Simple Estimators for Invertible Index Models

Hyungtaik Ahn
Dongguk University

Hidehiko Ichimura
University of Tokyo

James L. Powell
University of California, Berkeley

Paul A. Ruud
Vassar College

March 2015

Preliminary and Incomplete Draft

Abstract

This paper considers estimation of the unknown linear index coefficients of a model in which a number of nonparametrically-identified reduced form parameters are assumed to be smooth and invertible function of one or more linear indices. The results extend the previous literature by allowing the number of reduced form parameters to exceed the number of indices (i.e., the indices are "overdetermined" by the reduced form parameters. The estimation method is an extension of an approach proposed by Ahn, Ichimura, and Powell (2004) for monotone single-index regression models to a multi-index setting and extended by Blundell and Powell (2004) and Powell and Ruud (2008) to models with endogenous regressors and multinomial response, respectively. The estimator of the unknown index coefficients (up to scale) is the eigenvector of a matrix (defined in terms of a first-step nonparametric estimator of the reduced form parameters) corresponding to its smallest (in magnitude) eigenvalue. Under suitable conditions, the proposed estimator is root-n-consistent and asymptotically normal, and results of a small-scale simulation study illustrate its finite-sample performance.

JEL Classification: C24, C14, C13.

1. Introduction

The object of this paper is to construct a class of computationally-simple estimators for models in which a vector of nonparametric "reduced form" parameters $\boldsymbol{\gamma}(\mathbf{X})$ is assumed to depend upon the matrix $\mathbf{X}$ of regressors only through a vector $\mathbf{X}\boldsymbol{\theta}_0$ of indices–that is, $\boldsymbol{\gamma}(\mathbf{X}) = \mathbf{g}(\mathbf{X}\boldsymbol{\theta}_0)$–where $\mathbf{g}$ is an unknown function and the coefficient vector $\boldsymbol{\theta}_0$ is the unknown parameter of interest. In addition to the familiar smoothness restrictions imposed in the semiparametric literature, we assume that the function $\mathbf{g}$ is invertible in the vector of indices, a restriction that is satisfied for reduced form parameters of many econometric models in which the latent error terms are assumed to be independent of the regressors. The proposed estimator of $\boldsymbol{\theta}_0$ is the eigenvector of a random matrix $\hat{\mathbf{S}}$ corresponding to its smallest (in magnitude) eigenvalue, making it relatively simple to compute. As described in more detail below, the index coefficient estimator is a semiparametric analogue of Berkson's (1955) and Amemiya's (1976) minimum chi-square logit estimators, with the reduced form parameters being nonparametrically estimated – here, using general nonparametric regression methods instead of cell means for regressors with finite support. This paper consolidates the results in earlier working papers by Ahn, Ichimura, and Powell (1996), which considered single-index models, and by Powell and Ruud (2008), which considered estimation of multiple indices for multinomial choice models. It also extends those previous results to "overdetermined" models, i.e., invertible models in which the dimension of the reduced-form vector $\boldsymbol{\gamma}(\mathbf{X})$ exceeds the dimension of the index vector $\mathbf{X}\boldsymbol{\beta}_0$. Blundell and Powell (2004) considered a different extension of the Ahn-Ichimura-Powell estimation approach to binary response models with endogenous regressors, and that same extension could be adapted to the multi-index models considered here.

While a large literature exists for estimation of semiparametric single-index regression models – besides Ahn, Ichimura, and Powell (1996), Han (1987), Härdle and Stoker (1989), Härdle and Horowitz (1996), Ichimura (1993), Klein and Spady (1993), Manski (1975, 1985), Newey and Ruud (2005), and Powell, Stock and Stoker (1989), among many others – there are fewer results available on estimation of multiple-index regression models. Ichimura and Lee (1992) consider estimation of a semiparametric sample selection model with multiple indices in the selection equation, and Lee (1995) constructs a multinomial analogue to Klein and Spady's (1993) estimator of the semiparametric binary choice model, estimating the index coefficients by minimizing a semiparametric "profile likelihood" constructed using nonparametric estimators of the choice probabilities as functions of the indices. Lee demonstrates semi-

parametric efficiency of the estimator under the "distributional index" restriction that the conditional distribution of the errors depends on the regressors only through the indices. As Thompson (1993) shows, though, the semiparametric efficiency bound for multinomial choice under the assumption of independence of the errors – which coincides with the bound the weaker distributional index restriction for binary choice – differs when the number of choices exceeds two, indicating possible efficiency improvements from the stronger independence restriction. Ruud (2000) shows that the stronger independence restrictions yield choice probabilities that are invertible functions of the indices and whose derivative matrix (with respect to the indices) is symmetric, neither of which needs hold under only the distributional index restriction, and Berry, Gandhi, and Haile (2013) give general conditions on multivariate demand models that insure their invertibility with respect to the underlying latent utilities (here assumed to be linear in the regressors).

The next section defines the class of index models considered and the proposed estimator of the index model coefficients, and discusses how three examples–single-index regression, multinomial choice, and monotonic location-scale quantile regression–fit into the general framework. The large-sample behavior of the proposed estimator (root-N consistency and asymptotic normality) is derived in Section 3, and efficient choice of the "weight matrix" defining the estimator is discussed in Section 4.

2. The Model and Estimator

General Framework

The model for index restrictions considered here assumes that we have a random sample of N observations on a random vector $\mathbf{y}_i$ of dependent variables and a matrix $\mathbf{X}_i$ of random covariates (with q rows and $p+1$ columns; the dimension of $\mathbf{y}_i$ plays no direct role in the analysis). For this setup, a key restriction is that some r-dimensional parameter $\boldsymbol{\gamma}_i \equiv \boldsymbol{\gamma}(\mathbf{X}_i)$ of the conditional distribution of $\mathbf{y}_i$ given $\mathbf{X}_i$ (the "reduced form parameter vector") depends only on a q-dimensional linear combination $\mathbf{X}_i'\boldsymbol{\theta}_0$ of the regressors (the "structural index vector"), with the $(p+1)$-dimensional coefficient vector $\boldsymbol{\theta}_0$ being the unknown parameter of interest. That is, for some mapping

$$\boldsymbol{\gamma} \equiv \boldsymbol{\gamma}(\mathbf{X}) \equiv \mathbf{T}\left(F_{\mathbf{y}|\mathbf{X}}\left(\mathbf{y}|\mathbf{X}\right)\right) \tag{2.1}$$

of the conditional distribution $F_{\mathbf{y}|\mathbf{X}}$ of $\mathbf{y}_i$ given $\mathbf{X}_i$, the restriction

$$\boldsymbol{\gamma}_i \equiv \boldsymbol{\gamma}(\mathbf{X}_i) = \mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0) \tag{2.2}$$

holds with probability one for some smooth (unknown) function $\mathbf{g}: R^q \to R^r$.Another key restriction imposed here is that the function $g(\cdot)$ has a (left) inverse, so that

$$\begin{aligned} \mathbf{m}(\boldsymbol{\gamma}_i) &\equiv \mathbf{g}^{-1}(\boldsymbol{\gamma}_i) \\ &= \mathbf{X}_i\boldsymbol{\theta}_0 \equiv \boldsymbol{\mu}_i. \end{aligned} \tag{2.3}$$

With the smoothness restriction on $\mathbf{g}(\cdot)$, existence of the left-inverse function $\mathbf{m}(\cdot)$ requires $r \geq q$, i.e., at least as many reduced form parameters as structural indicies. We call the case with $r = q$ "just-determined" and with $r > q$ "overdetermined," in analogy with the traditional distinction between "just-" and "over-" identification in the traditional simultaneous equations literature.

Examples

A number of semiparametric models studied in econometrics fit into this framework. The *single index regression* literature has extensively considered the restriction (2.2) when $\boldsymbol{\gamma}(\mathbf{X}_i) = E[\mathbf{y}_i|\mathbf{X}_i]$ and $r = q = 1$, and a main objective here is to extend the analysis to multiple reduced form parameters ($r > 1$) and multiple indices ($q > 1$). A multivariate extension that generates the restrictions in (2.2) and (2.3) is the *semiparametric multinomial choice (MNC)* model under the assumption of independence of the errors and regressors. In this case the q-dimensional dependent variable $\mathbf{y}_i$ is a vector of indicator variables denoting which of $q+1$ mutually-exclusive and exhaustive alternatives (numbered from $j = 0$ to $j = q$) is chosen. Specifically, for individual i, alternative j is assumed to have an unobservable indirect utility y^*_{ij} for that individual, and the alternative with the highest indirect utility is assumed chosen. Thus an individual component y_{ij} of the vector $\mathbf{y}_i$ has the form

$$y_{ij} = 1\{y^*_{ij} \geq y^*_{ik} \text{ for } k = 0, ..., q\}, \tag{2.4}$$

with the convention that $\mathbf{y}_i = \mathbf{0}$ indicates choice of alternative $j = 0$. An assumption of joint continuity of the indirect utilities rules out ties (with probability one); in this model, the indirect utilities are further restricted to have the linear form

$$y^*_{ij} = \mathbf{x}'_{ij}\boldsymbol{\theta}_0 + \varepsilon_{ij} \tag{2.5}$$

for $j = 1, ..., q$, where the vector $\boldsymbol{\varepsilon}_i$ of unobserved error terms is assumed to be jointly continuously distributed and independent of the $q \times (p+1)$-dimensional matrix of regressors $\mathbf{X}_i$ (whose j^{th} row is $\mathbf{x}'_{ij}$). (For alternative $j = 0$, the standard normalization $y^*_{i0} = 0$ is imposed.) As Lee (1995) notes, the MNC model with independent errors restricts the conditional choice probabilities to depend upon the regressors only through the vector $\boldsymbol{\mu}_i \equiv \mathbf{X}_i\boldsymbol{\theta}_0$ of linear indices; that is, it takes the form

$$E[\mathbf{y}_i|\mathbf{X}_i] \equiv \boldsymbol{\gamma}_i = \mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0) \tag{2.6}$$

for some unknown function $\mathbf{g}(\cdot)$, so that

$$\mathbf{y}_i = \boldsymbol{\gamma}_i + \mathbf{u}_i = \mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0) + \mathbf{u}_i, \tag{2.7}$$

where $E[\mathbf{u}_i|\mathbf{X}_i] = \mathbf{0}$ by construction. In addition, the assumption of independence of the latent disturbances $\boldsymbol{\varepsilon}_i$ and the regressors $\mathbf{X}_i$ implies that the function $\mathbf{g}(\boldsymbol{\mu})$ is smooth and invertible in its argument $\boldsymbol{\mu}$ if $\boldsymbol{\varepsilon}_i$ has nonnegative density everywhere (Ruud 2000). A weaker condition yielding (2.6) is an assumption that the conditional distribution of $\boldsymbol{\varepsilon}_i$ given $\mathbf{X}_i$ only depends upon the vector of indices $\mathbf{X}_i\boldsymbol{\theta}_0$, but under this restriction the function $\mathbf{g}$ needs not be invertible in its argument, so invertibility would need to be imposed as an additional restriction for the method proposed here to apply.

A different, somewhat artificial example is a model that restricts several conditional quantiles of a scalar dependent variable y_i to depend upon only two indices, one for location and one for scale. Specifically, suppose the dependent variable y_i is determined through the structural equation

$$\begin{aligned} y_i &= \alpha\left(\mathbf{x}'_{i1}\boldsymbol{\theta}_1 + \sigma(\mathbf{x}'_{i2}\boldsymbol{\theta}_2)\varepsilon_i\right) \\ &= \alpha\left(\boldsymbol{\mu}_{i1} + \sigma(\boldsymbol{\mu}_{i2})\varepsilon_i\right), \end{aligned} \tag{2.8}$$

where the unknown functions $\alpha(\cdot)$ and $\sigma(\cdot)$ are assumed to be strictly increasing. Suppose the unobservable error ε_i is assumed to be continuously distributed given $\mathbf{x}_i$, and that r conditional quantiles are assumed to be independent of $\mathbf{x}_i$, i.e.,

$$\Pr\{\varepsilon_i \leq \eta_j|\mathbf{x}_i\} = \tau_j \tag{2.9}$$

for some fractions $0 < \tau_1 < ... < \tau_j < ... < \tau_r < 1$ and corresponding quantiles $\{\eta_j\}$ (assumed unique). Then the corresponding conditional quantiles of y_i given $\mathbf{x}_i$ satisfy

$$\begin{aligned} \gamma_j(\mathbf{x}_i) &\equiv F^{-1}_{y|\mathbf{x}}(\tau_j|\mathbf{x}_i) \\ &= \alpha\left(\mathbf{x}'_{i1}\boldsymbol{\theta}_1 + \sigma(\mathbf{x}'_{i2}\boldsymbol{\theta}_2)\eta_j\right) \end{aligned}$$

for $j = 1, ..., r$. As long as $r \geq 2$, the indices $\mu_{i1} = \mathbf{x}_{i1}'\boldsymbol{\theta}_1$ and $\mu_{i2} = \mathbf{x}_{i2}'\boldsymbol{\theta}_2$ can be written as functions of any two distinct conditional quantiles γ_{ij} and γ_{ik} of y_i:

$$\begin{aligned}\mu_{i1} &= \frac{\eta(\tau_j)\alpha^{-1}(\gamma_{ik}) - \eta(\tau_k)\alpha^{-1}(\gamma_{ij})}{\eta(\tau_j) - \eta(\tau_k)}, \\ \mu_{i2} &= \sigma^{-1}\left(\frac{\alpha^{-1}(\gamma_{ij}) - \alpha^{-1}(\gamma_{ik})}{\eta(\tau_j) - \eta(\tau_k)}\right),\end{aligned}$$

so that the (unknown) transformation from the indices to the reduced form parameters is indeed invertible. Of course, a natural assumption generating the identifying restriction (2.9) would be statistical independence of ε_i and $\mathbf{x}_i$, in which case the number of reduced-form quantiles r could be taken to be arbitrarily large, but here we take $r \geq 2$ to be fixed and defer efficiency considerations for this condition.

Identification

Returning to the general restrictions (2.1) through (2.3), if the inverse transformation $\mathbf{m}(\boldsymbol{\gamma}) = \mathbf{g}^{-1}(\boldsymbol{\gamma})$ $(= \mathbf{X}\boldsymbol{\theta}_0)$ were known, the coefficient vector $\boldsymbol{\theta}_0$ would be identified if any of the rows of the regressor matrix $\mathbf{X}$ had a full-rank covariance matrix. Furthermore, given a consistent estimator $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ and smoothness of the inverse transformation, the parameter vector $\boldsymbol{\theta}_0$ could be consistently estimated using generalized least-squares applied to the generated regression equation

$$\begin{aligned}\hat{\boldsymbol{\mu}}_i &\equiv \mathbf{g}^{-1}(\hat{\boldsymbol{\gamma}}_i) \equiv \mathbf{m}(\hat{\boldsymbol{\gamma}}_i) \\ &= \mathbf{X}_i\boldsymbol{\theta}_0 + [\mathbf{m}(\hat{\boldsymbol{\gamma}}_i) - \mathbf{m}(\boldsymbol{\gamma}_i)] \\ &\simeq \mathbf{X}_i\boldsymbol{\theta}_0 + \left[\frac{\partial \mathbf{m}(\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}'}\right](\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) \\ &\equiv \mathbf{X}_i\boldsymbol{\theta}_0 + \hat{\mathbf{u}}_i. \qquad (2.10)\end{aligned}$$

This estimation method was proposed by Berkson (1955) for the binary logit model, where the reduced form estimator $\hat{\boldsymbol{\gamma}}_i$ was a cell mean for the response probability when the regressors are constant within cells, and the method was generalized by Amemiya (1976) to general parametric multinomial choice models.

However, in the present setting where the inverse transformation $\mathbf{m}(\boldsymbol{\gamma}) = \mathbf{g}^{-1}(\boldsymbol{\gamma})$ is unknown, indentification of $\boldsymbol{\theta}_0$ is more delicate. The coefficient vector $\boldsymbol{\theta}_0$ is clearly only identified up to scale at best. and it will be convenient to normalize the first component of $\boldsymbol{\theta}_0$ to be one, with

$$\boldsymbol{\theta}_0 \equiv \begin{pmatrix} 1 \\ -\boldsymbol{\beta}_0 \end{pmatrix}, \qquad (2.11)$$

so that the object of estimation is the p-dimensional parameter $\boldsymbol{\beta}_0$. Writing the regressor matrix $\mathbf{X}_i \equiv [\mathbf{X}_{i1}, \mathbf{X}_{i2}]$, with the column vector $\mathbf{X}_{i1}$ corresponding to the normalized component in $\boldsymbol{\theta}_0$, the relation (2.3) can be rewritten as

$$\begin{aligned} \mathbf{m}(\boldsymbol{\gamma}_i) &= \mathbf{X}_i\boldsymbol{\theta}_0 \\ &= \mathbf{X}_{i1} - \mathbf{X}_{i2}\boldsymbol{\beta}_0 \\ &\Longrightarrow \\ \mathbf{X}_{i1} &= \mathbf{X}_{i2}\boldsymbol{\beta}_0 + \mathbf{m}(\boldsymbol{\gamma}_i) \\ &= \mathbf{X}_{i2}\boldsymbol{\beta}_0 + \mathbf{m}(\hat{\boldsymbol{\gamma}}_i) - \hat{\mathbf{u}}_i. \end{aligned} \tag{2.12}$$

In this representation the normalized regressor $\mathbf{X}_{i1}$ takes the form of the conditional mean of a vector of dependent variables from a semilinear regression model, with $\mathbf{X}_{i2}$ the regressors in the linear part and the identified $\boldsymbol{\gamma}_i$ being the regressors in the nonparametric component.

Identification of $\boldsymbol{\beta}_0$ for such models, which was considered by Robinson (1988) and Ahn and Powell (1993), requires sufficient variability in the rows of $\mathbf{X}_i$ given $\boldsymbol{\gamma}_i$; since $\mathbf{0} = \mathbf{Var}\left[\mathbf{X}_i\boldsymbol{\theta}_0|\boldsymbol{\gamma}_i\right]$, the parameter vector $\boldsymbol{\theta}_0$ is identified (up to scale) if $\mathbf{Var}\left[\mathbf{X}_i\boldsymbol{\delta}|\boldsymbol{\gamma}_i\right] \neq \mathbf{0}$ with positive probability unless $\boldsymbol{\delta}$ is proportional to $\boldsymbol{\theta}_0$. As discussed in detail by Lee (1995), this requirement can be ensured if the component vector $\mathbf{X}_{i1}$ is jointly continuously distributed on R^q conditional on the remaining components $\mathbf{X}_{i2}$ whose rows have full-rank covariance matrices. We impose this assumption in our analysis, which implies that the index $\boldsymbol{\mu}_i$ is jointly continuouslly distributed.

Following Ahn, Ichimura, and Powell (2004), the alternative approach to identification (and estimation) of $\boldsymbol{\theta}_0$ taken here uses the fact that, for values of $\boldsymbol{\gamma}_i$ that are (nearly) equal, the corresponding values of $\mathbf{X}_i\boldsymbol{\beta}_0$ will also be (nearly) equal. This implies that the vector $\boldsymbol{\theta}_0$ can be identified by matching observations (numbered i and j) with the same reduced form parameters $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j$ but different matrices of regressors. Conditional on

$$\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \tag{2.13}$$

relation (2.3) implies that

$$(\mathbf{X}_i - \mathbf{X}_j)\boldsymbol{\theta}_0 = \mathbf{0}. \tag{2.14}$$

It follows that

$$E[\mathbf{L}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\,|\,\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j]\boldsymbol{\theta}_0 = \mathbf{0} \tag{2.15}$$

for any random $(p+1) \times q$ matrix $\mathbf{L}_{ij}$ for which $E||\mathbf{L}_{ij}(\mathbf{X}_i - \mathbf{X}_j)||$ exists. A convenient class of such matrices is

$$\mathbf{L}_{ij} = (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{W}_{ij}, \tag{2.16}$$

for some suitable $q \times q$, nonnegative-definite "weight/trimming" matrix $\mathbf{W}_{ij} \equiv \mathbf{W}(\mathbf{X}_i, \mathbf{X}_j)$; this implies that the (identified) $(p+1) \times (p+1)$ matrix

$$\begin{aligned} \boldsymbol{\Sigma}_0 &\equiv E[(\mathbf{X}_i - \mathbf{X}_j)\mathbf{W}_{ij}(\mathbf{X}_i - \mathbf{X}_j) \,|\, \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j] \\ &= E[(\mathbf{X}_i - \mathbf{X}_j)\mathbf{W}_{ij}(\mathbf{X}_i - \mathbf{X}_j) \,|\, \boldsymbol{\mu}_i = \boldsymbol{\mu}_j] \end{aligned} \tag{2.17}$$

has

$$\boldsymbol{\theta}_0'\boldsymbol{\Sigma}_0\boldsymbol{\theta}_0 = \mathbf{0}, \tag{2.18}$$

that is, $\boldsymbol{\Sigma}_0$ has a zero eigenvalue with corresponding eigenvector equal to the true parameter $\boldsymbol{\theta}_0$. If the matrix $\mathbf{W}_{ij}$ is chosen so that the zero eigenvalue of $\boldsymbol{\Sigma}_0$ is unique – which requires a sufficiently rich support of the conditional distribution of $\mathbf{X}_i$ given $\mathbf{X}_i'\boldsymbol{\theta}_0$ – this suffices to identify $\boldsymbol{\theta}_0$ up to scale as the unique nontrivial solution to (2.18).

As discussed below, the weight/trimming matrix can be chosen for technical convenience and/or to improve the asymptotic efficiency of the corresponding estimator of $\boldsymbol{\beta}_0$. Because the index vector $\boldsymbol{\mu}_i$ will be jointly continuously distributed under our assumptions (to help ensure that $Var[\tilde{\mathbf{X}}_i\boldsymbol{\delta}] \neq \mathbf{0}$ for any nontrivial $\boldsymbol{\delta} \neq \boldsymbol{\theta}_0$), the conditioning event in the definition of $\boldsymbol{\Sigma}_0$ in (2.17) has zero probability, but we assume the relevant conditional expectations are sufficiently smooth so that $\boldsymbol{\Sigma}_0$ can be expressed as

$$\boldsymbol{\Sigma}_0 = \lim_{\varepsilon \to 0} E[(\mathbf{X}_i - \mathbf{X}_j)\mathbf{W}_{ij}(\mathbf{X}_i - \mathbf{X}_j)| \left\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\right\| < \varepsilon],$$

as usual for nonparametric regression problems.

Given a random sample of size N from this model, the preceding identification approach can be transformed into an estimation method for $\boldsymbol{\theta}_0$ by first estimating the unobservable reduced form vectors $\boldsymbol{\gamma}_i$ by some nonparametric method and then, as in Ahn, Ichimura, and Powell (1996), estimating a sample analogue to the matrix $\boldsymbol{\Sigma}_0$ using pairs of observations with estimated values $\hat{\boldsymbol{\gamma}}_i$ of $\boldsymbol{\gamma}_i$ that were approximately equal. For the multinomial choice example, where $\boldsymbol{\gamma}_i = E[\mathbf{y}_i|\mathbf{x}_i]$, the first-step nonparametric estimator of $\boldsymbol{\gamma}_i$ may be the familiar kernel regression estimator, which takes the form of a weighted average of the dependent variable,

$$\hat{\boldsymbol{\gamma}}_i \equiv \frac{\sum_{j=1}^N k_{ij} \cdot \mathbf{y}_j}{\sum_{i=1}^N k_{ij}}, \tag{2.19}$$

with weights k_{ij} given by

$$k_{ij} \equiv k\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h_1}\right), \tag{2.20}$$

for $\mathbf{x}_i$ a vector of the distinct components of the matrix $\mathbf{X}_i$, $k(\cdot)$ a "kernel" function which tends to zero as the magnitude of its argument increases, and $h_1 \equiv h_{1N}$ a first-step "bandwidth" which is chosen to tend to zero as the sample size n increases. More generally, the reduced form vector $\hat{\boldsymbol{\gamma}}_i$ may include other nonparametric estimates of the parameters of the conditional distribution of $\mathbf{y}_i$ given $\mathbf{X}_i$ (e.g., conditional quantiles, distribution functions, or higher moments) and can be estimated using alternatives to kernel estimation (e.g., nearest neighbor, local polynomial, splines, series, or sieve estimators). However it is constructed, we will assume the reduced form estimator $\hat{\boldsymbol{\gamma}}_i$ has an "asymptotically linear" representation of the form

$$\hat{\boldsymbol{\gamma}}_i = \boldsymbol{\gamma}_i + \frac{1}{N}\sum_{j=1}^{N} \mathbf{R}_{jN}(\mathbf{X}_i)\boldsymbol{\xi}_j + \mathbf{r}_{iN}, \tag{2.21}$$

where the i.i.d. "error term" $\boldsymbol{\xi}_j$ has $E\left[\boldsymbol{\xi}_j|\mathbf{X}_j\right] = 0$, $Var\left[\boldsymbol{\xi}_j|\mathbf{X}_j\right] \equiv \mathbf{V}_j$ and $E[\left\|\boldsymbol{\xi}_j\right\|^2] < \infty$, the matrix $\mathbf{R}_{jN}(\mathbf{X}_i)$ has second moment $E[\|\mathbf{R}_{jN}(\mathbf{X}_i)\|^2]$ that increases at a slower rate than $\sqrt{N}$, i.e.,

$$E[\|\mathbf{R}_{jN}(\mathbf{X}_i)\|^2] = o(\sqrt{N}) \tag{2.22}$$

and the remainder term $\mathbf{r}_{in}$ is negligible when multiplied by $\sqrt{N}$ (as described in the next section). For the kernel estimator example given in (2.19), given appropriate side conditions on the kernel and bandwidth terms, the components $\boldsymbol{\xi}_i$ and $\mathbf{R}_{jN}(\mathbf{X}_i)$ will take the form $\boldsymbol{\xi}_i = \mathbf{y}_i - E[\mathbf{y}_i|\mathbf{X}_i] = \mathbf{y}_i - \boldsymbol{\gamma}_i$ and $\mathbf{R}_{jN}(\mathbf{X}_i) = (h_1)^{-c}k_{ij}/f_{\mathbf{X}}(\mathbf{x}_i)$, where c is the number of continuously-distributed components of $\mathbf{x}_i$ and $f_{\mathbf{X}}$ is the joint density function of $\mathbf{x}_i$ (the product of the conditional density function of the continuous components given the discrete components and the probability mass function for the discrete components).

Given this nonoparametric estimator $\hat{\boldsymbol{\gamma}}_i$ of the reduced form parameter vector $\boldsymbol{\gamma}_i$, a second-step estimator of a matrix analogue to $\boldsymbol{\Sigma}_0$ can be constructed as

$$\hat{\mathbf{S}} \equiv \begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{h^q}\, K\left(\frac{\hat{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_j}{h}\right) \cdot (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j), \tag{2.23}$$

where $K(\cdot)$ is a univariate kernel analogous to k above, $h \equiv h_N$ is a bandwidth sequence for the second-step estimator $\hat{\mathbf{S}}$, and $\mathbf{A}_{ij} = \mathbf{A}_N(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{A}_{ji}$ is a $q \times q$, symmetric, nonnegative-definite

"weight/trimming" matrix which is constructed to equal zero for observations where $\hat{\boldsymbol{\gamma}}_i$ or $\hat{\boldsymbol{\gamma}}_j$ is imprecise (i.e., where $\mathbf{X}_i$ or $\mathbf{X}_j$ is outside some compact subset of its support). The term $K\left((\hat{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_j)/h\right)$ declines to zero as $\hat{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_j$ increases relative to the bandwidth h; thus, the conditioning event "$\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j$" in the definition of $\boldsymbol{\Sigma}_0$ is ultimately imposed as this bandwidth shrinks with the sample size (and the nonparametric estimator of $\boldsymbol{\gamma}_i$ converges to its true value in probability). It is worth noting that, even though r, the number of components of $\hat{\gamma}_i$, may exceed the number of indices q, the normalizing constant for the kernel weight in (2.23) is h^{-q} rather than h^{-r}, reflecting the assumption that the true reduced form parameters $\boldsymbol{\gamma}_i$ have a continous distribution only on an r-dimensional manifold in q-dimensional Euclidean space (since they depend only on the q-dimensional index vector $\boldsymbol{\mu}_i$).

Given the estimator $\hat{\mathbf{S}}$ of $\boldsymbol{\Sigma}_0$ – which corresponds to a particular structure for the population weight matrix $\mathbf{W}_{im}$ in the definition of $\boldsymbol{\Sigma}_0$, as discussed below – construction of an estimator of $\boldsymbol{\beta}_0$ follows exactly the same form as in Ahn, Ichimura, and Powell (1996) and Blundell and Powell (2004), exploiting a sample analogue of relation (2.17) based on the eigenvalues and eigenvectors of $\hat{\mathbf{S}}$. Since the estimator $\hat{\mathbf{S}}$ may not be in finite samples if the kernel function $K(\cdot)$ is not constrained to be nonnegative, the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is defined here as the eigenvector for the eigenvalue of $\hat{\mathbf{S}}$ that is closest to zero in magnitude. That is, defining $(\hat{\nu}_1, ..., \hat{\nu}_{p+1})$ to be the $p+1$ solutions to the determinantal equation

$$|\hat{\mathbf{S}} - \nu \mathbf{I}| = 0, \tag{2.24}$$

the estimator $\hat{\boldsymbol{\theta}}$ is defined as an appropriately-normalized solution to

$$(\hat{\mathbf{S}} - \hat{\nu}\mathbf{I})\hat{\boldsymbol{\theta}} = \mathbf{0}, \tag{2.25}$$

where

$$\hat{\nu} \equiv \arg\min_j \{|\hat{\nu}_j|\}. \tag{2.26}$$

Normalizing the first component of $\boldsymbol{\theta}_0$ to zero, with remaining coefficients defined as -$\boldsymbol{\beta}_0$, i.e.,

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} 1 \\ -\hat{\boldsymbol{\beta}} \end{pmatrix}, \qquad \boldsymbol{\theta}_0 = \begin{pmatrix} 1 \\ -\boldsymbol{\beta}_0 \end{pmatrix}, \tag{2.27}$$

and partitioning $\hat{\mathbf{S}}$ conformably as

$$\hat{\mathbf{S}} \equiv \begin{bmatrix} \hat{\mathbf{S}}_{11} & \hat{\mathbf{S}}_{12} \\ \hat{\mathbf{S}}_{21} & \hat{\mathbf{S}}_{22} \end{bmatrix}, \tag{2.28}$$

the solution to (2.25) takes the form

$$\hat{\boldsymbol{\beta}} \equiv [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1} \cdot \hat{\mathbf{S}}_{21} \tag{2.29}$$

for this normalization. An alternative "closed form" estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ can be defined as

$$\tilde{\boldsymbol{\beta}} \equiv [\hat{\mathbf{S}}_{22}]^{-1} \cdot \hat{\mathbf{S}}_{21}; \tag{2.30a}$$

this estimator exploits the fact (to be verified below) that $\hat{\nu}$ tends to zero in probability, since the smallest eigenvalue of the probability limit $\boldsymbol{\Sigma}_0$ of $\hat{\mathbf{S}}$ is zero.

As discussed by Ahn, Ichimura, and Powell (2004), the relation of $\hat{\boldsymbol{\beta}}$ to $\tilde{\boldsymbol{\beta}}$ here is analogous to the relationship between the two classical single-equation estimators for simultaneous equations systems, namely, limited-information maximum likelihood, which has an alternative derivation as a least-variance ratio (LVR) estimator, and two-stage least squares (2SLS), which can be viewed as a modification of LVR which replaces an estimated eigenvalue by its known (zero) probability limit. The analogy to these classical estimators extends to the asymptotic distribution theory for $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, which, under the conditions imposed below, will be asymptotically equivalent, like LVR and 2SLS. Their relative advantages and disadvantages are also analogous – e.g., $\tilde{\boldsymbol{\beta}}$ is slightly easier to compute, while $\hat{\boldsymbol{\beta}}$ will be equivariant with respect to choice of which (nonzero) component of $\boldsymbol{\theta}_0$ to normalize to unity. As noted earlier, both estimators are semiparametric analogues to the minimum chi-square estimators of Berkson (1955) and Amemiya (1976), using a particular semilinear regression estimator applied to (2.12), which treats the $\mathbf{g}^{-1}(\cdot)$ function as the unknown nonparametric component and the first component of $\mathbf{X}_i$ (with coefficient normalized to unity) as the dependent variable. The proposed estimator will be easier to calculate than Lee's (1995) estimator for the multinomial response model under index restrictions, which requires solution of a r-dimensional minimization problem with a criterion involving simultaneous estimation of J nonparametric regressions (with the J index functions as arguments) and minimization over the parameter vector $\boldsymbol{\beta}$. Unlike the estimator proposed here, however, Lee's estimator does not impose invertibility of the vector of choice probabilities in the vector of indices.

3. Large Sample Properties of the Estimator

Since the definition of the estimator $\hat{\boldsymbol{\theta}} = (1, -\hat{\boldsymbol{\beta}}')'$ is based on the same form of a "pairwise difference" matrix estimator $\hat{\mathbf{S}}$ analyzed in Ahn, Ichimura, and Powell (2004) and Blundell and Powell (2004),

the regularity conditions imposed here will be quite similar to those imposed in these earlier papers. Rather than restrict the first-stage nonparametric estimator of the choice probabilities $\boldsymbol{\gamma}(\mathbf{x}_i)$ to have a particular form (e.g., the kernel estimator in (2.19)), it is assumed to satisfy some higher-level restrictions on its rate of convergence and Bahadur representation which would need to be verified for the particular nonparametric estimation method utilized. Specifically, given a random sample of size N for $\{\mathbf{y}_i, \mathbf{X}_i\}$, it is assumed that the reduced-form estimator $\hat{\boldsymbol{\gamma}}(\mathbf{X}_i)$ has a relatively high convergence rate and the same asymptotic linear representation as for a kernel estimator. That is, defining the "trimming" indicator

$$t_{ij} \equiv 1\{||\mathbf{A}_N(\mathbf{X}_i, \mathbf{X}_j)|| \neq \mathbf{0}\}, \tag{3.31}$$

the condition

$$\max_{i,j} t_{ij}||\hat{\boldsymbol{\gamma}}(\mathbf{X}_i) - \boldsymbol{\gamma}(\mathbf{X}_i)|| = o_p(N^{-3/8}) \tag{3.32}$$

is imposed. This is a restriction on both the nonparametric estimation and the construction of the weight-matrix function $\mathbf{A}_N(\mathbf{X}_i, \mathbf{X}_j)$, which will generally requiring "trimming" of observations outside a bounded set of $\mathbf{X}_i$ values to ensure that (3.32) is satisfied. Similarly, the remainder term $\mathbf{r}_{iN}$ in the asymptotic linear representation (2.21) for the reduced-form estimator is assumed to converge to zero at a rate faster than $\sqrt{N}$ uniformly in i and j,

$$\max_{i,j} t_{ij}||\mathbf{r}_{iN}|| = o_p(N^{-3/4}). \tag{3.33}$$

Another restriction on the model is that the first column $\mathbf{x}_{i1}$ of the matrix of the regressors $\mathbf{X}_i$ is continuously distributed conditionally on the remaining components, and that the corresponding coefficient $\boldsymbol{\beta}_{0,1}$ is nonzero (and normalized to unity); as discussed by Lee (1995), this restriction helps ensure that the parameters are identified by ensuring that $\mathbf{X}_i$ is sufficiently variable conditional upon a given value of the reduced form vector $\boldsymbol{\gamma}_i = \mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0)$. Other conditions are imposed on the error distribution, kernel function K, and bandwidth h; a discussion of the relevant regularity conditions, which are extensions of conditions imposed in Ahn, Ichimura, and Powell (2004) and other single-index regression papers, is given in the appendix below.

Under the assumptions in the appendix, consistency of the estimator $\hat{\mathbf{S}}$ for a particular matrix $\boldsymbol{\Sigma}_0$ can be established. That is,

$$\hat{\mathbf{S}} \xrightarrow{p} \boldsymbol{\Sigma}_0, \tag{3.34}$$

where $\mathbf{\Sigma}_0$ is of the form given in (4.1) with

$$\mathbf{W}_{ij} = \lim_{N\to\infty} \mathbf{A}_{ij}\sqrt{\phi_i \phi_j}, \tag{3.35}$$

where the term ϕ_i (referenced here as the "gamma density") depends on the joint density $f_{\boldsymbol{\mu}}$ of the indices μ_i, the Jacobian matrix

$$\underset{(r\times q)}{\mathbf{G}_i} \equiv \frac{\partial \mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0)}{\partial \boldsymbol{\mu}'} = \left[\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\mu}'}\right]_i, \tag{3.36}$$

and the kernel function $K(\cdot)$ as follows:

$$\underset{(1\times 1)}{\phi_i} \equiv \delta_i \cdot f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i), \tag{3.37}$$

with

$$\underset{(1\times 1)}{\delta_i} \equiv \int_{R^q} K(\mathbf{G}_i\mathbf{u})d\mathbf{u} \tag{3.38}$$

being an "inverse Jacobian determinant" term. In the overdetermined case, the form of the kernel function $K(\cdot)$ affects the probability limit of $\hat{\mathbf{S}}$, but when $r = q$ the term (3.38) reduces to

$$\underset{(1\times 1)}{\delta_i} \equiv |G_i|^{-1} = \left|\frac{\partial \mathbf{m}(\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}'}\right|,$$

and the term ϕ_i becomes the joint density function of $\boldsymbol{\gamma}(\mathbf{X})$ evaluated at the realized value $\boldsymbol{\gamma}_i = \mathbf{g}(\mathbf{X}_i)$.

The consistency of $\hat{\mathbf{S}}$ for $\mathbf{\Sigma}_0$, along with the identification restriction that the true coefficient vector $\boldsymbol{\theta}_0$ is the unique solution of (2.17), implies consistency of the corresponding estimator $\hat{\boldsymbol{\theta}}$ up to scale. In order to derive the asymptotic normal representation for the unnormalized coefficients $\hat{\boldsymbol{\beta}}$,the regularity conditions also yield an asymptotically-linear representation for $\hat{\mathbf{S}}\boldsymbol{\theta}_0$ of the form

$$\sqrt{N}\hat{\mathbf{S}}\boldsymbol{\theta}_0 = \frac{2}{\sqrt{N}}\sum_{\iota=1}^{N} \phi_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{A}}_i\mathbf{M}_i\boldsymbol{\xi}_i + o_p(1), \tag{3.39}$$

where $\boldsymbol{\xi}_i$ is the reduced-form error term given in (2.21) and ϕ_i is the "gamma density" term defined in 3.37). The remaining terms in (3.39) are defined as follows:

$$\underset{(q\times r)}{\mathbf{M}_i} \equiv \delta_i^{-1}\left(\int_{R^q} \mathbf{u}\cdot\left[\frac{\partial K(\mathbf{G}_i\mathbf{u})}{\partial \mathbf{v}'}\right]d\mathbf{u}\right), \tag{3.40}$$

is an "inverse Jacobian matrix" term (with δ_i defined in (3.38);

$$\underset{(q\times q)}{\tilde{\mathbf{A}}_i} \equiv \lim_{N\to\infty} E\left[\mathbf{A}_{ij} \mid \mathbf{X}_i, \boldsymbol{\mu}_j\right]_{\boldsymbol{\mu}_j = \mathbf{X}_i\boldsymbol{\theta}_0} \tag{3.41}$$

is the "limiting weight matrix;" and

$$\underset{((p+1)\times 1)}{\tilde{\mathbf{X}}_i} \equiv \mathbf{X}_i - \tilde{\mathbf{A}}_i^{-1} \left(\lim_{N\to\infty} E\left[\mathbf{A}_{ij}\mathbf{X}_j \mid \mathbf{X}_i, \boldsymbol{\mu}_j\right]_{\boldsymbol{\mu}_j = \mathbf{X}_i\boldsymbol{\theta}_0} \right) \tag{3.42}$$

is a "residual regressor matrix."

Of the terms appearing in the Bahadur representation (3.39), the "inverse Jacobian matrix" $\mathbf{M}_i$ in (3.40) is perhaps the most obscure, because of its dependence on the kernel function $K(\cdot)$ (jusr like the "inverse Jacobian determinant" δ_i in (3.38)). In the just-determined case $r = q$, the matrix $\mathbf{M}_i$ will insdeed be independent of $K(\cdot)$, reducing to $\partial\mathbf{m}(\boldsymbol{\gamma}_i)/\partial\boldsymbol{\gamma}' = \mathbf{G}_i^{-1}$, the derivative of the inverse transformation between the reduced form parameters $\boldsymbol{\gamma}_i$ and index vector $\boldsymbol{\mu}_i$. However, in the overdetermined case $r > q$ this matrix reflects the effect of the r-dimensional deviations $\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_\iota$ of the reduced form estimators from their true values on the corresponding deviations $\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i$ in the index estimates, an effect by the nature of the kernel weights $K((\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_\iota)/h)$. For concreteness, suppose the kernel function takes the "elliptically symmetric" form

$$K(\mathbf{v}) = c_r k(\mathbf{v}'\mathbf{Q}\mathbf{v}) \left|\mathbf{Q}\right|^{1/2}, \tag{3.43}$$

where $k(v)$ is a scalar kernel, $\mathbf{Q}$ is a nonnegative-definite matrix and c_r is a normalizing constant depending only on k and the dimension r of the argument $\mathbf{v}$; in this case, the "inverse Jacobian deternuinant" will reduce to

$$\delta_i = \frac{c_r}{c_q} \frac{\left|\mathbf{Q}\right|^{1/2}}{\left|\mathbf{G}_i'\mathbf{Q}\mathbf{G}_i\right|^{1/2}} \tag{3.44}$$

and the "inverse Jacobian matrix" will become

$$\mathbf{M}_i = (\mathbf{G}_i'\mathbf{Q}\mathbf{G}_i)^{-1}\mathbf{G}_i'\mathbf{Q}. \tag{3.45}$$

This latter form is analogous to the relevant Jacobian matrix for GMM estimation, with $\mathbf{Q}$ representing the usual "limiting weight matrix" for the GMM criterion and $\mathbf{G}_i$ serving the role of the expected derivatives of the moment functions with respect to the parameters.

The "regression residual" term $\tilde{\mathbf{X}}_i$ in (3.42) is more familiar from the literature on single-index regression models, though its definition is complicated by the presence of the weight matrix $\mathbf{A}_{ij}$ and its limiting form $\tilde{\mathbf{A}}_i$. In the special case in which $\mathbf{A}_{ij}$ only depends on $\mathbf{X}_i$ and $\mathbf{X}_j$ through the reduced form parameters $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ (or, equivalently, through the indices $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$), the matrix $\tilde{\mathbf{X}}_i$ reduces to

$$\tilde{\mathbf{X}}_i = \mathbf{X}_i - E\left[\mathbf{X}_i \mid \mathbf{X}_i\boldsymbol{\theta}_0\right],$$

a familiar object in the large-sample theory for estimators of single-index or semiparametric regression models.

The representation (3.39) facilitates the derivation of an asymptotic distribution theory for the estimator $\hat{\boldsymbol{\theta}} = (1, \hat{\boldsymbol{\beta}}')'$ of the index coefficients. Under the assumptions imposed below, the terms in this normalized average have zero mean and finite variance, so the Lindeberg-Levy central limit theorem implies that $\hat{\mathbf{S}}\,\boldsymbol{\theta}_0$ is asymptotically normal. However, this asymptotic distribution will be singular: since

$$\begin{aligned} \boldsymbol{\theta}_0'\tilde{\mathbf{X}}_i' &= \boldsymbol{\theta}_0'\mathbf{X}_{\cdot i} - \tilde{\mathbf{A}}_i^{-1} \lim_{N\to\infty} E\left[\mathbf{A}_{ij}\boldsymbol{\theta}_0'\mathbf{X}_j' \mid \mathbf{X}_i, \mathbf{X}_j\boldsymbol{\theta}_0 = \mathbf{X}_i\boldsymbol{\theta}_0\right] && (3.46) \\ &= \mathbf{0}, && (3.47) \end{aligned}$$

it follows that

$$\sqrt{N}\boldsymbol{\theta}_0'\hat{\mathbf{S}}\,\boldsymbol{\theta}_0 = o_p(1). \tag{3.48}$$

This further implies that the smallest (in magnitude) eigenvalue $\hat{\nu}$ converges in probability to zero faster than the square root of the sample size, because

$$\sqrt{N}\,|\hat{\nu}| = \sqrt{N}\,\min_{\alpha\neq 0} |\boldsymbol{\alpha}'\hat{\mathbf{S}}\,\boldsymbol{\alpha}|/\|\boldsymbol{\alpha}\|^2 \leq \sqrt{N}\,|\boldsymbol{\theta}_0'\hat{\mathbf{S}}\,\boldsymbol{\theta}_0|/\|\boldsymbol{\theta}_0\|^2 = o_p(1). \tag{3.49}$$

To derive the asymptotic distribution of the proposed estimator $\hat{\boldsymbol{\beta}}$ in (2.29), the normalized difference of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ can be decomposed as

$$\begin{aligned} \sqrt{N}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,[\hat{\mathbf{S}}_{12} - (\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I})\,\boldsymbol{\theta}_0] \\ &= [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,\hat{\mathbf{s}} - \sqrt{n}\,\hat{\nu}\,[\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\boldsymbol{\beta}_0, && (3.50) \end{aligned}$$

where

$$\hat{\mathbf{s}} \equiv \hat{\mathbf{S}}_{12} - \hat{\mathbf{S}}_{22}\boldsymbol{\beta}_0 \equiv [\hat{\mathbf{S}}\,\boldsymbol{\theta}_0]_2, \tag{3.51}$$

i.e., $\hat{\mathbf{s}}$ is the subvector of $\hat{\mathbf{S}}\boldsymbol{\theta}_0$ corresponding to the free coefficients $\boldsymbol{\beta}_0$. Using conditions (3.49), the same arguments as in Ahn and Powell (1993) yield

$$\hat{\mathbf{S}}_{22} - \hat{\nu}\,\mathbf{I} \to^p \boldsymbol{\Sigma}_{22}, \tag{3.52}$$

where $\boldsymbol{\Sigma}_{22}$ is the lower $p \times p$ diagonal submatrix of $\boldsymbol{\Sigma}_0$, and also

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = [\boldsymbol{\Sigma}_{22}]^{-1}\sqrt{n}\,\hat{\mathbf{s}} + o(1), \tag{3.53}$$

from which the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ follow from the asymptotic normality of $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$. Specifically,

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow^d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{22}^{-1}), \tag{3.54}$$

where $\boldsymbol{\Omega}_{22}$ is the lower $p \times p$ diagonal submatrix of

$$\boldsymbol{\Omega} \equiv E[\boldsymbol{\psi}_i \boldsymbol{\psi}_i'], \tag{3.55}$$

where $\boldsymbol{\psi}_i$ is the influence function term in (3.39), i.e.,

$$\boldsymbol{\psi}_i \equiv 2\phi_i \tilde{\mathbf{X}}_i' \tilde{\mathbf{A}}_i \mathbf{M}_i \boldsymbol{\xi}_i, \tag{3.56}$$

A similar argument yields the asymptotic equivalence of the "closed form" estimator $\tilde{\boldsymbol{\beta}}$ of (2.30a) and the estimator $\hat{\boldsymbol{\beta}}$, since

$$\begin{aligned} \sqrt{N}\,(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) &= \sqrt{N}([\hat{\mathbf{S}}_{22} - \hat{\nu}\,\mathbf{I}]^{-1} - [\hat{\mathbf{S}}_{22}]^{-1})\hat{\mathbf{S}}_{12} \\ &= -\sqrt{N}\,\tilde{\nu}\,[\hat{\mathbf{S}}_{22} - \hat{\nu}\,\mathbf{I}]^{-1}\hat{\theta} \\ &= o_p(1), \end{aligned} \tag{3.57}$$

for $\tilde{\nu}$ an intermediate value between $\hat{\nu}$ and zero.

A final requirement for conducting the usual large-sample normal inference procedures is a consistent estimator of the asymptotic covariance matrix $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{22}^{-1}$ of $\hat{\theta}$. Estimation of $\boldsymbol{\Sigma}_{22}^{-1}$ is straightforward; by the results given above, either $[\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}$ or $\hat{\mathbf{S}}_{22}^{-1}$ will be consistent, with the former being more natural for $\hat{\theta}$ and the latter for $\tilde{\theta}$. Consistent estimation of the matrix $\boldsymbol{\Omega}$ is less straightforward; given a suitably-consistent estimator $\hat{\boldsymbol{\psi}}_i$ of the influence function term $\boldsymbol{\psi}_i$ of (3.56), a corresponding estimator of $\boldsymbol{\Omega}$ could be constructed as

$$\hat{\boldsymbol{\Omega}} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i'. \tag{3.58}$$

Estimation of $\boldsymbol{\psi}_i$ requires a corresponding estimator of the influence function term $\boldsymbol{\xi}_i$ for the estimator of the reduced-form parameters, which would depend upon the form of these parameters–for example, when $\boldsymbol{\gamma}_i = E\,[\mathbf{y}_i|\mathbf{X}_i]$, a natural estimator of $\boldsymbol{\xi}_i$ would be $\hat{\boldsymbol{\xi}}_i = \mathbf{y}_i - \hat{\boldsymbol{\gamma}}_i = \mathbf{y}_i - \hat{E}\,[\mathbf{y}_i|\mathbf{X}_i]$, Given an appropriate estimator of the first-stage error term $\hat{\boldsymbol{\xi}}_i$,a similar Taylor's series argument as in Ahn and Powell (1993) yields a candidate $\hat{\boldsymbol{\psi}}_i$ of an estimator of the second step influence function $\boldsymbol{\psi}_i$,

$$\hat{\boldsymbol{\psi}}_i \equiv \frac{2}{n-1}\sum_{i=1}^{n} \frac{1}{h^{q+1}}\,\mathbf{D}\left(\frac{\hat{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_j}{h}\right)\hat{\boldsymbol{\xi}}_i \cdot (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j), \tag{3.59}$$

with $\mathbf{D}(\mathbf{v}) \equiv \partial K(\mathbf{v})/\partial \mathbf{v}'$. Verification of the consistency of $\hat{\mathbf{\Omega}}$ under the conditions imposed below would follow the same strategy as described in Section 4 of.Ahn and Powell (1993).

4. The Ideal Weight Matrix

The results of the preceding section raise the question of the optimal choice of weight matrix $\mathbf{A}_{ij}$ to be used in the construction of $\hat{\mathbf{S}}$ in (2.23) above. The usual efficiency arguments suggest that the optimal choice would yield equality (more precisely, proportionality) of the matrices $\mathbf{\Sigma}_{22}$ and $\mathbf{\Omega}_{22}$ characterizing the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. The matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Omega}_0$ can be expressed as

$$\begin{aligned} \mathbf{\Sigma}_0 &= \lim_{N\to\infty} E[\phi_i(\mathbf{X}_i - \mathbf{X}_j)\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j) \,|\, \boldsymbol{\mu}_i = \boldsymbol{\mu}_j] && (4.1) \\ &= 2E\left[\phi_i \tilde{\mathbf{X}}_i' \tilde{\mathbf{A}}_i \tilde{\mathbf{X}}_i\right] && (4.2) \end{aligned}$$

and

$$\mathbf{\Omega}_0 = 4E\left[\phi_i^2 \tilde{\mathbf{X}}_i' \tilde{\mathbf{A}}_i \mathbf{M}_i \mathbf{V}_i \mathbf{M}_i' \tilde{\mathbf{A}}_i \tilde{\mathbf{X}}_i\right], \tag{4.3}$$

so the Gauss-Markov heuristic would suggest that the weight matrix $\mathbf{A}_{ij}^*$ would be efficient if the corresponding limiting matrix $\tilde{\mathbf{A}}_i^*$ equates $2\mathbf{\Sigma}_0$ and $\mathbf{\Omega}_0$, i.e., if

$$\tilde{\mathbf{A}}_i^* = \left(\phi_i \mathbf{M}_i \mathbf{V}_i \mathbf{M}_i'\right)^{-1}, \tag{4.4}$$

with corresponding matrices

$$2\mathbf{\Sigma}_0^* = \mathbf{\Omega}_0^* = 4E\left[\tilde{\mathbf{X}}_i' \left(\mathbf{M}_i \mathbf{V}_i \mathbf{M}_i'\right)^{-1} \tilde{\mathbf{X}}_i\right]. \tag{4.5}$$

Unfortunately, construction of a weight matrix $\mathbf{A}_N^*(\mathbf{X}_i, \mathbf{X}_j)$ yielding $\tilde{\mathbf{A}}_i^*$ of (4.4) through the relation in (3.41) is apparently infeasible in the general setting in which $\mathbf{V}_i = \mathbf{Var}[\boldsymbol{\xi}_i|\mathbf{X}_i]$ is a general function of the regressor matrix $\mathbf{X}_i$. However, in the special case that $\mathbf{V}_i$ is restricted to depend on $\mathbf{X}_i$ only through $\boldsymbol{\gamma}(\mathbf{X}_i)$ (or, equivalently, through the index vector $\mathbf{X}_i\boldsymbol{\theta}_0$), there is a matrix $\mathbf{A}_{ij}^*$ yielding the optimal limiting weight matrix $\tilde{\mathbf{A}}_i^*$ of (4.4); that is, if

$$\begin{aligned} \mathbf{V}_i &\equiv \mathbf{Var}[\boldsymbol{\xi}_i|\mathbf{X}_i] \\ &= \mathbf{Var}[\boldsymbol{\xi}_i|\mathbf{X}_i\boldsymbol{\theta}_0], && (4.6) \end{aligned}$$

then a sequence of matrices $\mathbf{A}_N^*(\mathbf{X}_i, \mathbf{X}_j)$ for which

$$\mathbf{A}_{ij}^* \equiv \lim_{N\to\infty} \mathbf{A}_N^*(\mathbf{X}_i, \mathbf{X}_j) = \left(\frac{\phi_i \mathbf{M}_i \mathbf{V}_i \mathbf{M}_i' + \phi_j \mathbf{M}_j \mathbf{V}_j \mathbf{M}_j'}{2} \right)^{-1} \tag{4.7}$$

would generate the limiting weight matrix in (4.4) and the corresponding equality of $2\boldsymbol{\Sigma}_0^*$ and $\boldsymbol{\Omega}_0^*$ in (4.5). The corresponding estimator $\hat{\boldsymbol{\beta}}^*$ of the unnormalized slope coefficients $\boldsymbol{\beta}$ would have the asymptotic normal distribution

$$\begin{aligned} \sqrt{N}\,(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) &\rightarrow {}^d\mathcal{N}(\mathbf{0}, \frac{1}{4}\,[\boldsymbol{\Omega}_{22}^*]^{-1}) \\ &= \mathcal{N}(\mathbf{0}, \left(E\left[\tilde{\mathbf{X}}_i' \left(\mathbf{M}_i \mathbf{V}_i \mathbf{M}_i' \right)^{-1} \tilde{\mathbf{X}}_i \right]_{22} \right)^{-1}), \end{aligned} \tag{4.8}$$

where now $\tilde{\mathbf{X}}_i = \mathbf{X}_i - E\left[\mathbf{X}_i \mid \mathbf{X}_i \boldsymbol{\theta}_0\right]$ due to the restricted form of $\mathbf{A}_N^*(\mathbf{X}_i, \mathbf{X}_j)$. In the just-identified setting with $r = q$, the matrices $\mathbf{M}_i$ and $\mathbf{V}_i$ are square, with $\mathbf{M}_i = \mathbf{G}_i^{-1}$, so the asymptotic distribution of $\hat{\boldsymbol{\beta}}^*$ would simplify to

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) \rightarrow^d \mathcal{N}(\mathbf{0}, \left(E\left[\tilde{\mathbf{X}}_i' \mathbf{G}_i' \mathbf{V}_i^{-1} \mathbf{G}_i \tilde{\mathbf{X}}_i \right]_{22} \right)^{-1}). \tag{4.9}$$

Even when condition (4.6) is imposed, the Gauss-Markov motivation for (4.4) would only establish efficiency of the estimator $\hat{\boldsymbol{\beta}}^*$ within the class of estimators defined by (2.23), (2.29), and (2.30a), not its efficiency among all regular estimators of $\boldsymbol{\beta}_0$ under the restrictions imposed here. Still, for a special case, a stronger semiparametric efficiency result holds. The binary response model

$$y_i = 1\{\mathbf{X}_i \boldsymbol{\theta}_0 + \varepsilon_i > 0\}$$

with independence of the error term ε_i and the (row vector of) regressors $\mathbf{X}_i$ will satisfy the index and invertibility restrictions imposed here, with $r = q = 1$ and $\gamma_i = E[y_i|\mathbf{X}_i] = g(\mathbf{X}_i\boldsymbol{\theta}_0)$, where g is the c.d.f. of -ε_i (assumed invertible and smooth), $\mathbf{G}_i = dg(\mathbf{X}_i\boldsymbol{\theta}_0)/d(\mathbf{X}_i\boldsymbol{\theta}_0) = \mathbf{G}_i'$, and $\mathbf{V}_i = \gamma_i(1-\gamma_i)$. For this model, Klein and Spady (1993) propose an estimator of $\boldsymbol{\theta}_0$ (up to scale) and shows that it achieves the relevant semiparametric efficiency bound, and with our normalization that efficient estimator has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}^*$ in (4.9), implying that $\hat{\boldsymbol{\beta}}^*$ is semiparametrically efficient for binary response under independence of the errors and regressors. For the just-determined multiple index model (2.6), Lee (1995) shows that $4\boldsymbol{\Omega}_{22}^*$ is the semiparametric analogue of the information matrix for estimation of $\boldsymbol{\beta}_0$ when only the index restrictions (2.6) (and smoothness of $\mathbf{g}(\cdot)$ are imposed, so $\hat{\boldsymbol{\beta}}^*$ would have the same asymptotic distribution as an efficient semiparametric estimator under these

index restrictions. However, the argument for consistency of $\hat{\boldsymbol{\beta}}^*$ exploits the additional restriction of invertibility of the conditional expections $\mathbf{g}(\mathbf{X}_i\boldsymbol{\theta}_0)$ in the vector of indices $\mathbf{X}_i\boldsymbol{\theta}_0$, which is not required for consistency of Lee's (1995) semiparametric maximum likelihood estimator. Indeed, Thompson (1993) shows that, for the multinomial response model (2.4)-(2.5), the semiparametric information matrix under the assumption of independence of the errors and regressors generally exceeds $4\boldsymbol{\Omega}_{22}^*$ when $r > 1$, implying that the semiparametric efficiency of $\hat{\boldsymbol{\beta}}^*$ only holds for $r = 1$ (binary response).

As noted earlier, there is a close connection between the estimation approach proposed here and the "minimum chi-squared logit" estimator proposed for the binary logit by Berkson (1955), and extended to general parametric multinomial choice models by Amemiya (1976). For that estimation approach, it is assumed that the matrix of regressors $\mathbf{X}_i$ is constant within each of a fixed number of groups, and a nonparametric estimator of the vector of choice probabilities $\boldsymbol{\gamma}_i$ for each group is constructed from the observed choice frequencies for each group. With this setup, the relavant asymptotic covariance matrix $\mathbf{V}_i$ takes the form

$$\mathbf{V}_i = diag\{\boldsymbol{\gamma}_i\} - \boldsymbol{\gamma}\boldsymbol{\gamma}', \tag{4.10}$$

and information matrix

$$I = E\left[\mathbf{X}_i'\mathbf{G}_i'\mathbf{V}_i^{-1}\mathbf{G}_i\mathbf{X}_i\right] \tag{4.11}$$

Amemiya (1976) shows that an efficient estimator of $\boldsymbol{\theta}_0$ for this setup is the coefficient vector of the generalized least squares regression of $\mathbf{g}^{-1}(\hat{\boldsymbol{\gamma}}_i)$ on $\mathbf{X}_i$, using the matrix $\mathbf{G}_i'\mathbf{V}_i^{-1}\mathbf{G}_i$ (or a feasible version that replaces the unknown probabilities by the group frequencies) as the weighting matrix. The estimator $\hat{\boldsymbol{\theta}}^*$ is a semiparametric analogue of this minimum chi-squared estimator, with asymptotic variance similar to (4.11) but with the regressors $\mathbf{X}_i$ replaced by the residual regressors $\tilde{\mathbf{X}}_i = \mathbf{X}_i - E[\mathbf{X}_i|\mathbf{X}_i\boldsymbol{\theta}_0]$.

When the index restrictiion (4.6) on the reduced-form error variance holds and the model is overdetermined ($r > q$), there is another avenue to reduce the asymptotic variance of $\hat{\boldsymbol{\beta}}$, namely, choice of kernel function $K(\cdot)$, which affects the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ through the inverse Jacobian matrix term $\mathbf{M}_i$. When this kernel function takes the elliptically-symmetric form (3.43) and the weight matrix is the infeasible $\mathbf{A}_{ij}^*$ of (4.4), the asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ is the inverse of the lower $p \times p$

diagonal submatrix of

$$\begin{aligned}\frac{1}{4}\mathbf{\Omega}^* &= E\left[\tilde{\mathbf{X}}_i'\left(\mathbf{M}_i\mathbf{V}_i\mathbf{M}_i'\right)^{-1}\tilde{\mathbf{X}}_i\right] \\ &= E\left[\tilde{\mathbf{X}}_i'(\mathbf{G}_i'\mathbf{Q}\mathbf{G}_i)\left(\mathbf{G}_i'\mathbf{Q}\mathbf{V}_i\mathbf{Q}\mathbf{G}_i\right)^{-1}(\mathbf{G}_i'\mathbf{Q}\mathbf{G}_i)\tilde{\mathbf{X}}_i\right], \end{aligned} \tag{4.12}$$

using the expression for $\mathbf{M}_i$ in (**??**). If $\mathbf{V}_i$ were constant ($\mathbf{V}_i \equiv \mathbf{V}_0$), the usual Gauss-Markov heuristic would choose $\mathbf{Q}$ to maximize this matrix by equating $\mathbf{G}_i'\mathbf{Q}\mathbf{G}_i$ and $\mathbf{G}_i'\mathbf{Q}\mathbf{V}\mathbf{Q}\mathbf{G}_i$, i.e., by setting $\mathbf{Q} = \mathbf{Q}^* = \mathbf{V}^{-1}$, so that the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}^*$ would reduce to $(1/4)\mathbf{\Omega}^* = E\left[\tilde{\mathbf{X}}_i'\mathbf{G}_i'\mathbf{V}_0^{-1}\mathbf{G}_i\tilde{\mathbf{X}}_i\right]$. When the covariance matrix $\mathbf{V}_i$ varies with $\mathbf{X}_i$ (but only as a function of $\boldsymbol{\gamma}_i = g(\mathbf{X}_i\boldsymbol{\theta}_0)$), this reduction would require a modification of the estimator $\hat{\mathbf{S}}$ in (2.23), replacing the the fixed kernel $K(\cdot)$ with a varying (elliptically symmetric) kernel $K_{ij}^*(\cdot)$ of the form

$$K_{ij}^*(\mathbf{v}) = c_r k(\mathbf{v}'\mathbf{Q}_{ij}^*\mathbf{v})\left|\mathbf{Q}_{ij}^*\right|^{1/2}, \tag{4.13}$$

with

$$\mathbf{Q}_{ij}^* \equiv \left(\frac{\mathbf{V}_i + \mathbf{V}_j}{2}\right)^{-1}. \tag{4.14}$$

A modification of the arguments for the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$ would yield the same results for the improved estimator $\hat{\boldsymbol{\beta}}^*$ using the weight matrix $\mathbf{A}_{ij}^*$ in (4.7) and the varying kernel $K_{ij}^*(\cdot)$ in (4.13), yielding the same form for the asymptotic distribution

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) \to^d \mathcal{N}(\mathbf{0}, \left(E\left[\tilde{\mathbf{X}}_i'\mathbf{G}_i'\mathbf{V}_i^{-1}\mathbf{G}_i\tilde{\mathbf{X}}_i\right]_{22}\right)^{-1}). \tag{4.15}$$

whether or not the model is just- or overdetermined.

Of course, even when condition (4.6) is imposed, the efficient estimator $\hat{\boldsymbol{\theta}}^*$ using the ideal weight matrix in (4.7) (and optimal kernel in (4.13)) would be infeasible in two respects – the limiting matrix $\mathbf{A}_{ij}^*$ does not specify the trimming mechanism needed to achieve the uniform nonparametric rate of convergence of $\hat{\boldsymbol{\gamma}}_i$ to $\boldsymbol{\gamma}_i$ specified in (**??**), and the construction of $\hat{\boldsymbol{\theta}}^*$ involves unknown nuisance parameters via the joint density function $f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)$ of the indices, the Jacobian matrix $\mathbf{G}_i = \partial\mathbf{g}(\boldsymbol{\mu}_i)/\partial\boldsymbol{\mu}'$, and the reduced-form covariance matrix $\mathbf{V}_i$. Following Ichimura (2004), the trimming requirement can be imposed by multiplying the limiting $\mathbf{A}_{ij}^*$ by a trimming term t_{ij} of trimming terms of the form

$$t_{ij} \equiv 1\{f_{\mathbf{X}}(\mathbf{X}_i) > b_N\} \cdot 1\{f_{\mathbf{X}}(\mathbf{X}_i) > b_N\}, \tag{4.16}$$

where $f_{\mathbf{X}}(\mathbf{X}_i)$ is the joint density function of the regressors and b_N is a bandwidth term declining to zero at an appropriate rate. Ichimura (2004) discusses conditions under which this trimming mechanism

will satisfy the requirements in (3.32) and also under which the asymptotic distribution $\hat{\mathbf{S}}$ will be unaffected by replacement of the density $f_{\mathbf{X}}(\mathbf{X}_i)$ in (4.16) by a consistent nonparametric (kernel) estimator $\hat{f}_{\mathbf{X}}(\mathbf{X}_i)$. As noted above, construction of a consistent estimator $\hat{\mathbf{V}}_i$ of $\mathbf{V}_i$ will depend upon the form of the reduced-form estimators $\hat{\boldsymbol{\gamma}}_i$; the remaining nuisance parameters $f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)$ and $\mathbf{G}_i$ could also be consistently estimated using a kernel density estimator applied to the estimated indices $\hat{\boldsymbol{\mu}}_i = \mathbf{X}_i\hat{\boldsymbol{\theta}}$ and the derivative of a kernel regression of $\hat{\boldsymbol{\gamma}}_i$ on $\hat{\boldsymbol{\mu}}_i$, respectively. Demonstration that an estimator using a feasible version $\hat{\mathbf{A}}_{ij}$ of $\mathbf{A}^*_{ij}$ would achieve the same asymptotic distribution (4.15) of the ideal estimator $\hat{\boldsymbol{\beta}}^*$ would involve strengthening of the regularity conditions and similar derivations to those in the Appendix.

5. Appendix

The derivations of the consistency of $\hat{\mathbf{S}}$ for $\boldsymbol{\Sigma}_0$ in (3.34) and the asymptotic linearity representation (3.39) for $\hat{\mathbf{S}}\boldsymbol{\theta}_0$ follow the same lines as in Ahn and Powell (1993) and Blundell and Powell (2004). Regarding $\hat{\mathbf{S}}$ as a function of the reduced-form estimators $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\boldsymbol{\gamma}}_j$, a second-order Taylor's series expansion around the true reduced-form parameters $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ yields

$$\hat{\mathbf{S}} = \mathbf{S}_0 + \mathbf{S}_1 + \mathbf{S}_2, \tag{5.1}$$

where

$$\mathbf{S}_0 \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{h^q} K\left(\frac{\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j}{h}\right) \cdot (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j), \tag{5.2}$$

$$\mathbf{S}_1 \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{h^{q+1}} \mathbf{D}\left(\frac{\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j}{h}\right) (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_j + \boldsymbol{\gamma}_j) \cdot (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j),$$

and

$$\mathbf{S}_2 \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{h^{q+2}} (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_j + \boldsymbol{\gamma}_j)'\mathbf{H}\left(\frac{\boldsymbol{\gamma}^*_i - \boldsymbol{\gamma}^*_j}{h}\right) (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_j + \boldsymbol{\gamma}_j) \cdot (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j),$$

where $\mathbf{D}(\mathbf{v}) \equiv \partial K(\mathbf{v})/\partial \mathbf{v}$ and $\mathbf{H}(\mathbf{v}) \equiv \partial^2 K(\mathbf{v})/\partial \mathbf{v}\partial \mathbf{v}'$, and where $\boldsymbol{\gamma}^*_i$ and $\boldsymbol{\gamma}^*_j$ are intermediate values.

The first term $\mathbf{S}_0$ is a second-order U-statistic with a summand depending on the sample size N through the bandwidth term h. We assume that the terms in $(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)$ have bounded second moments and have conditional expectations that are smooth functions of the underlying indices

$\mathbf{X}_i\boldsymbol{\theta}_0$ and $\mathbf{X}_j\boldsymbol{\theta}_0$, and also the kernel function $K(\cdot)$ satisfies some common regularity conditions (it integrates to one, is bounded and continuous with bounded derivatives). Under such conditions, the second moment of the terms in the summand for $\mathbf{S}_0$ will be of order h^{-q}, so by Lemma 3.1 of Powell, Stock, and Stoker (1989), the term $\mathbf{S}_0$ will satisfy a weak law of large numbers,

$$\mathbf{S}_0 - E\left[\mathbf{S}_0\right] \xrightarrow{p} \mathbf{0}, \tag{5.3}$$

provided

$$h = h_N = o(1), \qquad h^{-q} = o(N). \tag{5.4}$$

Calculating the expectation of the summand in $\mathbf{S}_0$, first conditioning on $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\theta}_0$ and $\boldsymbol{\mu}_j$ and then making the change of variables $\mathbf{u} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/h$, yields

$$\begin{aligned}
E\left[\mathbf{S}_0\right] &= E\left(h^{-q}K\left(\frac{\mathbf{g}(\boldsymbol{\mu}_i) - \mathbf{g}(\boldsymbol{\mu}_j)}{h}\right) E\left[(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\,\middle|\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\right]\right) \\
&= E\left(\int K\left(\frac{\mathbf{g}(\boldsymbol{\mu}_i) - \mathbf{g}(\boldsymbol{\mu}_i - h\mathbf{u})}{h}\right) E\left[(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\,\middle|\boldsymbol{\mu}_i, \boldsymbol{\mu}_j = \boldsymbol{\mu}_i - h\mathbf{u}\right]\right. \\
&\qquad \left. f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i - h\mathbf{u})d\mathbf{u}\right) \\
&\to E\left(\int K\left(\mathbf{G}_i\mathbf{u}\right)d\mathbf{u}\cdot E\left[(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\,\middle|\boldsymbol{\mu}_i, \boldsymbol{\mu}_j = \boldsymbol{\mu}_i\right] f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)\right) \qquad (5.5) \\
&= E\left(\delta_i f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\,\middle|\boldsymbol{\mu}_i, \boldsymbol{\mu}_j = \boldsymbol{\mu}_i\right) \\
&\equiv \boldsymbol{\Sigma}_0.
\end{aligned}$$

Under the rate condition (3.32) for the uniform convergence of the reduced form estimators, the remaining terms $\mathbf{S}_1$ and $\mathbf{S}_2$ in the decomposition (5.1) satisfy

$$\begin{aligned}
\|\mathbf{S}_1\| &\leq \frac{2C}{h^{q+1}}\left[\binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left\|(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\right\|\right] \\
&\qquad \cdot\left[\max_{i,j} 1\{||\mathbf{A}_N(\mathbf{X}_i, \mathbf{X}_j)|| \neq \mathbf{0}\}\cdot||\hat{\boldsymbol{\gamma}}(\mathbf{X}_i) - \boldsymbol{\gamma}(\mathbf{X}_i)||\right] \\
&= o_p\left(\frac{1}{N^{3/8}h^{q+1}}\right), \qquad (5.6) \\
\|\mathbf{S}_2\| &\leq \frac{2C}{h^{q+2}}\left[\binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left\|(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i - \mathbf{X}_j)\right\|\right] \\
&\qquad \cdot\left[\max_{i,j} 1\{||\mathbf{A}_N(\mathbf{X}_i, \mathbf{X}_j)|| \neq \mathbf{0}\}\cdot||\hat{\boldsymbol{\gamma}}(\mathbf{X}_i) - \boldsymbol{\gamma}(\mathbf{X}_i)||^2\right] \\
&= o_p\left(\frac{1}{N^{3/4}h^{q+2}}\right) \qquad (5.7)
\end{aligned}$$

(where C is an upper bound for the kernel K and its first two derivatives). As long as the bandwidth h converges to zero sufficently slowly so that

$$\frac{1}{h} = o\left(N^{1/4(q+2)}\right), \tag{5.8}$$

it will follow that

$$\|\mathbf{S}_1\| = o_p(1) \tag{5.9}$$

and

$$\|\mathbf{S}_2\| = o_p(N^{-1/2}), \tag{5.10}$$

implying

$$\hat{\mathbf{S}} \xrightarrow{p} \boldsymbol{\Sigma}_0. \tag{5.11}$$

With the identification requirement that the lower $p \times p$ submatrix $\boldsymbol{\Sigma}_{22}$ has full rank, consistency of $\hat{\boldsymbol{\beta}}$ follows from $\boldsymbol{\Sigma}_0\boldsymbol{\theta}_0 = 0$.

To obtain the asymptotic linearity condition (3.39), we use the same decomposition (5.1), yielding

$$\begin{aligned} \sqrt{N}\hat{\mathbf{S}}\boldsymbol{\theta}_0 &= \sqrt{\mathbf{N}}\mathbf{S}_0\boldsymbol{\theta}_0 + \sqrt{N}\mathbf{S}_1\boldsymbol{\theta}_0 + \sqrt{N}\mathbf{S}_2\boldsymbol{\theta}_0 & (5.12)\\ &= \sqrt{\mathbf{N}}\mathbf{S}_0\boldsymbol{\theta}_0 + \sqrt{N}\mathbf{S}_1\boldsymbol{\theta}_0 + o_p(1) & (5.13) \end{aligned}$$

by (5.10). The first term is a second-order smoothed U-statistic, so under condition (5.8), which implies (5.4), Lemma 3.1 of Powell, Stock,and Stoker yield

$$\sqrt{N}\mathbf{S}_0\boldsymbol{\theta}_0 = \sqrt{N}E[\mathbf{S}_0\boldsymbol{\theta}_0] + \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\left[\boldsymbol{\rho}_{iN} - E\left[\boldsymbol{\rho}_{iN}\right]\right] + o_p(1), \tag{5.14}$$

where

$$\begin{aligned} \boldsymbol{\rho}_{iN} &\equiv E\left[h^{-q}K\left(\frac{\mathbf{g}(\boldsymbol{\mu}_i) - \mathbf{g}(\boldsymbol{\mu}_j)}{h}\right)\cdot(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \mid \mathbf{X}_j\right] \\ &= \int K\left(\frac{\mathbf{g}(\boldsymbol{\mu}_i) - \mathbf{g}(\boldsymbol{\mu}_i - h\mathbf{u})}{h}\right) E\left[((\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}\left(h\mathbf{u}\right)\left|\mathbf{X}_i, \boldsymbol{\mu}_j = \boldsymbol{\mu}_i - h\mathbf{u}\right.\right] \\ &\qquad f_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i - h\mathbf{u})d\mathbf{u}, \end{aligned} \tag{5.15}$$

again using the change-of-variables $\mathbf{u} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/h$ for $\boldsymbol{\mu}_j$. Since the second moment of $\boldsymbol{\rho}_{iN}$ is of order h^2, it follows that

$$\sqrt{N}\mathbf{S}_0\boldsymbol{\theta}_0 = \sqrt{N}E[\mathbf{S}_0\boldsymbol{\theta}_0] + o_p(1), \tag{5.16}$$

that is, that the only contribution of the leading term on the right-hand side of (5.12) is a bias term $\sqrt{N}E[\mathbf{S}_0\boldsymbol{\theta}_0]$. Under the assumption that the expectation $E[\mathbf{S}_0\boldsymbol{\theta}_0] = E[\mathbf{S}_0\boldsymbol{\theta}_0](h)$ is a sufficiently smooth function of the bandwidth parameter h, it will admit a power-series expansion of the form

$$E[\mathbf{S}_0\boldsymbol{\theta}_0](h) = \sum_{j=1}^{L} \boldsymbol{\Gamma}_j h^j + o\left(\frac{1}{\sqrt{N}}\right) \tag{5.17}$$

as long as L is sufficiently large so that $h^{L+1} = o(N^{-1/2})$ while (5.8) is satisfied. Assuming the expansion (5.17) is available, a kernel function $K(\cdot)$ can be constructed to ensure that the coefficient matrices $\boldsymbol{\Gamma}_j$ equal zero for $j = 1, ..., L$ using the jackknife approach described in Honoré and Powell (2004). (That approach would replace the matrix $\hat{\mathbf{S}} = \hat{\mathbf{S}}(h)$ by a particular linear combination of matrices $\{\hat{\mathbf{S}}(c_j h)\}$ for distinct positive constants $c_1, ..., c_L$, which is algebraically equivalent to use of a "higher-order kernel" constructed using the generalized jackknife approach of Schucany and Sommers (1977).) With this choice of kernel, the bias term $E[\mathbf{S}_0\boldsymbol{\theta}_0] = o(N^{-1/2})$, which, together with (5.12) and (5.16), yields

$$\sqrt{N}\hat{\mathbf{S}}\boldsymbol{\theta}_0 = \sqrt{N}\mathbf{S}_1\boldsymbol{\theta}_0 + o_p(1). \tag{5.18}$$

To verify the asymptotic linearity expression (3.39) for the remaining term $\sqrt{N}\mathbf{S}_1\boldsymbol{\theta}_0$, we insert the corresponding asymptotic linearity expression (2.21) into the expression for $\mathbf{S}_1$, obtaining

$$\begin{aligned}
\sqrt{N}\mathbf{S}_1\boldsymbol{\theta}_0 \quad \equiv \quad & \frac{1}{\sqrt{N}}\begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{l=1}^{N} \frac{1}{h^{q+1}}\, \mathbf{D}\left(\frac{\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j}{h}\right)(\mathbf{R}_{lN}(\mathbf{X}_i)\boldsymbol{\xi}_j - \mathbf{R}_{lN}(\mathbf{X}_i)\boldsymbol{\xi}_j) \\
& \qquad \cdot(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
& + \sqrt{N}\begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{h^{q+1}}\, \mathbf{D}\left(\frac{\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j}{h}\right)(\mathbf{r}_{lN} - \mathbf{r}_{lN}) \\
& \qquad \cdot(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
\equiv \quad & \mathbf{T}_1 + \mathbf{T}_2.
\end{aligned} \tag{5.19}$$

The second term is asymptotically negligible under condition (3.33), since

$$\begin{aligned}
\|\mathbf{T}_2\| \quad \leq \quad & \frac{C\sqrt{N}}{h^{q+1}}\begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \left\|(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{A}_{ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right\| \cdot \max_{i,j} t_{ij}[||\mathbf{r}_{iN}|| + ||\mathbf{r}_{iN}||] \\
= \quad & O_p\left(N^{1/2}h^{-(q+1)}\right) \cdot o_p\left(N^{-3/4}\right) \\
= \quad & o_p\left(N^{-1/4}h^{-(q+1)}\right) \\
= \quad & o_p(1),
\end{aligned} \tag{5.20}$$

where the last line follows from condition (5.8) above (with C an upper bound for $\|\mathbf{D}(\cdot)\|$).

Finally, apart from negligible terms with $i = l$ or $j = l$, the term $\mathbf{T}_1$ can be rewritten as a third-order smoothed U-statistic, i.e., it is of the form

$$\mathbf{T}_1 = \sqrt{N} \left(\begin{array}{c} N \\ 3 \end{array} \right)^{-1} \sum_{i<j<l} \boldsymbol{\kappa}_{ijlN} + o_p(1), \tag{5.21}$$

where

$$\boldsymbol{\kappa}_{ijlN} \equiv \frac{1}{h^{q+1}} \mathbf{D}\left(\frac{\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j}{h}\right) (\mathbf{R}_{lN}(\mathbf{X}_i)\boldsymbol{\xi}_j - \mathbf{R}_{lN}(\mathbf{X}_i)\boldsymbol{\xi}_j) \cdot (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A}_{ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{5.22}$$

Though the kernel $\boldsymbol{\kappa}_{ijlN}$ is not symmetric in the indices i, j, and l, it can be symmetrized in the same way as in the proof of Theorem 5.1 of Ahn and Powell (1993), by rewriting $\mathbf{T}_1$ as

$$\mathbf{T}_1 = \sqrt{N} \left(\begin{array}{c} N \\ 3 \end{array} \right)^{-1} \sum_{i<j<l} \tilde{\boldsymbol{\kappa}}_{ijlN} + o_p(1), \tag{5.23}$$

where

$$\tilde{\boldsymbol{\kappa}}_{ijlN} \equiv \frac{1}{6} \left(\boldsymbol{\kappa}_{ijlN} + \boldsymbol{\kappa}_{iljN} + \boldsymbol{\kappa}_{jilN} + \boldsymbol{\kappa}_{jliN} + \boldsymbol{\kappa}_{lijN} + \boldsymbol{\kappa}_{ljiN} \right). \tag{5.24}$$

A crude bound for the second moment of $\tilde{\boldsymbol{\kappa}}_{ijlN}$ is

$$\begin{aligned} E[\|\tilde{\boldsymbol{\kappa}}_{ijl}\|^2] &= O\left(E[\|\mathbf{R}_{jN}(\mathbf{X}_i)\|^2]\right) \cdot O\left(h^{-2(q+1)}\right) \\ &= O(N^{1/2} h^{-2(q+1)}), \end{aligned}$$

under assumption (2.22), so this second moment will grow slower than the sample size N under (5.8), which implies

$$h^{-2(q+1)} = o(\sqrt{N}).$$

Thus the projection lemma for smoothed U-statistics (Lemma A.3 of Ahn and Powell 1993) implies that the third-order U-statistic component of $\mathbf{T}_1$ can be replaced by its projection, i.e.,

$$\mathbf{T}_1 = \frac{6}{\sqrt{N}} \sum_{i=1}^{N} E[\tilde{\boldsymbol{\kappa}}_{ijlN} | \boldsymbol{\xi}_i, \mathbf{X}_i] + o_p(1).$$

Calculations analogous to those on pages 26 through 28 of Ahn and Powell (1993) establish that

$$E\left[\left\| E[\tilde{\boldsymbol{\kappa}}_{ijlN} | \boldsymbol{\xi}_i, \mathbf{X}_i] - \frac{1}{6}\boldsymbol{\psi}_i \right\|\right] = o(N^{-1/2}),$$

where $\boldsymbol{\psi}_i$ is the influence function term defined in (3.56), establishing the asymptotic linearity relation (3.39).

6. References

Ahn, H., H. Ichimura, and J.L. Powell, 1996, "Simple Estimators for Monotone Index Models," manuscript, Department of Economics, University of California at Berkeley.

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

Amemiya, T., 1976, "The maximum likelihood, the minimum chi-square, and the nonlinear weighted least-squares estimator in the general qualitative response model," *Journal of the American Statistical Association,* 71, 347-351.

Berkson, J., 1955, "Maximum likelihood and minimum χ^2 estimates of the logistic function," *Journal of the American Statistical Association,* 50, 132-162.

Blundell, R.W. and J.L. Powell, 2004, "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

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

Härdle, W. and T.M. Stoker, 1989, "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association*, 84, 986-995.

Härdle, W. and J. Horowitz, 1996, "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632-1640.

Honoré, B.E. and J.L. Powell, 2005, "Pairwise Difference Estimators for Nonlinear Models," in Andrews, D.W.K. and J.H. Stock, eds., *,Identification and Inference for Econometric Models,* New York: Cambridge University Press.

Ichimura, H., 1993, "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71-120.

Ichimura, H,. 2004, "Computation of Asymptotic Distribution for Semiparametric GMM Estimators," manuscript , Department of Economics, University Colleg London.

Ichimura, H. and L-F. Lee, 1991, "Semiparametric least squares estimation of multiple index models: single equation estimation," in: W.A. Barnett, J.L. Powell, and G. Tauchen, eds., *Nonparametric and semiparametric methods in econometrics and statistics*, Cambridge: Cambridge University Press.

Klein, R.W. and R.S. Spady, 1993, "An efficient semiparametric estimator of the binary response

model," *Econometrica*, 61, 387-422.

Lee, L-F., 1995, "Semiparametric maximum likelihood estimation of polychotomous and sequential choice models", *Journal of Econometrics,* 65, 385-428.

Manski, C.F., 1975, "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205-228.

Manski, C.F., 1985, "Semiparametric Analysis of discrete response, asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 205-228.

Newey, W.K. and P.A. Ruud, 2005, "Density weighted least squares estimation," in Andrews, D.W.K. and J.H. Stock, eds., *Identification and Inference for Econometric Models,* New York: Cambridge University Press.

Newey, W.K. and T.M. Stoker, 1993 "Efficiency of weighted average derivative estimators and index models," *Econometrica,* 61, 1199-1223.

Powell, J.L., J.H. Stock and T.M. Stoker, 1989, "Semiparametric estimation of weighted average derivatives," *Econometrica* 57, 1403-1430.

Ruud, P.A., 1986, "Consistent estimation of limited dependent variable models despite misspecification of distribution," *Journal of Econometrics*, 32, 157-187.

Ruud, P.A., 2000, "Semiparametric estimation of discrete choice models," manuscript, Department of Economics, University of California at Berkeley.

Schucany, W.R. and J.P.Sommers, 1977, "Improvement of kernel type density estimators, *Journal of the American Statistical Association,* 72, 420-423.

Thompson, T.S. , 1993, "Some efficiency bounds for semiparametric discrete choice models," *Journal of Econometrics,* 58, 257–274