s_2^*} < p_2 < \min \left\{ \frac{1}{(m + 4) + 2(t_2 + 2)(m + 2)/r_2}, \frac{1}{2(m + 2)} \right\}.$$
Any \( p_1 \) and \( p_2 \) satisfying these inequalities will satisfy the rate requirements. This is so, since \( n a_n^q \to 0 \) if and only if \( p_1 > 1/q \), and \( n a_n^q \to \infty \) or \( \frac{n}{a_n^q} \to \infty \) if and only if \( p_1 < 1/q \). Similar inequalities hold for \( b_n \).

Denote \( A_n(x_i, \theta) = \left( \frac{1}{n} \right)^m \sum_{j \neq i} g(z_{ij}, \gamma) J \left( \frac{z_{ij} - x_i}{b_n} \right) \), and \( B_n(x_i, \theta) = \left( \frac{1}{n} \right)^m \sum_{j \neq i} J \left( \frac{z_{ij} - x_i}{b_n} \right) \).

Proposition A1.1-A1.4 can be applied to these functions under the conditions in Assumptions 4 and 5. The variances of \( A_n(x_i, \theta) \) and \( B_n(x_i, \theta) \) have the familiar order \( O\left( \frac{1}{n} \right) \) uniformly on \( S_x \times \Theta \), i.e.,

\[
\sup_{S_x \times \Theta} \text{var}(A_n(x, \theta)|x_i) = O\left( \frac{1}{n} b_n^m \right) \quad \text{and} \quad \sup_{S_x \times \Theta} \text{var}(B_n(x, \theta)|x_i) = O\left( \frac{1}{n} b_n^m \right).
\]

Proposition A1.1 implies that if \( \lim_{n \to \infty} \frac{n}{a_n^q + 2} = \infty \), \( \sup_{S_x \times \Theta} |A_n(x, \theta) - E(A_n(x, \theta)|x_i)| \to 0 \), and \( \sup_{S_x \times \Theta} |B_n(x, \theta) - E(B_n(x, \theta)|x_i)| \to 0 \). Proposition A1.2 guarantees that

\[
\sup_{x \in X} |E\left( A_n(x, \theta)|x_i \right) - A(x, \theta)| = O\left( \frac{1}{n} a_n^q \right) \quad \text{and} \quad \sup_{x \in X} |E\left( B_n(x, \theta)|x_i \right) - B(x, \theta)| = O\left( \frac{1}{n} a_n^q \right),
\]

where \( A_n(x, \theta) \) and \( B_n(x, \theta) \) are in (3.2) and (3.3). Hence \( A_n(x, \theta) \) converges to \( A(x, \theta) \) and \( B_n(x, \theta) \) converges to \( B(x, \theta) \) in probability uniformly on \( X \times \Theta \). Since \( p(x) \) is bounded away from zero on \( X \times \Theta \), by uniform continuity \( B_n(x, \theta) \) is bounded away from zero on \( X \times \Theta \) in probability. Uniform convergence of \( B_n(x, \theta) \) and \( A_n(x, \theta) \) implies that \( E_n(g(z_2, \gamma)|x_i, \theta) \) converges in probability to \( E(g(z_2, \gamma)|x_i, \theta) \) uniformly on \( X \times \Theta \). Similarly, let

\[
A_n(x_i, \theta) = \frac{1}{(n-1)a_n^k} \sum_{j \neq i} f(z_{ij}, \beta) K \left( \frac{z_{ij} - x_i}{a_n} \right) \quad \text{and} \quad B_n(x_i, \theta) = \frac{1}{(n-1)a_n^k} \sum_{j \neq i} K \left( \frac{z_{ij} - x_i}{a_n} \right).
\]

It follows under Assumptions 2 and 3 that \( \sup_{S_x \times \Theta} |A_n(x, \theta)| = O\left( \frac{1}{n} a_n^q \right), \sup_{S_x \times \Theta} |B_n(x, \theta)| = O\left( \frac{1}{n^2} a_n^q \right) \), \( \sup_{S_x \times \Theta} |E(A_n(x, \theta)|x_i) - A(x, \theta)| = O\left( \frac{1}{n} a_n^q \right) \), \( \sup_{S_x \times \Theta} |E(B_n(x, \theta)|x_i) - B(x, \theta)| = O\left( \frac{1}{n} a_n^q \right) \), where \( A(x, \theta) = E(f(z_i, \beta)|x_i) h(x_i) \), and \( B(x) = h(x) \). As \( \lim_{n \to \infty} n^{-1} a_n^{k+2} \) is \( \infty \), \( E_n(f(z_i, \beta)|x_i) \) converges in probability to \( E(f(z_i, \beta)|x_i) \) uniformly on \( X \times \Theta \).

The first and second order derivatives of \( A_n(x_i, \theta) \) are \( \frac{\partial A_n(x_i, \theta)}{\partial \theta_{i1}} = \frac{1}{(n-1)a_n^k} \sum_{j \neq i} \frac{\partial f(z_{ij}, \theta)}{\partial \theta_{i1}} K \left( \frac{z_{ij} - x_i}{a_n} \right) \) and \( \frac{\partial^2 A_n(x_i, \theta)}{\partial \theta_{i1}^2} = \frac{1}{(n-1)a_n^k} \sum_{j \neq i} \frac{\partial^2 f(z_{ij}, \theta)}{\partial \theta_{i1}^2} K \left( \frac{z_{ij} - x_i}{a_n} \right) \). The variances of these derivatives have order \( O\left( \frac{1}{n} \right) \) uniformly on \( S_x \times \Theta \), and the biases are \( O\left( \frac{1}{n} a_n^q \right) \). As \( \lim_{n \to \infty} n^{-1} a_n^{k+2} \) is \( \infty \), \( \frac{\partial A_n(x_i, \theta)}{\partial \theta_{i1}} \) converges in probability to \( \frac{\partial A(x, \theta)}{\partial \theta_{i1}} \) uniformly on \( X \times \Theta \), where \( \frac{\partial A(x, \theta)}{\partial \theta_{i1}} = E \left( \frac{\partial f(z_i, \beta)}{\partial \theta_{i1}} \right) h(x) \); and \( \frac{\partial^2 A_n(x_i, \theta)}{\partial \theta_{i1} \partial \theta_{i2}} \) converges in probability to \( \frac{\partial^2 A(x, \theta)}{\partial \theta_{i1} \partial \theta_{i2}} \) uniformly on \( X \times \Theta \), where \( \frac{\partial^2 A(x, \theta)}{\partial \theta_{i1} \partial \theta_{i2}} = E \left( \frac{\partial^2 f(z_i, \beta)}{\partial \theta_{i1} \partial \theta_{i2}} \right) h(x) \). On the other hand,

\[
\frac{\partial A_n(x_i, \theta)}{\partial \theta_{i1}} = \frac{1}{(n-1)b_n^m} \sum_{j \neq i} \frac{\partial g(z_{ij}, \gamma)}{\partial \theta_{i1}} J \left( \frac{z_{ij} - x_i}{b_n} \right) + \frac{1}{(n-1)^m b_n^{m+1}} \sum_{j \neq i} g(z_{ij}, \gamma)(x_i - x_j) \frac{\partial J \left( \frac{z_{ij} - x_i}{b_n} \right)}{\partial \theta_{i1}},
\]

and

\[
\frac{\partial B_n(x_i, \theta)}{\partial \theta_{i1}} = \frac{1}{(n-1)b_n^m} \sum_{j \neq i} (x_i - x_j) \frac{\partial J \left( \frac{z_{ij} - x_i}{b_n} \right)}{\partial \theta_{i1}}.
\]
Their variances have order $O\left(\frac{1}{n^{b_2}}\right)$ uniformly on $S_x \times \Theta$, and their biases have $O(\epsilon_r^2)$ uniformly on $S_x \times \Theta$. Proposition A1.1 implies that as $\lim_{n \to \infty} \frac{n^{-2b_2}}{n^{2(m+2)+2(1+1)(m+1)/r_2}} = \infty$,

$$
\sup_{x \times \Theta} \left\| \frac{\partial B_{j,n}(x_i, \theta)}{\partial \theta_l} - \frac{\partial B_j(x_i, \theta)}{\partial \theta_l} \right\| = o_p(1), \quad \text{and} \quad \sup_{x \times \Theta} \left\| \frac{\partial A_{j,n}(x_i, \theta)}{\partial \theta_l} - \frac{\partial A_j(x_i, \theta)}{\partial \theta_l} \right\| = o_p(1),
$$

where the explicit expressions of $\frac{\partial B_j(x_i, \theta)}{\partial \theta_l}$ and $\frac{\partial A_j(x_i, \theta)}{\partial \theta_l}$ are in (3.4) and (3.5). The second order derivatives are relatively complicated. However, they can be analyzed similarly by Propositions A1.1-A1.4. It follows that as $\lim_{n \to \infty} \frac{n^{-4b_2}}{n^{2(m+4)+2(1+2)(m+2)/r_2}} = \infty$,

$$
\sup_{x \times \Theta} \left\| \frac{\partial^2 B_{j,n}(x_i, \theta)}{\partial \theta_l \partial \theta_l'} - \frac{\partial^2 B_j(x_i, \theta)}{\partial \theta_l \partial \theta_l'} \right\| = o_p(1), \quad \text{and} \quad \sup_{x \times \Theta} \left\| \frac{\partial^2 A_{j,n}(x_i, \theta)}{\partial \theta_l \partial \theta_l'} - \frac{\partial^2 A_j(x_i, \theta)}{\partial \theta_l \partial \theta_l'} \right\| = o_p(1).$$

More properties on nonparametric kernel regression functions with index restrictions and their derivatives can be found in Ichimura and Lee [1991].
Appendix 3: Proofs of Main Results

Proof of Proposition 3.1: Let \( Q_{I,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i)G^2[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta), x_i, \alpha] \).

By the uniform law of large numbers of Appendix 1, \( E_n(f(z_1, \beta)|x_i) \) and \( E_n(g(z_2, \gamma)|x_i \delta, \theta) \) converge in probability, respectively, to \( E(f(z_1, \beta)|x_i) \) and \( E(g(z_2, \gamma)|x_i \delta, \theta) \) uniformly in \((x_i, \theta) \in X \times \Theta\). Since \( G \) is a uniformly continuous function on its arguments, \( G^2[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta, x_i, \alpha)] \) converges in probability to \( G^2[E(f(z_1, \beta)|x_i), E(g(z_2, \gamma)|x_i \delta, \theta, x_i, \alpha)] \) uniformly in \((x_i, \theta) \in X \times \Theta\). Let

\[
Q_{I,n}^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i)G^2[E(f(z_1, \beta)|x_i), E(g(z_2, \gamma)|x_i \delta, \theta, x_i, \alpha)].
\]

It follows that

\[
\sup_{\theta \in \Theta} |Q_{I,n}(\theta) - Q_{I,n}^*(\theta)| \leq \sup_{x \in X} |G^2[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta, x_i, \alpha)] - G^2[E(f(z_1, \beta)|x_i), E(g(z_2, \gamma)|x_i \delta, \theta, x_i, \alpha)]| = o_p(1),
\]

i.e., \( Q_{I,n}(\theta) - Q_{I,n}^*(\theta) \) converges in probability to zero uniformly on \( \Theta \). The classical uniform law of large number (e.g., Amemiya [1985]) implies that \( Q_{I,n}^*(\theta) \) converges in probability to \( Q_I^*(\theta) \) uniformly on \( \Theta \), where

\[
Q_I^*(\theta) = E \left( I_X(x)G^2[E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x \delta, \theta, x, \alpha)] \right).
\]

Under our identification condition, \( Q_I^*(\theta) \) has a unique minimum at \( \theta = \theta_0 \). The consistency of \( \hat{\theta}_I \) follows from the uniform convergence of \( Q_{I,n}(\theta) \) to \( Q_I^*(\theta) \) and the unique minimizer of \( Q_I^*(\theta) \) being \( \theta_0 \).

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

\[
\sum_{i=1}^{n} I_X(x_i)G_\theta[E_n(f(z_1, \beta(\hat{\theta}_I))|x_i), E_n(g(z_2, \beta(\hat{\theta}_I))|x_i \delta(\hat{\theta}_I)), x_i, \alpha(\hat{\theta}_I)] = 0.
\]

By a Taylor series expansion,

\[
\sqrt{n}(\hat{\theta}_I - \theta_0) = -\left\{ \frac{1}{n} \sum_{i=1}^{n} I_X(x_i)G_\theta[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta), x_i, \alpha]G_\theta[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta), x_i, \alpha] \right\}^{-1}
\times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i)G_\theta[E_n(f(z_1, \beta_0)|x_i), E_n(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] \right\}.
\]

Explicitly,

\[
G_\theta[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x_i \delta), x_i, \alpha] = \left[ E_n \left( \frac{\partial f(z_1, \beta)}{\partial \theta} \Big| x_i \right), \frac{\partial E_n(g(z_2, \gamma)|\theta)|x_i \delta)}{\partial \theta} \right] \nabla E_n[f(z_1, \beta)|x_i], E_n(g(z_2, \gamma)|x_i \delta), x_i, \alpha],
\]

where \( \nabla \) denotes the gradient operator.
where $\nabla G[\cdot] \ denotes \ the \ gradient \ vector \ of \ G \ w.r.t. \ E_n(f(z_1, \beta)|x), E_n(g(z_2, \gamma)|x \delta), \ and \ \alpha$; and

$$
G_{\theta \theta', [E_n(f(z_1, \beta)|x), E_n(g(z_2, \gamma)|x \delta), x_i], \alpha] = \left( \begin{array}{c}
\left( \frac{\partial f(z_1, \beta)}{\partial \theta} \right) | x_i , \frac{\partial E_n(g(z_2, \gamma)|x \delta)}{\partial \theta}, \frac{\partial \alpha}{\partial \theta} \right)
\end{array} \right) \nabla^2 G[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x \delta), x_i, \alpha],
$$

where $\nabla^2 G[\cdot]$ denotes the hessian matrix of $G$ w.r.t. $E_n(f(z_1, \beta)|x), E_n(g(z_2, \gamma)|x \delta)$, and $\alpha$. Since $\tilde{\theta}$ converges in probability to $\theta_0$,

$$
G[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x \delta), x_i, \alpha] \rightarrow \frac{G[E(f(z_1, \beta_0)|x_i), E(g(z_2, \gamma_0)|x \delta_0), x_i, \alpha_0]}{0} = 0,
$$

and $G_{\theta \theta'} E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x \delta), x_i, \alpha] \rightarrow G_{\theta \theta'} [E(f(z_1, \beta_0)|x_i), E(g(z_2, \gamma_0)|x \delta_0), x_i, \alpha_0],$

uniformly in $x_i \in X$. It follows that

$$
\sqrt{n}(\tilde{\theta} - \theta_0) = - \left\{ E(I_X(x)G_{\theta}[E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x \delta), x, \alpha])
\times G_{\theta'}[E(f(z_1, \beta)|x), E(g(z_2, \gamma)|x \delta), x, \alpha)] + o_p(1) \right\}^{-1}
$$

$$
\times \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(x_i)G_{\theta}[E_n(f(z_1, \beta)|x_i), E_n(g(z_2, \gamma)|x \delta_0), x_i, \alpha_0]
\times G[E_n(f(z_1, \beta_0)|x_i), E_n(g(z_2, \gamma_0)|x \delta_0), x_i, \alpha_0] \right\} \right\}.
$$

Denote $A_n, B_n, A_{j,n}, B_{j,n}$ by

$$
A_n(x_i) = \frac{1}{(n-1)a_n} \sum_{j \neq i}^{n} f(x_j, \beta_0) K \left( \frac{x_i - x_j}{a_n} \right),
$$

$$
B_n(x_i) = \frac{1}{(n-1)a_n} \sum_{j \neq i}^{n} K \left( \frac{x_i - x_j}{a_n} \right),
$$

$$
A_{j,n}(x_i) = \frac{1}{(n-1)b_n} \sum_{j \neq i}^{n} g(x_j, \gamma_0) J \left( \frac{x_i \delta_0 - x_j \delta_0}{b_n} \right),
$$

and $B_{j,n}(x_i) = \frac{1}{(n-1)b_n} \sum_{j \neq i}^{n} J \left( \frac{x_i \delta_0 - x_j \delta_0}{b_n} \right)$. As $n$ goes to infinity, $A_n(x_i)$ converges in probability to $A(x_i)$ uniformly in $x_i \in X$, where $A(x_i) = E(f(z_1, \beta_0)|x_i) h(x_i)$; $B_n(x_i)$ converges in probability uniformly to $B(x_i)(h(x_i))$; $A_{j,n}(x_i)$ converges in probability uniformly to $A_{j,n}(x_i) = E(g(z_2, \gamma_0)|x \delta_0) p(x \delta_0)$; and $B_{j,n}(x_i)$ converges in probability uniformly to $B_{j,n}(x_i)$. Since $\frac{\partial E_n(f(z_1, \beta_0)|x_i)}{\partial \theta} = \frac{\partial A_{j,n}(x_i)}{\partial \theta}$, and

$$
\frac{\partial E_n(g(z_2, \gamma_0)|x \delta_0), x_i, \alpha_0]}{\partial \theta} = \frac{1}{B_{j,n}(x_i)} \frac{\partial A_{j,n}(x_i)}{\partial \theta} - \frac{A_{j,n}(x_i)}{B_{j,n}(x_i)} \frac{\partial B_{j,n}(x_i)}{\partial \theta},
$$

$G_{\theta}[E_n(f(z_1, \beta_0)|x_i), E_n(g(z_2, \gamma_0)|x \delta_0), x_i, \alpha_0]$ is a function of $A_n(x_i), B_n(x_i), A_{j,n}(x_i), B_{j,n}(x_i), \frac{\partial A_{j,n}(x_i)}{\partial \theta}$, $\frac{\partial A_{j,n}(x_i)}{\partial \theta}$, and $\frac{\partial B_{j,n}(x_i)}{\partial \theta}$. Without loss of generality, $f$ and $g$ will be treated as functions (instead of vector-valued functions) to simplify notation in the remaining arguments. By the mean value theorem,
where the remainder $R_{n,1}(x_i)$ is

$$R_{n,1}(x_i) = \nabla' G_\theta \left[ \bar{E}_n(f(z_1, \beta_\theta)|x_i), \bar{E}_n(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0 \right]$$

\[ \times \left[ \frac{1}{B_n(x_i)} (A_n(x_i) - A(x_i)) - \frac{\bar{A}_n(x_i)}{B_n^2(x_i)} (B_n(x_i) - B(x_i)), \right. \]

\[ \frac{1}{B_n(x_i)} \left( \frac{\partial A_n(x_i)}{\partial \theta'} - \frac{\partial A(x_i)}{\partial \theta'} \right) - \frac{\partial \bar{A}_n(x_i)}{\partial \theta'} \left( \frac{\partial B_n(x_i)}{\partial \theta'} - \frac{\partial B(x_i)}{\partial \theta'} \right), \]

\[ \frac{1}{B_n(x_i)} (A_{J_n}(x_i) - A_J(x_i)) - \frac{\bar{A}_{J_n}(x_i)}{B_{J_n}^2(x_i)} (B_{J_n}(x_i) - B_J(x_i)), \]

\[ - \frac{1}{B_{J_n}^2(x_i)} \frac{\partial \bar{B}_{J_n}(x_i)}{\partial \theta'} (A_{J_n}(x_i) - A_J(x_i)) \]

\[ - \left( \frac{1}{B_{J_n}^2(x_i)} \frac{\partial \bar{A}_{J_n}(x_i)}{\partial \theta'} - \frac{\partial \bar{A}_n(x_i)}{\partial \theta'} \right) \left( \frac{\partial \bar{B}_{J_n}(x_i)}{\partial \theta'} - \frac{\partial \bar{B}(x_i)}{\partial \theta'} \right) \]

\[ + \frac{1}{B_{J_n}^2(x_i)} \left( \frac{\partial A_{J_n}(x_i)}{\partial \theta'} - \frac{\partial A_J(x_i)}{\partial \theta'} \right) \left( \frac{\partial \bar{B}_{J_n}(x_i)}{\partial \theta'} - \frac{\partial \bar{B}(x_i)}{\partial \theta'} \right), \]

where $\nabla' G_\theta \left[ \cdot \right]$ denotes the derivative vector of $G_\theta \left[ \cdot \right]$ w.r.t. the arguments $E_n(f(z_1, \beta_\theta)|x_i), E_n(\frac{\partial f(z_1, \beta_\theta)}{\partial \theta}|x_i),$ $E_n(g(z_2, \gamma_\theta)|xi_0),$ and $\frac{\partial E_n(g(z_2, \gamma_\theta)|xi_0)}{\partial \theta}$, in that order, and $B_n(x_i)$ lies between $B_n(x_i)$ and $B(x_i)$, etc. On the other hand, since $G[E(f(z_1, \beta_\theta)|x_i), E(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0] = 0$, by a Taylor series expansion up to the second order,

$$G[E_n(f(z_1, \beta_\theta)|x_i), E_n(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0]$$

$$= \nabla'_{1,2} G \left[ E(f(z_1, \beta_\theta)|x_i), E(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0 \right]$$

\[ \times \left[ \left( A_n(x_i) - E(f(z_1, \beta_\theta)|x_i)B_n(x_i) \right) \frac{1}{h(x_i)} \left( A_{J_n}(x_i) - E(g(z_2, \gamma_\theta)|xi_0)B_{J_n}(x_i) \right) \frac{1}{p(x_i \delta_0)} \right], \]

\[ + R_{n,2}(x_i), \]

where

$$R_{n,2}(x_i) = [T_n(x_i) - T(x_i)] \nabla^2 G(\bar{T}_n(x_i))[T_n(x_i) - T(x_i)]', \]

$$T_n(x_i) = [A_n(x_i), B_n(x_i), A_{J_n}(x_i), B_{J_n}(x_i)], \quad T(x_i) = [A(x_i), B(x_i), A_J(x_i), B_J(x_i)], \quad \text{and} \quad \nabla^2 G(\cdot) \text{ is the hessian matrix of } G \text{ with respect to } T_n(x_i). \] The above equation is valid because

$$A_n(x_i) - E(f(z_1, \beta_\theta)|x_i)B_n(x_i) = [A_n(x_i) - A(x_i)] - E(f(z_1, \beta_\theta)|x_i)[B_n(x_i) - B(x_i)], \]

and $A_{J_n}(x_i) - E(f(z_1, \beta_\theta)|x_i)B_{J_n}(x_i) = [A_{J_n}(x_i) - A_J(x_i)] - E(f(z_1, \beta_\theta)|x_i)[B_{J_n}(x_i) - B_J(x_i)].$ It follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) G_\theta \left[ E_n(f(z_1, \beta_\theta)|x_i), E_n(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0 \right] \right. \]

\[ \times G[\bar{E}_n(f(z_1, \beta_\theta)|x_i), \bar{E}_n(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0] \}

\[ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) G_\theta \left[ E(f(z_1, \beta_\theta)|x_i), E(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0 \right] \right. \]

\[ \times \nabla'_{1,2} G[\bar{E}_n(f(z_1, \beta_\theta)|x_i), \bar{E}_n(g(z_2, \gamma_\theta)|xi_0), x_i, \alpha_0] \]

\[ \times \left. \left[ \left( A_n(x_i) - E(f(z_1, \beta_\theta)|x_i)B_n(x_i) \right) \frac{1}{h(x_i)} \left( A_{J_n}(x_i) - E(g(z_2, \gamma_\theta)|xi_0)B_{J_n}(x_i) \right) \frac{1}{p(x_i \delta_0)} \right] \right\} \}

\[ + R_n, \]

28
where \( R_n = L_{n,1} + L_{n,2} + L_{n,3} \) with

\[
L_{n,1} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) G_\theta[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] R_{n,2}(x_i),
\]

\[
L_{n,2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \nabla_1^2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0]
\times \left[ (A_n(x_i) - E(f(z_1, \beta_0)|z_1) B_n(x_i)) \frac{1}{h(x_i)} \right] \left( A_{J,n}(x_i) - E(g(z_2, \gamma_0)|x_i \delta_0) B_{J,n}(x_i) \right) \frac{1}{p(x_i \delta_0)} R_{n,1}(x_i),
\]

and \( L_{n,3} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) R_{n,1}(x_i) R_{n,2}(x_i) \). Since each term within the summations of \( L_{n,1}, L_{n,2}, \) and \( L_{n,3} \) 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 and 5, that \( L_{n,1} \) converges to zero in probability for all \( l \).

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

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I_X(x_i) G_\theta[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] \nabla_1^2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0]
\times (A_n(x_i) - E(f(z_1, \beta_0)|z_1) B_n(x_i)) \frac{1}{h(x_i)} \right] \right\} R_{n,1}(x_i),
\]

where \( w_i = (x_i, z_i) \),

\[
U_1(w_i, w_j, a_n) = I_X(x_i) G_\theta[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] \nabla_1^2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0]
\times (f(z_1, \beta_0) - E(f(z_1, \beta_0)|z_1)) \frac{1}{h(x_i)} a_n K \left( \frac{x_1 - x_j}{a_n} \right),
\]

and

\[
U_2(w_i, w_j, b_n) = I_X(x_i) G_\theta[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] \nabla_1^2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0]
\times (g(z_2, \gamma_0) - E(g(z_2, \gamma_0)|x_i \delta_0)) \frac{1}{p(x_i \delta_0)} \frac{1}{b_n^m} J \left( \frac{x_1 \delta_0 - x_j \delta_0}{b_n} \right).
\]

\[E[U_l(w_i, w_j, a_n)|w_i]\] converges to zero for \( l = 1, 2 \), but

\[
\lim_{n \to \infty} E[U_1(w_j, w_i, a_n)|w_i] = I_X(x_i) G_\theta[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0] \nabla_1^2 G[E(f(z_1, \beta_0)|z_1), E(g(z_2, \gamma_0)|x_i \delta_0), x_i, \alpha_0]
\times (f(z_1, \beta_0) - E(f(z_1, \beta_0)|z_1)),
\]

and

\[
\lim_{n \to \infty} E[U_2(w_j, w_i, b_n)|w_i] = E \left\{ I_X(x) G_\theta[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|x \delta_0), x, \alpha_0] \nabla_1^2 G[E(f(z_1, \beta_0)|z), E(g(z_2, \gamma_0)|x \delta_0), x, \alpha_0] \right| x_i \delta_0 \}
\times (g(z_2, \gamma_0) - E(g(z_2, \gamma_0)|x_i \delta_0)).
\]

29
The asymptotic distribution follows from the central limit theorem for U-statistics in Proposition A1.6 of Appendix 1 (see, also, Powell, Stock, and Stoker [1989]). Q.E.D.

Proof of Corollary 3.1: By the uniform law of large numbers of Appendix 1,

\[ Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i) C^2(E_n(f(z_1, \beta) | z_i), z_i, \alpha) \]

converges in probability to \( Q^*(\theta) = E(I_X(x) C^2[E(f(z_1, \beta) | x), x, \alpha]) \), uniformly in \( \theta \in \Theta \). Under our identification condition, \( Q^*(\theta) \) has a unique minimum at \( \theta = \theta_0 \). The consistency of \( \hat{\theta}_S \) follows from the uniform convergence of \( Q_n(\theta) \) to \( Q(\theta) \) and the unique minimum of \( Q^*(\theta) \) being \( \theta_0 \). The asymptotic distribution of \( \hat{\theta}_S \) follows from Proposition 3.1 with \( C^2(z) = 0 \). Q.E.D.

Proof of Proposition 3.2: Let

\[ Q_{F,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i)(y_i - F[E_n(f_1(z_1, \beta) | x_i), E_n(g(z_2, \gamma) | z_i, \alpha)]^2, \]

and \( Q_{F,n}^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i)(y_i - F[E_n(f_1(z_1, \beta) | x_i), E(g(z_2, \gamma) | z_i, \alpha)]^2. \) Since

\[ \sup_{\theta \in \Theta} |Q_{F,n}(\theta) - Q_{F,n}^*(\theta)| \]

\[ \leq 2 \sup_{x \times \Theta} |F[E_n(f_1(z_1, \beta) | x_i), E_n(g(z_2, \gamma) | x_i, \alpha)] - F[E(f_1(z_1, \beta) | x_i), E(g(z_2, \gamma) | x_i, \alpha)]| \]

\[ + \sup_{x \times \Theta} |F^2[E_n(f_1(z_1, \beta) | x_i), E_n(g(z_2, \gamma) | x_i, \alpha)] - F^2[E(f_1(z_1, \beta) | x_i), E(g(z_2, \gamma) | x_i, \alpha)]| \]

\[ = o_p(1), \]

\( Q_{F,n}(\theta) \) converges in probability to \( Q_{F,n}^*(\theta) \) uniformly in \( \theta \in \Theta \). It is apparent that \( Q_{F,n}^*(\theta) \) converges in probability to \( Q_F^*(\theta) \) uniformly in \( \theta \), where

\[ Q_F^*(\theta) = E(I_X(x)(y - F[E(f_1(z_1, \beta) | x), E(g(z_2, \gamma) | x, \alpha)]^2). \]

\( Q_F^*(\theta) \) is uniquely minimized at \( \theta_0 \). Therefore, \( \hat{\theta}_A \) is consistent.

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

\[ \sqrt{n}(\hat{\theta}_A - \theta_0) = \{ E(I_X(x) F_\theta [E(f_1(z_1, \beta) | x), E(g(z_2, \gamma) | z_0, \alpha)] \]

\[ \times F_\theta [E(f_1(z_1, \beta) | x), E(g(z_2, \gamma) | z_0, \alpha)] + o_p(1) \}^{-1} \]

\[ \times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ I_X(x_i) F_\theta [E_n(f_1(z_1, \beta) | x_i), E_n(g(z_2, \gamma) | z_i, \alpha)] \}

\[ \times (y_i - F[E_n(f_1(z_1, \beta) | x_i), E_n(g(z_2, \gamma) | z_i, \alpha)] \}

\[ \times \left( (\hat{A}_1, n(x_i) - E(f_1(z_1, \beta) | x_i) B_n(x_i) \frac{1}{h(x_i)}, (A_{j,n}(x_i) - E(g(z_2, \gamma) | z_i, \alpha) B_{j,n}(x_i)) \frac{1}{p(x_i, \alpha)} \right)^j \}

\[ = \frac{1}{\sqrt{n(n-1)}} \sum_{i=1}^{n} \sum_{j=1}^{n} U(w_i, w_j, b_n), \]
where \( w_i = (x_i, z_i) \),

\[
U(w_i, w_j, b_n) = I_X(x_i) F_\theta \left[ E(f_1(z_1, \beta_0) | x_i), E(g(z_2, \gamma_0) | x_i, \alpha_0) \right] \times \left\{ \epsilon_i - \nabla_1 F \left[ E(f_1(z_1, \beta_0) | x_i), E(g(z_2, \gamma_0) | x_i, \alpha_0) \right] \right\} \left[ \frac{1}{h(x_i)} \frac{1}{a_n} \int \frac{1}{b_n} K \left( \frac{x_i - x_j}{a_n} \right) \right] \\
	imes \left[ \frac{1}{p(x_i) b_n} \int \frac{1}{b_n} J \left( \frac{x_i \delta_0 - x_j \delta_0}{b_n} \right) \right].
\]

As \( \lim_{n \to \infty} E[U(w_i, w_j, b_n) | w_i] = I_X(x_i) F_\theta \left[ E(f_1(z_1, \beta_0) | x_i), E(g(z_2, \gamma_0) | x_i, \alpha_0) \right] \epsilon_i \) and

\[
\lim_{n \to \infty} E[U(w_j, w_k, b_n) | w_i] = -I_X(x_i) F_\theta \left[ E(f_1(z_1, \beta_0) | x_i), E(g(z_2, \gamma_0) | x_i, \alpha_0) \right] \\
\times \nabla_1 F \left[ E(f_1(z_1, \beta_0) | x_i), E(g(z_2, \gamma_0) | x_i, \alpha_0) \right] \left[ f_1(z_1, \beta_0) - E(f_1(z_1, \beta_0) | x_i) \right] \\
\times \nabla_2 F \left[ E(f_1(z_1, \beta_0) | x), E(g(z_2, \gamma_0) | x, \alpha_0) \right] \left[ g(z_2, \gamma_0) - E(g(z_2, \gamma_0) | x, \alpha_0) \right]
\]

the asymptotic distribution of \( \tilde{\theta}_A \) follows from the U-statistic central limit theorem. Q.E.D.

Proof of Proposition 4.1: Let \( Q_{n, w}(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i) \frac{1}{v_n(x_i, \theta)} G^2[E_n(f(z_1, \beta) | x_i), x_i, \alpha] \). By the uniform law of large number with bandwidth (Proposition A1.1 of Appendix 1) under Assumption 6, \( \hat{\sigma}_n^2(x) \) converge in probability to \( \sigma^2(x) \) uniformly on \( X \). It follows that \( v_n(x_i, \hat{\theta}) \) converges in probability to \( v(x_i) \) uniformly on \( X \). Since \( X \) is compact and \( v(x) \) is continuous, \( v(x) \) is bounded away from zero on \( X \). Consequently, \( v_n(x_i, \hat{\theta}) \) is bounded away from zero in probability on \( X \), and \( \sup_{x \in X} \frac{1}{v_n(x_i, \hat{\theta})} = O_p(1) \).

These conditions imply

\[
\sup_{x \in X} \left| \frac{1}{v_n(x_i, \hat{\theta})} - \frac{1}{v(x_i)} \right| = \sup_{x \in X} \left| \frac{v(x_i) - v_n(x_i, \hat{\theta})}{v(x_i) v_n(x_i, \hat{\theta})} \right| \leq O_p(1) \cdot \sup_{x \in X} |v(x_i) - v_n(x_i, \hat{\theta})| = o_p(1).
\]

Let \( Q_{n, w}^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_X(x_i) \frac{1}{v(x_i)} G^2[E(f(z_1, \beta) | x_i), x_i, \alpha] \). Then,

\[
\sup_{\theta \in \Theta} |Q_{n, w}(\theta) - Q_{n, w}^*(\theta)| \leq \sup_{x \in X} \left| \frac{1}{v_n(x_i, \hat{\theta})} G^2[E_n(f(z_1, \beta) | x_i), x_i, \alpha] - \frac{1}{v(x_i)} G^2[E(f(z_1, \beta) | x_i), x_i, \alpha] \right| \]

\[
\leq \sup_{\theta \in \Theta} \left| \frac{1}{v_n(x_i, \hat{\theta})} \right| G^2[E_n(f(z_1, \beta) | x_i), x_i, \alpha] - G^2[E(f(z_1, \beta) | x_i), x_i, \alpha] \times \sup_{\theta \in \Theta} \left| \frac{1}{v_n(x_i, \hat{\theta}) v(x_i)} \right| G^2[E(f(z_1, \beta) | x_i), x_i, \alpha] \cdot |v_n(x_i, \hat{\theta}) - v(x_i)| = o_p(1).
\]

On the other hand, \( Q_{n, w}(\theta) \) converges in probability to \( Q_{w}^*(\theta) \) uniformly on \( \Theta \), where

\[
Q_{w}^*(\theta) = E \left( I_X(x) \frac{1}{v(x)} G^2[E(f(z_1, \beta) | x), x, \alpha] \right).
\]
The consistency follows because $\theta_o$ is the unique minimizer of $Q^*_w(\theta)$. The WSMD estimator satisfies the first order condition:

$$\sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta)} G_\theta \left[ E_n(f(z_1, \beta(\vartheta))|z_i), z_i, \alpha(\vartheta) \right] G \left[ E_n(f(z_1, \beta(\vartheta))|z_i), z_i, \alpha(\vartheta) \right] = 0.$$ 

It follows by a Taylor series expansion that

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta)} G_\theta \left[ E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o \right] G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o]$$

$$+ \left\{ \frac{1}{n} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta)} G_\vartheta \left[ E_n(f(z_1, \beta)|z_i), z_i, \alpha \right] G \left[ E_n(f(z_1, \beta)|z_i), z_i, \alpha \right] \right\} \cdot \sqrt{n}(\vartheta - \vartheta)_{0},$$

where $\vartheta$ lies between $\vartheta_0$ and $\vartheta_o$. The uniform law of large numbers with a bandwidth implies that

$$\sqrt{n}(\vartheta - \vartheta_0) = - \left\{ E \left( I_X(z) \frac{1}{v(x)} G_\theta [E(f(z_1, \beta_o)|x), x, \alpha_o] \cdot G [E(f(z_1, \beta_o)|x), x, \alpha_o] + o_p(1) \right) \right\}^{-1}$$

$$\times \frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta)} G [G_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] \cdot [G_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] + o_p(1).$$

By a Taylor expansion, $\frac{1}{v_n(z_i, \vartheta)} = \frac{1}{v_n(z_i, \vartheta_0)} - \frac{1}{v_n^2(z_i, \vartheta_0)} \frac{\partial v_n(z_i, \vartheta)}{\partial \vartheta} (\vartheta - \vartheta_0)$. Therefore,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta_0)} G_\theta \left[ E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o \right] G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o]$$

$$- \frac{1}{\sqrt{n}} \sum_{i=1}^n G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] \cdot G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] \frac{1}{v_n^2(z_i, \vartheta_0)} \frac{\partial v_n(z_i, \vartheta)}{\partial \vartheta} \cdot \sqrt{n}(\vartheta - \vartheta_0)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta_0)} G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] \cdot G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o] + o_p(1).$$

Denote

$$A_n^{(2)}(z_i) = \frac{1}{(n - 1) \alpha_n^2} \sum_{j \neq i} f(z_j, \beta_o) f'(z_j, \beta_o) K \left( \frac{z_i - z_j}{\alpha_n} \right),$$

$$D_n(z_i) = (A_n(z_i), B_n(z_i), \text{vec}^T[A_n^{(2)}(z_i)]),$$

and $D(z_i) = (A'(z_i), B(z_i), \text{vec}^T[A^{(2)}(z_i)])$. By a Taylor expansion of $\frac{1}{v_n(z_i, \vartheta_0)}$ w.r.t. $D_n(z_i)$ at $D(z_i)$,

$$\frac{1}{v_n(z_i, \vartheta_0)} = \frac{1}{v(z_i)} + R_n, \text{ where } R_n = \frac{\nabla v_n(z_i, \vartheta_0)(D_n(z_i) - D(z_i))}{v_n(z_i, \vartheta_0)},$$

and $\nabla v_n(z_i, \vartheta_0)$ denotes the gradient of $v_n(z_i, \vartheta_0)$ w.r.t. $D_n(z_i)$, evaluated at some point between $D_n(z_i)$ and $D(z_i)$.

Combining the terms,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v_n(z_i, \vartheta_0)} G_\theta \left[ E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o \right] G [E_n(f(z_1, \beta_o)|z_i), z_i, \alpha_o]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n I_X(z_i) \frac{1}{v(z_i, \vartheta_0)} G [E(f(z_1, \beta_o)|z_i), z_i, \alpha_o] \cdot \nabla [G[E(f(z_1, \beta_o)|z_i), z_i, \alpha_o]] \cdot \frac{1}{h(z_i)} + R_n^{(\tau)},$$

32
where $R_n^{(7)} = \sum_{l=1}^{7} S_{n,l}$ with

$$S_{n,1} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \frac{1}{v(x_i, \theta_0)} \gamma [E(f(z_1, \beta_0)|x_i), x_i, \alpha_0] R_{n,2}(x_i),$$

$$S_{n,2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \frac{1}{v(x_i)} \nabla'_1 G [E_n(f(z_1, \beta_0)|x_i), x_i, \alpha_0] [A_n(x_i) - E(f(z_1, \beta_0)|x_i)] B_n(x_i) \frac{1}{h(x_i)} R_{n,1}(x_i),$$

$$S_{n,3} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \frac{1}{v(x_i)} R_{n,1}(x_i) R_{n,2}(x_i),$$

$$S_{n,4} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \gamma [E(f(z_1, \beta_0)|x_i), x_i, \alpha_0] \cdot \nabla'_1 G [E(f(z_1, \beta_0)|x_i), x_i, \alpha_0]$$

$$\times [A_n(x_i) - E(f(z_1, \beta_0)|x_i)] B_n(x_i) \frac{1}{h(x_i)} R_{n,3}(x_i),$$

$$S_{n,5} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \gamma [E(f(z_1, \beta_0)|x_i), x_i, \alpha_0] R_{n,3}(x_i) R_{n,3}(x_i),$$

$$S_{n,6} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) \nabla'_1 G [E_n(f(z_1, \beta_0)|x_i), x_i, \alpha_0] [A_n(x_i) - E(f(z_1, \beta_0)|x_i)] B_n(x_i) \frac{1}{h(x_i)} R_{n,1}(x_i) R_{n,3}(x_i),$$

and $S_{n,7} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_X(x_i) R_{n,1}(x_i) R_{n,2}(x_i) R_{n,3}(x_i)$. Since all $S_{n,l}$, $l = 1, \ldots, 7$, involve errors of higher order terms, Proposition A1.5 implies that all of them converge in probability to zero. Using arguments similar to those in the proof of Proposition 3.2, the result follows from Proposition A1.6. Q.E.D.
Appendix 4: SMD Estimation with Finite Restrictions

Suppose that \( z_1, 1 \leq l \leq L \), where \( L \) does not depend on \( n \), are predetermined regressor vectors in the interior of the support of \( z \). With \( L \) equations, an SMD estimation procedure is

\[
\min_{\theta \in \Theta} \sum_{l=1}^{L} G^2[E_n(f(z_1, \beta(\theta)))|z_1], E_n(g(z_2, \gamma(\theta))|z_1 \delta(\theta), z_1, \alpha(\theta)]. \tag{A4.1}
\]

As the points of \( z_1 \) are in the interior of the support of \( z \), trimming of the regressors as in (2.12) is done implicitly. Let \( \hat{\theta}_L \) denote the SMD estimator from (A4.1). The identification of \( \theta_0 \) requires the condition that for any \( \theta \neq \theta_0 \),

\[
G[E(f(z_1, \beta(\theta)))|z_1], E(g(z_2, \gamma(\theta))|z_1 \delta(\theta), x_1, \alpha(\theta)] \neq \mathbf{0},
\]

for some \( l \in \{1, \cdots, L\} \). The consistency of \( \hat{\theta}_L \) follows from the uniform convergence of \( S_{L,n}(\theta) \) to \( S_L(\theta) \), where

\[
S_{L,n}(\theta) = \sum_{l=1}^{L} G^2[E_n(f(z_1, \beta)|z_1)], E_n(g(z_2, \gamma)|z_1 \delta, x_1, \alpha],
\]

and \( S_L(\theta) = \sum_{l=1}^{L} G^2[E(f(z_1, \beta)|z_1)], E(g(z_2, \gamma)|z_1 \delta, x_1, \alpha] \), uniformly one8 as \( n \) goes to infinity and from the identification condition.

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

\[
\sum_{l=1}^{L} G_{\theta}[E_n(f(z_1, \beta(\hat{\theta}_L))|z_1)], E_n(g(z_2, \gamma(\hat{\theta}_L))|z_1 \delta(\hat{\theta}_L)), x_1, \alpha(\hat{\theta}_L)]]
\]

\[
\times G[E_n(f(z_1, \beta(\hat{\theta}_L))|z_1)], E_n(g(z_2, \gamma(\hat{\theta}_L))|z_1 \delta(\hat{\theta}_L)), x_1, \alpha(\hat{\theta}_L)] = 0.
\]

By a mean value theorem applied twice to \( G \),

\[
\sqrt{n a_n^k(\hat{\theta}_L - \theta_0)}
\]

\[
= -\left\{ \sum_{l=1}^{L} G_{\theta}[E_n(f(z_1, \beta)|z_1)], E_n(g(z_2, \gamma)|z_1 \delta, x_1, \alpha]G_{\theta}[E_n(f(z_1, \beta)|z_1)], E_n(g(z_2, \gamma)|z_1 \delta, x_1, \alpha] \right\}^{-1}
\]

\[
\times \sum_{l=1}^{L} \{ G_{\theta}[E_n(f(z_1, \beta)|z_1)], E_n(g(z_2, \gamma)|z_1 \delta, x_1, \alpha] \nabla_{1,2} G[E_n(f(z_1, \beta_0)|z_1)], E_n(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0]
\]

\[
\times \left( \sqrt{n a_n^k[E_n(f(z_1, \beta_0)|z_1)] - E(f(z_1, \beta_0)|z_1]) \right) \nabla_{1,2} G[E_n(f(z_1, \beta_0)|z_1)], E_n(g(z_2, \gamma_0)|z_1 \delta_0)] \}
\]

where \( \delta \) lies between \( \hat{\theta}_L \) and \( \theta_0 \), \( \hat{E}_n(\cdots) \) lies between \( E_n(\cdots) \) and \( E(\cdots) \), and \( \nabla_{1,2} G(\cdots) \) denotes the gradient vector of \( G \) w.r.t. to its first two vectors of arguments. It is well known from the nonparametric regression estimation literature (see, e.g., Bierens [1987]) that, with a proper rate of convergence for \( a_n \) and a higher order kernel to correct the bias,

\[
\sqrt{n b_n^k(E_n(f(z_1, \beta_0)|z_1)] - E(f(z_1, \beta_0)|z_1)]) \overset{D}{=} N \left( 0, \frac{1}{h(z_1)]} [E(f^2(z_1, \beta_0)|z_1)] - E^2(f(z_1, \beta_0)|z_1)] \right) \int K^2(u)du
\]

and

\[
\sqrt{n b_n^m[E_n(g(z_2, \gamma_0)|z_1 \delta_0)] - E(g(z_2, \gamma_0)|z_1 \delta_0)) \]

\[
\overset{D}{=} N \left( 0, \frac{1}{h(z_1)]} [E(g^2(z_2, \gamma_0)|z_1 \delta_0)] - E^2(g(z_2, \gamma_0)|z_1 \delta_0)] \right) \int K^2(u)du
\]

Since \( G \) and \( G_{\theta} \) are continuous functions,

\[
\nabla_{1,2} G[E_n(f(z_1, \beta_0)|z_1)], E_n(g(z_2, \gamma_0)|z_1 \delta, x_1, \alpha] \overset{D}{=} \nabla_{1,2} G(E(f(z_1, \beta_0)|z_1)], E(g(z_2, \gamma_0)|z_1 \delta_0), x_1, \alpha_0],
\]

34
and \( G_\theta \left[ E_n \left( f(z_1, \hat{\beta}) | x_1 \right), E_n \left( g(z_2, \hat{\gamma}) | x_1 \hat{\delta} \right), x_1, \hat{\alpha} \right] \) go to zero as \( n \) goes to infinity. Consequently, \( \sqrt{n} a_n^k \cdot (\hat{\theta}_L - \theta_0) \overset{D}{\rightarrow} N(0, \Omega) \), where

\[
\Omega = \left\{ \sum_{l=1}^{L} G_\theta \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] G_\psi \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] \right\}^{-1}
\]

and

\[
\Sigma = \int K^2(u) du \cdot \sum_{l=1}^{L} \frac{1}{h(x_l)} G_\theta \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right]
\]

\[
\times \nabla^2 G \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] \left[ E(f(z_1, \beta_0) | x_1) - E \left( f(z_1, \beta_0) \big| x_1 \right) \right] \left[ E \left( g(z_2, \gamma_0) \big| x_1 \delta_0 \right) - E \left( g(z_2, \gamma_0) \big| x_1 \delta_0 \right) \right]
\]

The nonparametric regression error \( E_n \left( g(z_2, \gamma_0) | x_1 \delta_0 \right) - E \left( g(z_2, \gamma_0) | x_1 \delta_0 \right) \) does not influence the asymptotic distribution of \( \hat{\theta}_L \). This is so, because \( x_\delta_0 \) has a smaller dimension than \( x \) which usually implies that the rate of convergence of \( E_n \left( g(z_2, \gamma_0) | x_1 \delta_0 \right) \) to \( E \left( g(z_2, \gamma_0) | x_1 \delta_0 \right) \) is faster than the rate of convergence of \( E_n \left( f(z_1, \beta_0) | x_1 \right) \) to \( E \left( f(z_1, \beta_0) | x_1 \right) \). The asymptotic distribution of \( \hat{\theta}_L \) is dominated by the stochastic element with the slowest rate of convergence. The rate of convergence of \( O \left( \frac{1}{\sqrt{n} a_n^k} \right) \) for \( \hat{\theta}_L \) reflects the rate of convergence of the nonparametric regression estimate \( E_n \left( f(z_1, \beta) | x_1 \right) \).

The SMD estimation procedure in (A4.1) is an unweighted procedure. It is possible to derive a weighted estimation procedure to improve its asymptotic efficiency. The weighted procedure is also useful for the construction of a goodness-of-fit statistic. At each \( x_1 \), define

\[
v_n(x_1) = \nabla^2 G \left[ E_n \left( f(z_1, \hat{\beta}) | x_1 \right), E_n \left( g(z_2, \hat{\gamma}) | x_1 \hat{\delta} \right), x_1, \hat{\alpha} \right] \frac{1}{h_n(x_1)} \left[ E_n \left( f^2(z_1, \hat{\beta}) | x_1 \right) - E_n \left( f^2(z_1, \hat{\beta}) | x_1 \right) \right]
\]

\[
\times \int K^2(u) du \cdot \nabla^2 G \left[ E_n \left( f(z_1, \hat{\beta}) | x_1 \right), E_n \left( g(z_2, \hat{\gamma}) | x_1 \hat{\delta} \right), x_1, \hat{\alpha} \right],
\]

where \( h_n(x_1) \) is a kernel estimate of the density of \( x \) at \( x_1 \), and the estimates \( \hat{\alpha}, \hat{\beta}, \) and \( \hat{\delta} \) are evaluated at \( \hat{\theta}_L \). The inverse of \( v_n(x_1) \) can be used as a weighting function in the following estimation procedure:

\[
\min_{\theta \in \Theta} \frac{1}{\sum_{l=1}^{L} v_n(x_1)} G_\theta \left[ E_n \left( f(z_1, \beta(\theta)) | x_1 \right), E_n \left( g(z_2, \gamma(\theta)) | x_1 \delta(\theta) \right), x_1, \alpha(\theta) \right].
\]

Let \( \hat{\theta}_{L,w} \) denote the weighted SMD estimator from (A4.2). The estimator \( \hat{\theta}_{L,w} \) has the same rate of convergence as the unweighted SMD estimator \( \hat{\theta}_L \) but is relatively more efficient: \( \sqrt{n} a_n^k \cdot (\hat{\theta}_{L,w} - \theta_0) \overset{D}{\rightarrow} N(0, \Omega) \), where

\[
\Omega = \left\{ \sum_{l=1}^{L} \frac{1}{v(x_1)} G_\theta \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] G_\psi \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] \right\}^{-1},
\]

and

\[
v(x_1) = \nabla^2 G \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right] \frac{1}{h(x_1)} \left[ E \left( f^2(z_1, \beta_0) | x_1 \right) - E^2 \left( f(z_1, \beta_0) | x_1 \right) \right]
\]

\[
\times \int K^2(u) du \cdot \nabla^2 G \left[ E(f(z_1, \beta_0) | x_1), E(g(z_2, \gamma_0) | x_1 \delta_0), x_1, \alpha_0 \right].
\]
The optimally weighted MD estimation procedure in Berkson [1944], Taylor [1953], etc., is also known as a minimum-chi-square procedure because its minimized weighted objective function multiplied by its corresponding sample size is asymptotically distributed chi-square with a degree of freedom equal to the number of equations minus the number of restricted parameters. With a finite number of restrictions, the semiparametric weighted MD estimation procedure (A4.2) is also a minimum chi-square procedure because

\[
\frac{1}{n} \sum_{i=1}^{L} \frac{1}{v_n(x_i)} G^2[E_n(f(z_1, \beta(\hat{\theta}_{L,w}))[x_1], E_n(g(z_2, \gamma(\hat{\theta}_{L,w}))[x_1 \delta(\hat{\theta}_{L,w})], z_1, \alpha(\hat{\theta}_{L,w})] \overset{D}{\longrightarrow} \chi^2(L - \dim \theta).}
\]

This result can be shown as follows. By a Taylor expansion of \(G\) at \(\theta\),

\[
G[E_n(f(z_1, \alpha(\hat{\theta}_{L,w}))[x_1], E_n(g(z_2, \gamma(\hat{\theta}_{L,w}))[x_1 \delta(\hat{\theta}_{L,w})], z_1, \alpha)] = \nabla_1 G[\hat{E}_n(f(z_1, \beta_0))[x_1], \hat{E}_n(g(z_2, \gamma_0))[x_1 \delta_0], z_1, \alpha_0] + G_\beta [E_n(f(z_1, \beta))[x_1], E_n(g(z_2, \gamma))[x_1 \delta], z_1, \alpha](\hat{\theta}_{L,w} - \theta_0)
\]

\[
\times [E_n(f(z_1, \beta_0))[x_1] - E(f(z_1, \beta_0))[x_1], E_n(g(z_2, \gamma_0))[x_1 \delta_0] - E(g(z_2, \gamma_0))[x_1 \delta_0)]' + G_\beta [E_n(f(z_1, \beta))[x_1], E_n(g(z_2, \gamma))[x_1 \delta], z_1, \alpha](\hat{\theta}_{L,w} - \theta_0).
\]

As the nonparametric regression functions converge in probability to their limiting functions, it follows that

\[
\sqrt{\frac{n}{a_n^k}} G[E_n(f(z_1, \beta[\hat{\theta}_{L,w}])[x_1], E_n(g(z_2, \gamma[\hat{\theta}_{L,w}])[x_1 \delta(\hat{\theta}_{L,w})], z_1, \alpha[\hat{\theta}_{L,w}]) = \nabla_1 G[E(f(z_1, \beta))[x_1], E(g(z_2, \gamma))[x_1 \delta], z_1, \alpha] \sqrt{\frac{n}{a_n^k}} (E_n(f(z_1, \beta_0))[x_1] - E(f(z_1, \beta_0))[x_1]) + G_\beta [E(f(z_1, \beta))[x_1], E(g(z_2, \gamma))[x_1 \delta], z_1, \alpha] \sqrt{\frac{n}{a_n^k}}(\hat{\theta}_{L,w} - \theta_0) + o_p(1).
\]

Let \(G_L\) be a vector of dimension \(L\) with its \(l\)th element being

\[
G[E_n(f(z_1, \beta[\hat{\theta}_{L,w}])[x_1], E_n(g(z_2, \gamma[\hat{\theta}_{L,w}])[x_1 \delta(\hat{\theta}_{L,w})], z_1, \alpha[\hat{\theta}_{L,w}])].
\]

Let \(G_L, G_\theta\) be the \(L \times \dim \theta\) matrix of the derivatives of \(G\) w.r.t. \(\theta\). Let \(H_n\) be a vector of dimension \(L\) with its \(l\)th element being \(\nabla_1 G[E(f(z_1, \beta))[x_1], E(g(z_2, \gamma))[x_1 \delta], z_1, \alpha] \sqrt{\frac{n}{a_n^k}}(E_n(f(z_1, \beta_0))[x_1] - E(f(z_1, \beta_0))[x_1])\). Finally, let \(V\) be an \(L \times L\) diagonal matrix with elements \(v(l, l), l = 1, \ldots, L\). In terms of these matrices and vectors, \(\sqrt{\frac{n}{a_n^k}} G_L = H_n + G_L \theta \cdot \sqrt{\frac{n}{a_n^k}}(\hat{\theta}_{L,w} - \theta_0) + o_p(1)\). Since \(\sqrt{\frac{n}{a_n^k}}(\hat{\theta}_{L,w} - \theta_0) = -G_L V^{-1} (G_L, V_\theta)^{-1} G_L, V^{-1} H_n + o_p(1)\), it follows that \(\sqrt{\frac{n}{a_n^k}} G_L = V^{1/2} M V^{-1/2} H_n\), where \(M = I - V^{-1/2} G_L, V^{-1/2} G_L, V^{-1} H_n\), and \(V^{-1/2} H_n\) is asymptotically normal \(N(0, I)\).
References

Recent Crest Working Papers


91-18: Ted Bergstrom, "When Non-Transitive Relations Take Maxima and Competitive Equilibria Can't be Beat."


