Estimation of a Nonlinear Panel Data Model with Predetermined Variables and Semiparametric Individual Effects*

Wayne-Roy Gayle†, Soiliou Daw Namoro‡

September 24, 2008

Abstract

In this paper, we explore identification and efficient semiparametric estimation of a class of nonlinear panel data index models with small-T, which includes a class of single-index panel discrete-choice models. The model allows for the inclusion of predetermined variables, lagged dependent variables, and a nonparametric specification of the individual-specific effects. The paper provides a root-N consistent, asymptotically normal and efficient estimator for the finite-dimensional parameters, and a consistent estimator of the unknown index function. The estimator developed in this paper may be computed with any smoother, be it sieves or kernel smoothers. We propose a powerful new kernel-based modified backfitting algorithm to compute the estimator. The algorithm fully implements the identifying restrictions of the model. We study the small sample properties of the estimator via Monte Carlo techniques. The results indicate that the estimator performs well in recovering the finite dimensional parameters of interest. The simulation results also show that, in small samples, the estimator outperforms more parametric models with various mis-specifications of the index function.
Keywords: Semiparametric estimation, dynamic panel data, nonlinear models.
JEL classification: C13, C14, C23

*The authors are grateful to Jean-Francois Richard, Mehmet Caner, Robert Miller, Holger Sieg, George-Levi Gayle and three anonymous referees for insightful comments and discussions. Comments by the participants of the 12-th Conference on Panel Data at Copenhagen, Denmark 2005 were greatly appreciated. All remaining errors are our own.

†Department of Economics, University of Virginia, 2015 Ivy Road, Room 312, Charlottesville, VA 22904-4182, E-mail: wg4b@virginia.edu

‡Economics Department, University of Pittsburgh, 230 S. Bouquet St., Pittsburgh, PA. 15260, E-mail: snamoro@pitt.edu

1 Introduction

This paper is concerned with identification and estimation of the following semiparametric regression model

$$y_{it} = \Phi_t \left(x_{it}\beta + f(z_i) \right) + \varepsilon_{it} \qquad (i = 1, \cdots, N,\ t = 1, \cdots, T), \tag{1.1}$$

where x_{it} is a K-dimensional vector of random variables that may contain lags of the dependent variable as well as other predetermined variables; z_i is an L-dimensional vector of time-constant random variables; and ε_{it} is an individual-time specific idiosyncratic shock assumed to be mean independent of the other explanatory variables. The parameters of interest are β, $\Phi := \{\Phi_t,\ t = 1,\ \cdots,\ T\}$ and f, where β is a K-dimensional vector, the Φ_t's are strictly increasing and smooth unknown functions, and f is an unknown function.

The estimator developed in this paper builds on previous work of Chamberlain (1980), Newey (1994a), Chen (1998), and Arellano and Carrasco (2003) (to name a few), concerning the estimation of binary choice panel data models with individual-specific effects. The common strategy of these papers, as well as ours, is to impose restrictions on the conditional distribution of the individual-specific effects, conditioned on the observed regressors. However, the estimator developed here differs in a variety of ways. Our own interest goes beyond the binary choice framework. Any model that can be presented in the form of equation (1.1) can be estimated using the algorithm developed in this paper. In the next section, we provide two examples of how equation (1.1) may be derived from more familiar single-index panel data models. The assumptions required on the individual specific effects will depend on the nature of the observed regressors.

The estimator proposed in this paper treats both the index functions Φ_t and the function f as unknown functions. The models proposed in Chamberlain (1980) assumes that the index function Φ_t is known, and that $f(z_i)$ is known up to a set of finite dimensional parameters. Newey (1994a) extends this framework to allow for f to be an unknown function, while maintaining the parametric specification of the index function. These models assume that the the explanatory variables are all strictly exogenous. The model presented in this paper is therefore an extension of the model presented in Newey (1994a) to allow for predetermined variables and an unspecified time specific index function. In the discrete choice

framework, Chen (1998) also extends the framework of Newey (1994a) by relaxing the parametric specification of the index function, but maintains the assumption that all of the explanatory variables are strictly exogenous. Including lagged-dependent variables into the set of regressors requires stronger assumptions on the relationship between the individual specific effects and the regressors, as discussed in the next section. Arellano and Carrasco (2003) develops a panel data discrete choice model that allows for the individual specific effect to be related to the explanatory variables in a less restrictive way than suggested in this paper. They also allow for all the explanatory variables to be predetermined. However the model presented by Arellano and Carrasco (2003) requires that the index function is known.

Semiparametric panel data models specified similar to equation (1.1) with an unknown index function can be estimated by a series or sieve minimum distance estimator (see Newey and Powell (2003), Ai and Chen (2003), and Chen (2007)). Gayle and Viauroux (2007) show that the resulting estimator of the finite dimensional parameters are $\sqrt{N}$-consistent with a Gaussian limiting distribution. In this paper, we present a general minimum distance estimator and a kernel-based algorithm to compute this estimator. The algorithm may also be implemented using sieve based smoothers. The algorithm presented here adopts the backfitting algorithm of Buja et al. (1989) Mammen et al. (1999) and Mammen et al. (2001) to the panel data context. A key extension provided by our algorithm is the estimation of additive models with monotone components, where the additive components are specified as the difference between two monotone components. We provide sufficient conditions under which the algorithm converges. We show that the resulting estimator of β is $\sqrt{N}$-consistent with a Gaussian limiting distribution. The semiparametric efficiency bound is derived and we show that the proposed estimator achieves this bound.

The paper provides two Monte Carlo exercises that confirm the convergence rate of the proposed estimator. In the first exercise, the dependent variable is continuous, all the explanatory variables are strictly exogenous, and the index function is asymmetric about zero. We show that wrongly assuming a symmetric index function such as a "stretched" normal distribution function significantly biases the estimates of the finite dimensional parameters. The second exercise simulates a dynamic probit model with unconditional heteroskedasticity. The proposed model also works well in this environment, and outperforms a model where the index function is known, but the error term is assumed to be homoskedastic.

The rest of paper is organized as follows: the following section motivates equation (1.1) by describing how it is derived from various economic models. Section 3 discusses identification while Section 4 presents the estimator. Section 5 presents the algorithm used to compute the estimate. Section 6 derives the large sample properties of the estimator and propose estimators of the asymptotic variances and average derivatives. Section 7 is devoted to the Monte carlo simulations and Section 8 concludes. All the proofs and auxiliary lemmas are to be found in the appendix of the paper.

2 The Model

In this section, we discuss two examples of how equation (1.1) is derived from more primitive models. The first example discusses relaxing the log-linearity assumption in the classical Mincer wage regression, and the second example is a dynamic panel data discrete-choice model.

EXAMPLE 1. *Semiparametric panel data Mincerian wage equation with semiparametric individual effects*. Consider the wage equation for N individuals observed over T consecutive time periods

$$\ln W_{it} = F_t(\beta_1 S_{it} + \beta_2 E_{it} + \beta_3 E_{it}^2 + x_{it}\beta_4 + \mu_i) + u_{it} \quad (i = 1, \cdots, N; \quad t = 1, \cdots, T), \quad (2.1)$$

where for individual i in period t, W_{it} is the average hourly wage rate, S_{it} is the level of completed schooling, E_{it} is the level of labor market experience, and x_{it} are other observed individual-time varying characteristics. The x_{it}'s as well as S_{it} and E_{it} may be predetermined in that they may be partially determined by lagged values of u_{it}. In this context, μ_i is interpreted is the individual's time invariant, unobserved ability. To keep things simple, assume that u_{it} has zero mean and is mean independent of all the explanatory variables.[1] The restriction of F_t, $t = 1, \cdots, T$ to the identity function results in the popular log-linear panel data wage equation (see Altug and Miller (1990) and Altug and Miller (1998)for example).

Assume that there exists a set of proxies z_i such that the individual specific effect can be

[1]This assumption abstracts away from sample selection considerations where the distribution of observed wages is potentially different from the wage offer distribution.

decomposed as $\mu_i = f(z_i) + v_i$, where v_i is independent of $((x_{it}, S_{it}, E_{it}),\ t = 0, \cdots, T)$. One alternative is to specify z_i to be the time average of the strictly exogenous explanatory variables (see Mundlak (1978), Newey and McFadden (1994), and Gayle and Viauroux (2007)). However this choice leads to a time inconsistency problem where it is not clear how to treat a new year of observation, say $T+1$ given that the model, and $f(z_i)$ in particular, is estimated with the first T cross sections. An alternative that avoids this problem is to assume that z_i is composed of time invariant measures of ability such as IQ and Armed Forces Qualification Test (AFQT) scores. Equation (2.1) can be written as

$$E[\ln W_{it} | S_{it}, E_{it}, x_i, z_i, v_i] = F_t(\beta_1 S_{it} + \beta_2 E_{it} + \beta_3 E_{it}^2 + x_{it}\beta_4 + f(z_i) + v_i). \tag{2.2}$$

Assume that the density of v_i, f_v is continuous. This density is not a function of the explanatory variables by assumption. We can therefore integrate out v_i in equation (2.2) to get

$$E[\ln W_{it} | S_{it}, E_{it}, x_i, z_i] = \Phi_t(\beta_1 S_{it} + \beta_2 E_{it} + \beta_3 E_{it}^2 + x_{it}\beta_4 + f(z_i)). \tag{2.3}$$

By defining $\varepsilon_{it} := \ln W_{it} - E[\ln W_{it} | S_{it}, E_{it}, x_i, z_i]$ we obtain equation (1.1).

EXAMPLE 2. *Dynamic panel data binary choice model with semiparametric individual effects*. For the second example, consider the model for N individuals observed over T consecutive time periods

$$y_{it} = 1\{\alpha y_{it-1} + w_{it}\gamma + \mu_i - u_{it} \geq 0\} \qquad (i = 1, \cdots, N; \quad t = 1, \cdots, T), \tag{2.4}$$

where w_{it} is a set of strictly exogenous variables. Define $w_i := (w_{i0}, \cdots, w_{iT})$. Assume that u_{it} is distributed according to the cdf Φ_t, which is not a function of (y_{it-1}, w_i, μ_i). This assumption is substantive as it rules out conditional heteroskedasticity of u_{it} conditional on (y_{it-1}, w_i, μ_i). However, it does allow for unconditional heteroskedasticity. Under homoskedasticity, Manski (1987) derives an estimator under weaker assumptions on the individual-specific effect. Honoré and Kyriazidou (2000) extends the model of Manski (1987) to include the lagged dependent variable. However, the resulting estimators are not $\sqrt{N}$-consistent, and the asymptotic distribution is generally unknown.

The suggestion of this paper is to assume that there exists a set of strictly exogenous time-invariant regressors z_i such that $\mu_i := f(z_i)$. This is a stronger assumption than the one

made in Example 1, in that the model does not allow for the existence of the pure random effects v_i. To is why, note that the lagged dependent variable y_{it-1} would necessarily depend on v_i, which would violate the independence assumption required to derive the estimator. The assumption made here implies that

$$y_{it} = 1\{\alpha y_{it-1} + w_{it}\gamma + f(z_i) - u_{it} \geq 0\} \qquad (i = 1, \cdots, N; \quad t = 1, \cdots, T). \tag{2.5}$$

Defining $x_{it} := (y_{it-1}, w_{it})$ and $\varepsilon_{it} := y_{it} - E[y_{it}|x_{it}, z_i]$ obtains equation (1.1). The estimator derived in this paper uses only the information provided in equation (1.1). The resulting minimum distance estimator therefore does not require modeling the initialization of y_{it}. This implies that the resulting estimator is not subject to the initial conditions problem (see Honore and Tamer (2006)) in that it is robust to mis-specification of the distribution of y_{i0} conditioned on μ_i.

These two examples show that under certain assumptions and by appropriately defining z_i, equation (1.1) is implied by a variety of models that are popular in applied work. Equation (1.1) is also of interest in its own right. It extends the GLM model of Chen (1995) by relaxing the parametric specification of the link function.

Returning to equation (1.1), define the conditioning vector $w_{it} := (x_{it}, z_i)$. By taking conditional expectations of y_{it} conditioned on w_{it} in equation (1.1) we obtain

$$P_{it} := E(y_{it} \mid w_i) = \Phi_t(x_{it}\beta + f(z_i)), \quad (i = 1, \cdots, N; \quad t = 1, \cdots T). \tag{2.6}$$

We formalize the monotonicity constraint on the index function that will be maintained in this paper in the following assumption.

Assumption 2.1. *For $t = 1, \cdots, T$, the index function $\Phi_t : \Re \longrightarrow \Re$ is strictly increasing.*

Under assumption 2.1 the index function can be inverted. Define the inverse index function $\varphi_{t0} := \Phi_{t0}^{-1}$. Equation (2.6) implies that

$$\varphi_{t0}(P_{it}) = x_{it}\beta_0 + f_0(z_i), \quad (i = 1, \cdots, N; \quad t = 1, \cdots T), \tag{2.7}$$

which in turn implies

$$\Delta[\varphi_{t0}(P_{it})] = \Delta x_{it}\beta_0, \quad (i = 1, \cdots, N; \quad t = 2, \cdots T), \tag{2.8}$$

where $\Delta[\varphi_{t0}(P_{it})] := \varphi_{t0}(P_{it}) - \varphi_{t-1,0}(P_{it-1})$ and $\Delta x_{it}\beta_0 := (x_{it}\beta_0 - x_{it-1}\beta_0)$. The time invariant restriction on $f_0(z_i)$ is implicitly imposed by the first differencing of equation (2.7), and will therefore not need to be made explicit in estimation. Since $f_0(z_i)$ will not be estimated jointly with the other parameters of the model, the computational cost due to the possibly large dimension of z_i is incurred only once in the estimation of P_i.

3 Identification

Define $\varphi := (\varphi_1, \cdots, \varphi_T)$. The parameter vector we are interested in identifying is denoted by $\pi = (\beta, \varphi, f)$. The goal of the section is to prove that under a set of assumptions, there is a unique parameter vector $\pi_0 = (\beta_0, \varphi_0, f_0)$ that satisfies equation (2.6). Let $\|\cdot\|$ be a norm on $\Re^K$. The restrictions are formally stated in the the following assumption.

Assumption 3.1. *1. For at least one $k \in [1, \cdots, K]$, x_{ik} is not contained in z_i. Without loss of generality, let $k = K$.*

2. $rank(E[\Delta x_{it}' \Delta x_{it}]) = K$.

3. $\|\beta\| = 1$ and $E[\varphi(P_i)] = 0$.

Assumption 3.1.1 is satisfied if the set of regressors contain predetermined variables and z_i is composed of all strictly exogenous variables for individual i. It is also satisfied if z_i is composed of time invariant characteristic of the individual, as discussed in the first example. In the case where all the explanatory variables are strictly exogenous, this assumption means that one of the regressors is excluded from z_i. A similar conditional independence assumption is used in Honoré and Lewbell (2002) to obtain identification of their finite dimensional parameter vector. Honoré and Lewbell (2002) impose no other restrictions on the dependence between the individual effect and the other regressors. The estimator proposed in this paper therefore makes more restrictive assumptions on the dependence between the individual effect and the other regressors than the estimator developed

in Honoré and Lewbell (2002). However, the estimator proposed in this paper provides a convenient framework for predictions and simulations (conditioned on the observables). As shown in Newey (1994a), this conditional independence assumption is not needed if it is assumed that the index function is known.

Part 2 of assumption 3.1 is the full rank assumption needed for identification of the model π. It requires that x_{it} does not contain time-constant random variables. However, the effect of time-constant random variables can be controlled for by including them in z_i.

Part 3 of assumption 3.1 are the scale and location normalizations required for point identification of the model π. The assumption that $\|\beta\| = 1$ fixes the scale of the parameter in the model. This normalization is frequent in single index models (see Manski, 1985 and Manski, 1987 for example). An alternative normalization (see Horowitz, 1992 and Ichimura, 1993) is to assume that the first component of x_{it} has a probability distribution conditional on the remaining components that is absolutely continuous with respect to the Lebesgue measure, and then assume that $|\beta_1| = 1$. Identification of the model can also be proven under this alternative normalization. The assumption that $E[\varphi(P_i)] = 0$ fixes the location of the φ's and f. This is one of many alternative normalizations that can be imposed. This particular normalization is chosen because of it is easy to implement in proposed algorithm.

Assume that the parameter vector π_0 satisfies the restrictions in Assumption 3.1. Let the alternative model $\pi_1 = (\beta_1, \Phi_1, f_1)$ be observationally equivalent to π_0 in that

$$P_{it} = \Phi_{t1}\left(x_{it}\beta_1 + f_1(z_i)\right), \quad (i = 1, \cdots, N; \quad t = 1, \cdots T). \tag{3.1}$$

The identification theorem is stated as follows.

Theorem 3.2. *(Identification) If* (i) $(\Phi_{t1},\ t = 1, \cdots, T)$ *satisfy assumption 2.1, and* (ii) π_1 *satisfies assumption 3.1, then* $\beta_0 = \beta_1$, $f_0 = f_1$, *and for* $t = 1, \cdots, T$, $\varphi_{t,0} = \varphi_{t,1}$.

Proof. See appendix A.1 □

4 The Estimator

Suppose a sample of N independent realizations $(y_{it}, x_{it}, z_i \; t = 1, \cdots, T; \;\; i = 1, \cdots, N)$ are drawn from the distribution of the $T \times (K + L + 1)$-dimensional random matrix (y, x, z) with support $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$, where $\mathcal{Y} \subseteq \Re$, $\mathcal{X} \subseteq \Re^K$, and $\mathcal{Z} \subseteq \Re^L$. Let and $w := (x, z)$ and let $f_w(w)$ be the probability density function of the distribution function defined on $\mathcal{X} \times \mathcal{Z}$ with respect some dominating measure.

Because the predicted outcomes $P_{it} := E[y_{it}|w_{it}] = \int (y f(y, w_{it}) \nu(\mathrm{d}y)) / f_w(w_{it})$ has the density density f_w in the denominator, f_w must be bounded away from zero. We therefore impose a fixed trimming condition by defining the compact subset $\mathcal{W} \subset \mathcal{X} \times \mathcal{Z}$ where $f_w(w)$ is bounded away from zero on $\mathcal{W}$. This fixed trimming condition imply that there is a compact connected subset $\mathcal{K} \subset \Re$ in which all the P's lie. Let $\Lambda^2_{c_2}(\mathcal{K}) := \{ f \in \mathcal{C}^2(\mathcal{K}) : \| f \|_{s,2} \leq c_2 < \infty \}$, where $\| \cdot \|_{s,2}$ is the supremum Sobolev norm (see Newey (1994b)), and $\mathcal{S}_{\mathcal{K}}$ be a compact subset of $\Lambda^2_{c_2}(\mathcal{K})$, composed of strictly increasing functions. Define the function Δ as $a := (a_1, \ldots, a_T)' \longmapsto \Delta a := (a_2 - a_1, \ldots, a_T - a_{T-1})'$ and let

$$
\begin{aligned}
\mathcal{F} &:= \left\{ a \mapsto \Delta f(a) | a \in \Re^T, f(a) = (f_1(a_1) \cdots, f_T(a_T))', f_t : \Re \mapsto \Re \right\}, \\
\mathcal{F}_c &:= \left\{ a \mapsto \Delta f(a) \in \mathcal{F} \,|\, f_t \in \mathcal{S}_{\mathcal{K}}, t = 1, \cdots, T \right\}.
\end{aligned}
$$

Assume that $\theta_0 := (\beta_0', \varphi_0)' \in \Theta := \mathcal{B} \times \mathcal{F}_c$, where $\mathcal{B} \subseteq \Re^K$ is compact and convex with non-empty interior. We remark that the vector $\Delta x \beta$ is an element of the space $\mathcal{F}$, and $\Delta[\varphi(P)] := (\varphi_2(P_2) - \varphi_1(P_1), \cdots, \varphi_T(P_T) - \varphi_{T-1}(P_{T-1}))'$ is an element of $\mathcal{F}_c$. We further require that the induced density $f_P(P)$ also be bounded away from zero on $\mathcal{K}$. This holds in general given boundedness conditions on f_w and y (see Mood et al. (1974), sections 5 and 6 for detailed discussions). Define the indicator function $\tau_{it} = 1\{ w_{it} \in \mathcal{W} \}$, let $\tau_i := \prod_{t=1}^T \tau_{it}$ and define the residual vector $\rho(w, \theta) := (\Delta[\varphi(P)] - \Delta x \beta)$. Let $\breve{\theta}$ minimize the following objective function

$$
Q_0(\theta) := E\left[\tau \rho(w, \theta)' [\Sigma]^{-1} \rho(w, \theta) \right], \tag{4.1}
$$

where Σ is a $(T-1)$-dimensional symmetric, positive definite weighting matrix for any given w. In general, $\breve{\theta}$ will be set valued. However, the identification results of theorem 3.2 imply that the transformation $\theta_0 := (\breve{\beta}/a, \{(\breve{\varphi}_t - c_t)/a, \; t = 1, \cdots, T\})$, where $a := \|\breve{\beta}\|$ and $c_t := E[\tau_i \breve{\varphi}_t(P_{it})]$, maps $\breve{\theta}$ onto a singleton.

Estimation of θ_0 from the sample analog of equation (4.1) is infeasible because the predicted outcomes P_{it} are unknown. To overcome this problem, we replace P_{it} with a consistent kernel estimator $\hat{P}_{it}$. Let σ_1 be a positive constant. Define the function $K_{it}(w) := \sigma_1^{-(K+L)} K_1(\sigma_1^{-1}(w - w_{it}))$, where K_1 is a Kernel. Let $q_{it} = (1, y_{it})$ and define $\hat{\gamma}(w) = (\hat{\gamma}_1(w), \hat{\gamma}_2(w))$ by

$$\hat{\gamma}(w) \quad := \quad (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} q_{it} K_{it}(w).$$

Then the estimated conditional mean is defined by $\hat{P}_{it} = \hat{\gamma}_2(w_{it}) / \hat{\gamma}_1(w_{it})$, and the estimate of the probability density function (pdf) of w, $f_w(w)$ is $\hat{f}_w(w) = \hat{\gamma}_1(w)$. We assume also that $\mathcal{W}$ is chosen so that the estimated density $\hat{f}_w(w)$ is bounded away from zero on $\mathcal{W}$.

To our knowledge, there exists no estimator defined as the infimum of a sample analog to equation (4.1) that uses kernels to estimate the index functions. There are now well established methods for estimating θ_0 using sieves as the smoother for the infinite dimensional parameters (see Newey and Powell (2003), Ai and Chen (2003), and Chen (2007) for examples). These SMD estimators also have the desirable property of semiparametric efficiency given appropriate choice of the weighting matrix. In this section, we propose an estimator that is based purely on kernels that also achieves the semiparametric efficiency bound. Indeed, the estimator presented in this section can be implemented with any type of smoother as discussed in Mammen et al. (2001). In order to define the estimator, let

$$\begin{aligned}
\mathcal{F}^N &:= \{m = (m^i, i = 1, \cdots, N) : m^i \in \mathcal{F}\}, \\
\mathcal{F}_m^N &:= \{m \in \mathcal{F}^N : m^i \text{ does not depend on } i\}, \\
\mathcal{F}_c^N &:= \left\{m \in \mathcal{F}_m^N : m^i \in \mathcal{F}_c\right\}.
\end{aligned}$$

We remark that the vector $(x_i\beta,\ i = 1, \cdots, N)$ is an element of $\mathcal{F}^N$. $\mathcal{F}^N$ is a vector space when endowed with the operations "$+$" and "$\cdot$" defined as

$$\begin{aligned}
m + g &:= (m^i + g^i,\ i = 1, \cdots, n), \text{ for } m, g \in \mathcal{F}, \\
\alpha \cdot m &:= (\alpha m^i,\ i = 1, \cdots, n), \text{ for } \alpha \in \Re,\ m \in \mathcal{F}.
\end{aligned}$$

Define $\omega_{it}(P) := \sigma_2^{-1} K_2(\sigma_2^{-1}(\hat{P}_{it} - P))$, where σ_2 is a positive constant and K_2 is a ker-

nel. Then $\hat{f}_t(P_t) := N^{-1}\sum_{i=1}^N \omega_{it}(P_t)$ is the estimated marginal density of P_t. Let $\omega_i(P) := \prod_{t=1}^T \omega_{it}(P_t)$. The estimated joint density of $P = (P_1, \cdots, P_T)$ is given by $\hat{f}(P) := N^{-1}\sum_{i=1}^N \omega_i(P)$. Define the inner product on $\mathcal{F}^N$ by

$$\langle m, g\rangle_T := \int \frac{1}{N}\sum_{i=1}^N \tau_i m^i(P)' W g^i(P)\omega_i(P)\mathrm{d}P,$$

for some positive definite matrix W. This inner product induces the following semi-norm on $\mathcal{F}^N$

$$\|m\|_T^2 := \int \frac{1}{N}\sum_{i=1}^N \tau_i m^i(P)' W m^i(P)\omega_i(P)\mathrm{d}P.$$

Define the sample residual vector of functions $\rho(w_i, P, \theta) := (\Delta[\varphi(P)] - \Delta x_i\beta)$, and let $\breve{\theta} \in \Theta$ be the solution to

$$\min_{\theta\in\Theta} \hat{Q}_N(\theta) = \min_{\theta\in\Theta} \int N^{-1}\sum_{i=1}^N \tau_i \rho(w_i, P, \theta)'\hat{\Sigma}^{-1}\rho(w_i, P, \theta)\omega_i(P)\mathrm{d}P, \tag{4.2}$$

where $\hat{\Sigma}$ is a consistent estimator of Σ. Again, $\breve{\theta}$ will typically be set valued. The feasible semiparametric minimum distance estimator of θ_0 is given by $\hat{\theta} := (\breve{\beta}_N/a_N, \{(\breve{\varphi}_t - c_{N,t})/a_N, t = 1, \cdots, T\})$, where $a_N := \|\breve{\beta}_N\|$ and $c_{N,t} := N^{-1}\sum_i \tau_i \breve{\varphi}_t(\hat{P}_{it})$.

Remark 4.1. It is not obvious that the solution $\breve{\theta}$ defined in (4.2) exists, and if it does, whether it is unique. We therefore state the following lemma.

Lemma 4.2. *The minimization problem (4.2) has a unique solution.*

Proof. See appendix A.2. □

Remark 4.3. For the semi-norm defined above to be well-defined, we require that $\omega_{it} \geq 0$, and $\omega_{it} = 0$ on a set of measure zero. An important consequence of this restriction is that higher order kernels cannot be used in the definition of ω_{it}. It would seem therefore that the estimator of the finite dimensional parameter cannot obtain the parametric rate of convergence. However, the proposed estimator works by first concentrating out the index functions (estimating them as function of β) and then estimate β. This allows us to apply Proposition 2 of Newey (1994a) concerning how estimation of the nuisance parameter affects the asymptotic variance of the estimator of the finite dimensional parameter, the result being that $\hat{\beta}$ is $\sqrt{N}$-consistent.

Remark 4.4. In order to compute 4.2, a consistent estimator $\hat{\Sigma}$ of Σ is required. This is achieved by implementing a two step approach where in the first step $\hat{\Sigma}$ is replaced with the $(T-1)$-dimensional identity matrix. This obtains consistent estimates of θ_0. These consistent estimates are then used to construct $\hat{\Sigma}$, which is used in the second stage estimator. Details of the construction are found in the next Section 6.

5 Computing the Estimator

It is possible to define a feasible empirical analog to (4.1) by implementing the method series or sieve estimation developed in by Newey and Powell (2003), Ai and Chen (2003) and Chen (2007). To the best of our knowledge however, there has been no work in the econometrics literature that shows how to compute panel data estimators such as equation (4.2) using kernel estimators with monotonicity and additivity constraints. Since kernel estimation is still the workhorse in the nonparametric literature, we find it pertinent to present such a method. The method presented in this section develops a technique that makes use of the method of alternating projections (Bauschke and Borwein, 1996; Deutsch, 2001) and backfitting algorithm developed in Hastie and Tibshirani (1986), Buja et al. (1989), Mammen et al. (1999) and Mammen et al. (2001). To begin, we impose the restrictions on the kernel K_2 that will be needed for the derivation of the algorithm and to prove its convergence.

Assumption 5.1. *For $d \geq 2$, $K_2(s)$ is differentiable of order d, the d-th derivatives bounded uniformly, $K_2(s)$ is zero outside a bounded set, $K_2(s) \geq 0$, $\int K_2(s)ds = 1$, $\int sK_2(s)ds = 0$, and $\int |K_2(s)|^2 ds < \infty$.*

We begin by defining the projection of $\Delta x\beta$ onto $\mathcal{F}_c^N$ for a fixed β. This projection is defined as the fixed point to a backfitting algorithm. Proposition 1 of Mammen et al. (2001) suggests that this projection can be decomposed into three cascading projections. The first is the projection of $\Delta x\beta$ onto the set $\times_{s=1}^{T} C^2$ to obtain the $T-1$-dimensional unconstrained estimator $\hat{m}(\beta) := (\hat{m}_2(\beta), \cdots, \hat{m}_T(\beta))'$ defined by

$$\hat{m}(\beta) = \arg \min_{\tilde{m} \in \times_{s=1}^{T} C^2} \int \frac{1}{N} \sum_{i=1}^{N} \tau_i (\tilde{m} - \Delta x_i \beta)' \hat{\Sigma}^{-1} (\tilde{m} - \Delta x_i \beta) \omega_i(P) dP.$$

The solution can be computed for each P individually, suggesting the following minimization problem

$$\hat{m}(P;\beta) = \arg \min_{\tilde{m} \in \times_{s=1}^{T} \mathcal{C}^2} \frac{1}{N} \sum_{i=1}^{N} \tau_i (\tilde{m} - \Delta x_i \beta)' \hat{\Sigma}^{-1} (\tilde{m} - \Delta x_i \beta) \omega_i(P), \tag{5.1}$$

with the solution given by $\hat{m}_t(P,\beta) := N^{-1} \sum_{i=1}^{N} \tau_i \Delta x_{it} \beta \omega_i(P) / \hat{f}_P(P), t = 2, \cdots, T$.

We next define the empirical projection estimator $(\breve{\varphi}_1, \breve{\varphi}_2)$ as minimizers of

$$\|\hat{m}_2(\beta) - \tilde{\varphi}_2 + \tilde{\varphi}_1\|_T^2 = \int [\hat{m}_2(P;\beta) - \tilde{\varphi}_2 + \tilde{\varphi}_1]^2 \hat{f}(P) dP, \tag{5.2}$$

with the solution characterized by the following:

$$\begin{aligned}
\breve{\varphi}_2 &= \int \hat{m}_2(P;\beta) \frac{\hat{f}(P)}{\hat{f}_2(P_2)} dP_{-2} + \int \breve{\varphi}_1(P_1) \frac{\hat{f}(P)}{\hat{f}_2(P_2)} dP_{-2}, \\
\breve{\varphi}_1 &= \int \breve{\varphi}_2(P_2) \frac{\hat{f}(P)}{\hat{f}_1(P_1)} dP_{-1} - \int \hat{m}_2(P;\beta) \frac{\hat{f}(P)}{\hat{f}_2(P_2)} dP_{-1},
\end{aligned}$$

where dP_{-t} is the Lebesgue measure on the vector $(P_s,\ s \neq t)$. Straightforward calculations show that this system of equations reduces to

$$\begin{aligned}
\breve{\varphi}_2 &= \frac{1}{N} \sum_{i=1}^{N} \tau_i \Delta x_{i2} \beta \frac{\omega_{i2}(P_2)}{\hat{f}_2(P_2)} + \frac{1}{N} \sum_{i=1}^{N} \tau_i \left(\int \breve{\varphi}_1(P_1) \omega_{i1}(P_1) dP_1 \right) \frac{\omega_{i2}(P_2)}{\hat{f}_2(P_2)}, \\
\breve{\varphi}_1 &= \frac{1}{N} \sum_{i=1}^{N} \tau_i \left(\int \breve{\varphi}_2(P_2) \omega_{i2}(P_2) dP_2 \right) \frac{\omega_{i1}(P_1)}{\hat{f}_1(P_1)} - \frac{1}{N} \sum_{i=1}^{N} \tau_i \Delta x_{i2} \beta \frac{\omega_{i1}(P_1)}{\hat{f}_1(P_1)}.
\end{aligned}$$

To minimize computation costs, we will approximate $(\int \breve{\varphi}_1(P_1) \omega_{i1}(P_1) dP_1)$ by $\breve{\varphi}_1(P_{i1})$. Under Assumption 5.1, Gayle (2008) shows that the difference in these two quantities is $o_P(1)$. We also approximate $(\int \breve{\varphi}_2(P_2) \omega_{i2}(P_2) dP_2)$ by $\breve{\varphi}_2(P_{i2})$. With these approximations we have the following solutions

$$\begin{aligned}
\breve{\varphi}_2(P_2) &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \tau_i \omega_{i2}(P_2) \cdot (\Delta x_{i2} \beta + \breve{\varphi}_1(P_{i1})) / \hat{f}_2(P_2), \\
\breve{\varphi}_1(P_1) &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \tau_i \omega_{i1}(P_1) \cdot (\breve{\varphi}_2(P_{i2}) - \Delta x_{i2} \beta) / \hat{f}_1(P_1).
\end{aligned} \tag{5.3}$$

The third step is to project these solutions into the space of increasing functions. The

results of Brunk (1958), and Mammen et al. (2001) imply that

$$\begin{aligned}\varphi_2^*(P) &= \inf_{v\geq P}\sup_{u\leq P}\frac{\int_{s=u}^{v}\breve{\varphi}_2(s)\hat{f}_2(s)\mathrm{d}s}{\int_{s=u}^{v}\hat{f}_2(s)\mathrm{d}s},\\ \varphi_1^*(P) &= \inf_{v\geq P}\sup_{u\leq P}\frac{\int_{s=u}^{v}\breve{\varphi}_1(s)\hat{f}_1(s)\mathrm{d}s}{\int_{s=u}^{v}\hat{f}_1(s)\mathrm{d}s}.\end{aligned} \tag{5.4}$$

For fixed β, the backfitting algorithm therefore works as follows.

Inner Backfitting Algorithm (IBA)
Step 1. Obtain an initial estimator $(\varphi^{*[0]}(P_{i1}),\ i=1,\cdots,N)$.
Step 2. Apply the following loop:
Do for $r\geq 1$

$$\begin{aligned}\breve{\varphi}_2^{[r]}(P) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\tau_i\omega_{i2}(P)\cdot(\Delta x_{i2}\beta+\varphi_1^{*[r-1]}(P_{i1}))/\hat{f}_2(P)\\ \varphi_2^{*[r]}(P) &= \inf_{v\geq P}\sup_{u\leq P}\frac{\int_{s=u}^{v}\tilde{\varphi^{[r]}}_2(s)\hat{f}_2(s)\mathrm{d}s}{\int_{s=u}^{v}\hat{f}_2(s)\mathrm{d}s}\\ \varphi_2^{*[r]}(P) &= \varphi_2^{*[r]}(P)-\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\varphi_2^{*[r]}(P_{2i})\\ \breve{\varphi}_1^{[r]}(P) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\tau_i\omega_{i1}(P)\cdot(\varphi_2^{*[r]}(P_{i2})-\Delta x_{i2}\beta)/\hat{f}_1(P)\\ \varphi_1^{*[r]}(P) &= \inf_{v\geq P}\sup_{u\leq P}\frac{\int_{s=u}^{v}\breve{\varphi}_1^{[r]}(s)\hat{f}_1(s)\mathrm{d}s}{\int_{s=u}^{v}\hat{f}_1(s)\mathrm{d}s}\\ \varphi_1^{*[r]}(P) &= \varphi_1^{*[r]}(P)-\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\varphi_1^{*[r]}(P_{1i})\end{aligned} \tag{5.5}$$

until convergence in $(\varphi_2^*,\varphi_1^*)$ is reached.
Convergence of the IBA is stated in the following theorem.

Theorem 5.2. *(Convergence of IBA) Suppose that the assumptions of 5.1 hold. Then there exists a solution* $(\varphi_1^*,\varphi_2^*)$ *of the system of equations (5.5).*

Proof. See appendix A.3. □

Given the estimates $(\varphi_1^*,\varphi_2^*)$, and for fixed β, estimates of $(\varphi_t^*,\ t=3,\cdots,T)$ are derived by similiar computations as follows:

$$\begin{aligned}\breve{\varphi}_t(P) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\tau_i\omega_{i2}(P)\cdot(\Delta x_{it}\beta+\varphi_{t-1}^*(P_{i1}))/\hat{f}_2(P),\\ \varphi_t^*(P) &= \inf_{v\geq P}\sup_{u\leq P}\frac{\int_{s=u}^{v}\breve{\varphi}_t(s)\hat{f}_2(s)\mathrm{d}s}{\int_{s=u}^{v}\hat{f}_2(s)\mathrm{d}s},\end{aligned} \tag{5.6}$$

followed by the mean normalization. Given estimates of $\varphi_t, t = 1, \cdots, T$, the next step is to project this solution (an element of $\mathcal{F}_c$) onto $x\mathcal{B}$. This amount to substituting the $\Delta\varphi_t^*(P)$'s into (5.2) and solving the problem for β. This stage of the problem has a closed form solution given as follows:

$$\breve{\beta} = \left[\sum_{i=1}^{N} \tau_i \Delta x_i' \hat{\Sigma}^{-1} \Delta x_i\right]^{-1} \left[\sum_{i=1}^{N} \tau_i \Delta x_i' \hat{\Sigma}^{-1} \Delta[\varphi^*(\hat{P}_i)]\right]. \tag{5.7}$$

For an arbitrary initial choice of β, say $\breve{\beta}^{[0]}$, the outer backfitting algorithm therefore works as follows.

Outer Backfitting Algorithm (OBA)
Do for $s \geq 1$
Step 1. Compute the updated estimates $(\varphi_1^{*[s]}, \varphi_2^{*[s]})$ by implementing the IBA initialized by $\varphi_1^{*[s-1]}$ and β fixed at $\breve{\beta}^{[s-1]}$.
Step 2. Compute the updated estimates of $(\varphi_t^{*[s]},\ t = 3, \cdots, T)$ by implementing the system (5.6).
Step 3. Update β using equation (5.7), i.e.,

$$\breve{\beta}^{[s]} = \left[\sum_{i=1}^{N} \tau_i \Delta x_i' \hat{\Sigma}^{-1} \Delta x_i\right]^{-1} \left[\sum_{i=1}^{N} \tau_i \Delta x_i' \hat{\Sigma}^{-1} \Delta\left[\varphi^{*[s]}(\hat{P}_i)\right]\right].$$

until convergence in β is reached.

The final step in computing the estimator is to impose the normalization constraints. For $\hat{a} := \|\breve{\beta}\|$ the normalized estimates of the parameters of the model are given by $\hat{\beta} = \breve{\beta}/\hat{a}$ and $\hat{\varphi}_t = \varphi_t^*/\hat{a},\ t = 1, \cdots, T$.

To see that the sequence $\left\{\breve{\beta}^{[s]}, \Delta[\varphi^{*[s]}], s \geq 0\right\}$ defined by the OBA does converge, note that the solution is characterized by the system of equations

$$\begin{aligned} \Delta[\varphi^*(P)] &= \arg \inf_{m \in \mathcal{F}_c^N} \|\Delta x \breve{\beta} - m(P)\|_T, \\ \Delta x \breve{\beta} &= \arg \inf_{a \in x\mathcal{B}} \|a - \Delta[\varphi^*(P)]\|_T. \end{aligned}$$

This makes it clear that the iteration of the OBA defines a series of alternating projections between two convex and closes sets $x\mathcal{B}$ and $\mathcal{F}_c^N$. This intuition is formally stated in the following theorem.

Theorem 5.3. *(Convergence of the OBA) Suppose the assumptions of 5.1 hold. Then there exists a solution of the OBA.*

Proof. See appendix A.4. □

The final issue to cover is that of obtaining estimates of the weighting matrix $\hat{\Sigma}$. Assuming that this can be calculated from consistent estimates of β and φ, we propose a two-step procedure similar to that of the two-step efficient GMM estimator. The first stage replaces the weighting matrix with the identity matrix to obtain initial consistent estimates of β and φ. These first-stage estimates are then used to compute an estimate of the weighting matrix, which is used in the second stage to obtain the second-stage estimator of the parameters of interest. In the next section, we derive the efficient weighting matrix Σ and a propose a consistent estimator of this weighting matrix that can be computed from initial consistent estimates of β and φ.

6 Asymptotic properties of the estimator

In order to derive the asymptotic properties of the estimator, some regularity conditions must be imposed. We turn first to the nuisance parameter, the first stage kernel estimator of $P_{it0} = E[y_{it}|w_{it}]$. We impose conditions that ensure uniform convergence of the nonparametric estimate $\hat{P}_{it0}$. Define $\gamma_0 := (\gamma_{10}, \gamma_{20})$ where $\gamma_{10} := f_w(w_{it})$ and $\gamma_{20} := f_w(w_{it})E[y_{it}|w_{it}]$. Clearly $P_{it0} = \gamma_{20}/\gamma_{10}$. We make the following assumptions

Assumption 6.1. *1. $K_1(u)$ is differentiable of order $d \geq 2$, the derivatives d are bounded, $K_1(u)$ is zero outside a bounded set, $\int K_1(u)du = 1$, there is a positive integer m such that for all $j < m$, $\int K_1(u)[\otimes_{\ell=1}^{j} u]du = 0$. 2. There is a version of $\gamma_o(w)$ that is continuously differentiable to order d with bounded derivatives on an open set containing $\mathcal{W}$. 3. There*

is $p \geq 4$ *such that* $E[\|\tilde{y}\|^p] < \infty$ *and* $E[\|\tilde{y}\|^p|w]\, f_0(w)$ *is bounded. 4. The bandwidth* $\sigma_1 = \sigma_1(N)$ *satisfies* $N^{1-(2/p)}\sigma_1^{K+L}/\ln N \longrightarrow \infty$, $\sqrt{N}\sigma_1^{2m} \longrightarrow 0$, *and* $\sqrt{N}\ln N/(N\sigma_1^{K+L+2d}) \longrightarrow 0$

Assumption 6.1 ensures that the nuisance parameters $\hat{P}_i$ converges to the true conditional expectation at a fast enough rate to ensure $\sqrt{N}$-convergence of the finite dimensional parameter estimate $\hat{\beta}$. This result is proven and discussed in Newey and McFadden (1994) and Newey (1994b). Define $m_{t,0}(P;\beta) := E[\Delta x_t|P], t = 2,\cdots,T$. We require the following assumptions on the $m_{t,0}$ and the bandwidth σ_2.

Assumption 6.2. *1. For* $t = 2,\cdots,T$ *and fixed* $\beta \in \mathcal{B}$, *there is a version of* $m_{t,0}(P;\beta)$ *that is continuously differentiable to order* 2 *with bounded derivatives on an open set containing* $\mathcal{W}$. *2. For* $t = 2,\cdots,T$, $E[\|\Delta x_{it}\|^2] < \infty$. *3.* $\sigma_2 \to 0$ *and* $n\sigma_2^{T+1} \to \infty$ *as* $n \to \infty$.

Assumption 6.2 is standard in the nonparametric literature to obtain consistency of the estimators of nonparametric components φ_t (See Pagan and Ullah (1999) and Hardle et al. (2004) for discussions). Define the distance d on Θ as follows:

$$d[(\beta,\phi),(\alpha,\psi)] := \|\beta-\alpha\|_K + \sum_{t=1}^{T} \|\phi_t(P) - \psi_t(P)\|_{s,2}$$

where $\|\cdot\|_K$ is the Euclidean norm on $\Re^K$ and $\|\cdot\|_{s,2}$ is the supremum Sobolev norm of smoothness 2. In what follows, we denote the first stage estimator by $(\tilde{\beta},\tilde{\varphi})$ and the second stage estimator by $(\hat{\beta},\hat{\varphi})$. We now state the consistency and asymptotic normality theorems.

Theorem 6.3. *Let the assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied. Then* $\tilde{\beta} \xrightarrow{p} \beta_0$, *and for* $t = 1,\cdots,T$, $\|\tilde{\varphi}_t(P) - \varphi_{t,0}(P)\|_{s,2} \xrightarrow{p} 0$.

Proof. See Appendix A.5. □

Define

$$R(P_i) := \begin{bmatrix} -\varphi'_{t-1}(P_{i1}) & \varphi'_t(P_{i2}) & 0 & \cdots & 0 & 0 \\ 0 & -\varphi'_{t-1}(P_{i2}) & \varphi'_{t-1}(P_{i3}) & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -\varphi'_{T-1}(P_{i,T-1}) & \varphi'_T(P_{iT}) \end{bmatrix}, \quad (6.1)$$

$\varepsilon_{it} := y_{it} - P_{it}$ and $\varepsilon_i := (\varepsilon_{i1}, \cdots, \varepsilon_{iT})'$. The weighting matrix that is used to define the second stage estimator is $\Sigma := E[R(P_i)\varepsilon_i\varepsilon_i' R(P_i)']$. The proposed estimator for Σ is given by

$$\hat{\Sigma} := \frac{1}{N}\sum_{i=1}^{N} \hat{R}(\hat{P}_i)\hat{\varepsilon}_i\hat{\varepsilon}_i'\hat{R}(\hat{P}_i)', \tag{6.2}$$

where, analogously,

$$\hat{R}(\hat{P}_i) := \begin{bmatrix} -\tilde{\varphi}_{t-1}'(\hat{P}_{i1}) & \tilde{\varphi}_t'(\hat{P}_{i2}) & 0 & \cdots & 0 & 0 \\ 0 & -\tilde{\varphi}_{t-1}'(\hat{P}_{i2}) & \tilde{\varphi}_{t-1}'(\hat{P}_{i3}) & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -\tilde{\varphi}_{T-1}'(\hat{P}_{i,T-1}) & \tilde{\varphi}_T'(\hat{P}_{iT}) \end{bmatrix}, \tag{6.3}$$

$\hat{\varepsilon}_{it} := y_{it} - \hat{P}_{it}$, and $\hat{\varepsilon}_i := (\hat{\varepsilon}_{i1}, \cdots, \hat{\varepsilon}_{iT})'$. The proof of the asymptotic properties of the second stage estimator requires the following lemma.

Lemma 6.4. *Let assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied. Then* $\|\Sigma\| < \infty$, *and* $\hat{\Sigma} \xrightarrow{p} \Sigma$.

Proof. See Appendix A.6. □

We now state the consistency theorem for the second stage theorem.

Theorem 6.5. *Let the assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied.Then* $\hat{\beta} \xrightarrow{p} \beta_0$, *and for* $t = 1, \cdots, T$, $\|\hat{\varphi}_t(P) - \varphi_{t,0}(P)\|_{s,2} \xrightarrow{p} 0$.

Proof. See Appendix A.7. □

Finally, we state the theorem defining asymptotic normality of both the first and second stage finite dimensional estimators.

Theorem 6.6. *If the assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 are satisfied, then*

$$\sqrt{N}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V_1),$$

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_2),$$

where

$$V_1 := \left[E[\tau_i h_{i0}' h_{i0}]\right]^{-1} E[\tau_i h_{i0}' \Sigma^{-1} h_{i0}] \left[E[\tau_i h_{i0}' h_{i0}]\right]^{-1},$$

$$V_2 := \left[E[\tau_i h_{i0}' \Sigma^{-1} h_{i0}]\right]^{-1},$$

and $h_{i0} := \frac{\partial}{\partial \beta} \Delta\varphi_0(P_{i0}; \beta_0) - \Delta x_i$

Proof. See Appendix A.8. □

6.1 Semiparametric Efficiency Bound

We now tackle the question of whether the proposed estimator of the finite-dimensional parameter β is efficient. The model for which we compute the efficiency bound is the implied model given in equation (1.1). Chamberlain (1993) shows that models defined in section 2 are not efficient when no restrictions are made on the index function and the individual specific effects. The assumptions made in this model are therefore substantive from this point of view. The variance bound that we compute for equation (1.1) is the one that would be attained within an SMD framework. This is not surprising given proposition 1 of Newey (1994a), which states that the asymptotic variance of the semiparametric estimator depends on the nonparametric function that is being estimated, and not on the type of smoother used to estimate estimate it. Thus our estimation framework is as efficient as any competing extremum estimator for the condition given in (1.1), but retains the property that it is independent of the choice of smoother. This results in the following theorem.

Theorem 6.7. *The estimator* $\hat{\beta}$ *of the finite dimensional parameter* β *developed in section 4 is semiparametric efficient with variance bound given in theorem 6.6.*

Proof. See Appendix A.9. □

6.2 Estimating the asymptotic variance

In order to estimate the asymptotic variances V_1 and V_2, one needs to obtain estimates of h_{i0} in both cases, and Σ in the latter case. The feasible estimator of Σ is already defined to be $\hat{\Sigma}$ in equation (6.2) and its convergence to Σ is already shown in Lemma 6.4. An estimator of

$h_{i0} = \frac{\partial}{\partial \beta}\Delta\varphi_0(P_{i0};\beta_0) - \Delta x_i$ requires an estimator of $\frac{\partial}{\partial \beta}\Delta\varphi_0(P_{i0};\beta_0)$. To this end, note that the model (2.8) implies that for $t = 2,\cdots,T$,

$$\Delta\varphi_{t0}(P_{it0};\beta_0) = E[\tau_i \Delta x_{it}\beta_0 | P_{it0}, P_{it-1,0}] = E[\Delta x_{it}|P_{it0}, P_{it-1,0}]\beta_0.$$

This implies that an estimator of $\frac{\partial}{\partial \beta}\Delta\varphi_{t0}(P_{it0};\beta_0)$ can be defined as

$$\frac{\partial}{\partial \beta}\Delta\hat{\varphi}_t(\hat{P}_{it};\hat{\beta}) := \hat{E}[\tau_i \Delta x_{it}|\hat{P}_{it}, \hat{P}_{it-1}],$$

where $\hat{E}[\cdot|\hat{P}_{it}, \hat{P}_{it-1}]$ is some estimator of the conditional expectation, such as a kernel estimator. Given the choice of this conditional expectations estimator, and the convergence results above, it is straightforward to show that $\frac{\partial}{\partial \beta}\Delta\hat{\varphi}_t(\hat{P}_{it};\hat{\beta}) = \frac{\partial}{\partial \beta}\Delta\varphi_{t0}(P_{it0};\beta_0) + o_P(1)$, $t = 2,\cdots,T$. Let $\hat{h}_i = \frac{\partial}{\partial \beta}\Delta\hat{\varphi}(\hat{P}_i;\hat{\beta}) - \Delta x_i$, then feasible estimators of V_1 and V_2 are defined as

$$\hat{V}_1 := \left[N^{-1}\sum_{i=1}^N \tau_i \hat{h}_i'\hat{h}_i\right]^{-1}\left[N^{-1}\sum_{i=1}^N \tau_i \hat{h}_i'\hat{\Sigma}^{-1}\hat{h}_i\right]\left[N^{-1}\sum_{i=1}^N \tau_i \hat{h}_i'\hat{h}_i\right]^{-1},$$

and

$$\hat{V}_2 := \left[N^{-1}\sum_{i=1}^N \tau_i \hat{h}_i'\hat{\Sigma}^{-1}\hat{h}_i\right]^{-1}.$$

We end with the following proposition

Proposition 6.8. *Let assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied, and assume* $\hat{E}[\tau_i \Delta x_{it}|\hat{P}_{it}, \hat{P}_{it-1}] = E[\tau_i \Delta x_{it}|\hat{P}_{it}, \hat{P}_{it-1}] + o_P(1)$. *Then* $\hat{V}_1 \xrightarrow{p} V_1$ *and* $\hat{V}_2 \xrightarrow{p} V_2$.

6.3 Estimating marginal effects

Recall the model implies that the marginal effect of the covariates depends on t. The estimated coefficients $\hat{\beta}$ are not sufficient to characterize these marginal effects of the regressors on the dependent variable. We therefore present feasible estimates of the marginal effects that does not require estimation of the individual-specific effects $f_0(z_i)$. Differenti-

ating equation (1.1) with respect to x_{it} obtains

$$\begin{aligned}\frac{\partial y_{it}}{\partial x_{it}} &= \Phi_{t0}'\left(x_{it}\beta_0 + f_0(z_i)\right)\beta_0 \\ &= (\varphi_{t0}^{-1})'(\varphi_{t0}(P_{it0}))\beta_0 \\ &= \frac{1}{\varphi_{t0}'(P_{it0})}\beta_0.\end{aligned}$$

This implies that a feasible analog estimator of the average time-specific derivative $E_t[\tau_i \partial y_{it}/\partial x_{it}]$ can be obtained as

$$\hat{E}_t[\tau_i \partial y_{it}/\partial x_{it}] = \frac{1}{N}\sum_{i=1}^{N}\frac{\tau_i}{\hat{\varphi}_t'(\hat{P}_{it})}\hat{\beta}.$$

Likewise, a feasible estimator of the average marginal effect can be obtained as

$$\hat{E}[\tau_i \partial y_{it}/\partial x_{it}] = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\tau_i}{\hat{\varphi}_t'(\hat{P}_{it})}\hat{\beta}.$$

The derivative $\hat{\varphi}_t'$ can be taken directly by applying the results of Delfour and Solesio (1987). An alternative approach to estimating the derivative is to differentiate the equality $\Delta\varphi_{t0}(P_{it0};\beta_0) = E[\tau_i \Delta x_{it}\beta_0 | P_{it0}, P_{it-1,0}]$ to obtain

$$\varphi_{t0}'(P_{it0}) = \frac{\partial}{\partial P}\left(\varphi_{t0}(P;\beta_0) - \varphi_{t-1,0}(P_{it-1,0};\beta_0)\right)\Big|_{P=P_{it0}} = \frac{\partial}{\partial P}E[\tau_i \Delta x_{it}\beta_0 | P, P_{it-1,0}]\Big|_{P=P_{it0}},$$

which by the analogy principle implies that an estimator of the derivative is given by

$$\hat{\varphi}_t'(\hat{P}_{it}) = \frac{\partial}{\partial P}\hat{E}[\tau_i \Delta x_{it}\hat{\beta} | P, \hat{P}_{it-1}]\Big|_{P=\hat{P}_{it}}.$$

An immediate consequence of the above results is the following proposition.

Proposition 6.9. *Let assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied, and assume that*

$$\frac{\partial}{\partial P}\hat{E}[\tau_i \Delta x_{it}\hat{\beta} | P, \hat{P}_{it-1}]\Big|_{P=\hat{P}_{it}} = \frac{\partial}{\partial P}E[\tau_i \Delta x_{it}\hat{\beta} | P, \hat{P}_{it-1}]\Big|_{P=\hat{P}_{it}} + o_P(1).$$

Then

$$\sqrt{N}(\hat{E}[\tau_i \partial y_{it}/\partial x_{it}] - E[\tau_i \partial y_{it}/\partial x_{it}]) \xrightarrow{d} N(0, E[\tau_i(\varphi_{t0}'(P_{it0}))^{-1}]^2 V_2).$$

7 Experimental Evidence

In this section, we examine the small sample properties of the estimator via Monte Carlo experiments. Two model specifications are investigated: the first is a static panel data model with a continuous dependent variable and an asymmetric index function. The results are compared to the case where a known symmetric index function is assumed. The second investigates the performance of the estimator for a dynamic panel data probit model with unconditional heteroskedasticity. The results in this second exercise is compared to the case where the investigator correctly assume that the distribution of the error term is Gaussian, but also assumes homoskedasticity of the error term.

7.1 Static panel data model

Consider the following data generating process:

$$y_{it} = \Phi_t(x_{1it}\beta_1 + x_{2it}\beta_2 + f(z_i)) + v_{it}, \quad i = 1, \cdots, N, \quad t = 1, 2, 3. \tag{7.1}$$

In this model, x_{1it} and x_{2it} are both independently distribution as $U(-5,\ 10)$, v_{it} is independently distributed as $N(0,\ 2)$, and $z_i = (x_{2i1} + x_{2i2} + x_{2i3})/3$. The index function is chosen to be asymmetric about zero with range [0, 10]. Specifically, the index function is given by:

$$\begin{aligned} \Phi_t(x) &= \frac{10}{1 + \exp\left(-x\ \lambda(x)\ \sqrt{t}\right)}, \\ \lambda(x) &= 0.2 - \frac{0.1}{1 + \exp(-5x)}. \end{aligned} \tag{7.2}$$

The individual specific function is given by:

$$f(z_i) = 6\left(\frac{\exp(z_i)}{1 + \exp(z_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{\exp(z_i)}{1 + \exp(z_i)}\right). \tag{7.3}$$

Finally, $(\beta_1,\ \beta_2) = (0.6,\ 0.8)$. We perform 100 Monte Carlo replications of the model with three sample sizes N: 100, 200, and 400. The mean bias and the root mean squared error (RMSE) are calculated for each sample size. We also perform the same Monte carlo

Table 1: Small sample properties of the estimator of the static model.

	β_1		β_2	
	Mean Bias	RMSE	Mean Bias	RMSE
N=100				
KMD	0.0061	0.0951	-0.0137	0.0746
EKMD	-0.0026	0.0912	-0.0062	0.0680
RMD	-0.0225	0.2630	-0.0521	0.2026
N=200				
KMD	0.0085	0.0669	-0.0109	0.0527
EKMD	-0.0071	0.0605	-0.0090	0.0469
RMD	0.0085	0.1843	-0.0444	0.1637
N=400				
KMD	0.0044	0.0453	-0.0054	0.0347
EKMD	-0.0016	0.0410	-0.0028	0.0309
RMD	0.0048	0.1407	-0.0236	0.1102

exercise under the assumption that the index function is given by 10Φ where Φ in this case is the standard normal CDF. The results are presented in Table 1. The first stage unrestricted estimator is denoted by KMD, the second stage unrestricted estimator is denoted by EKMD, and the restricted model is denoted by RMD.

The comparison between KMD and EKMD show that KMD always has a higher mean bias and RMSE for both parameters. However, while the difference in the mean bias is substantial, the difference in the RMSE is relatively small. Table 1 also verifies $\sqrt{N}$-convergence for both KMD and EKMD. The restricted estimator RMD performs worse than the unrestricted estimators. Indeed, the RMSE of the RMD is always 3 to 4 times larger than the RME of the EKMD. The results also verifies that lack of parametric rate of convergence of the restricted model. Our experience from this exercise is that the algorithm proposed in Section 5 converges fast, with the IBA converging typically in 1 to 3 iterations, and the OBA converging in 3 to 5 iterations.

Table 2: Small sample properties of the estimator of the dynamic probit model.

	α		β	
	Mean Bias	RMSE	Mean Bias	RMSE
N=100				
KMD	0.0509	0.3618	-0.1549	0.2335
EKMD	-0.0043	0.4154	-0.1521	0.2564
RMD	0.1736	0.2806	-0.2429	0.3186
N=200				
KMD	0.0081	0.2592	-0.0624	0.1515
EKMD	-0.0889	0.2654	-0.0105	0.1395
RMD	0.1865	0.2746	-0.2349	0.2767
N=400				
KMD	-0.0284	0.1732	0.0035	0.1086
EKMD	-0.0024	0.1444	-0.0177	0.1022
RMD	0.1060	0.1753	-0.1165	0.1687

7.2 Dynamic panel data probit model

For the second simulation exercise, consider the following model:

$$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + f(z_i) + u_{it} > 0\}, \quad i = 1, \cdots, N; \quad t = 1, 2, 3. \tag{7.4}$$

Here, x_{it} and z_i are independently distributed $N(0,1)$. The random shock u_{it} is independently distributed $N(0,\ 0.3 + 0.1 \cdot t)$. The process is initialized with $y_{i0} = 0$, and the individual specific function is given by:

$$f(z_i) = \frac{\exp(z_i)}{1 + \exp(z_i)} - 0.5. \tag{7.5}$$

Again, we perform 100 Monte Carlo replications of the model with three sample sizes N: 100, 200, and 400. We also perform the same Monte Carlo exercise under the assumption that the investigator knows that u_{it} is normally distributed, but incorrectly assumes the distribution $N(0,\ 0.5)$. Finally, the finite dimensional parameters are $(\alpha,\ \beta) = (0.6,\ 0.8)$. The results are presented in Table 2. As in the previous exercise, the first stage unrestricted estimator is denoted by KMD, the second stage unrestricted estimator is denoted by EKMD, and the restricted model is denoted by RMD.

The results in Table 2 indicate that the noise introduced by estimating the weighting matrix results in the EKMD performing worse that KMD in very small samples (N=100). The results also show that the gains in variance reduction from a parametric specification of the index function may outweigh the increased bias for very small samples. However, as the sample size grows, the EKMD performs uniformly better than both the KMD and the RMD. The results of Table 2 again verifies $\sqrt{N}$-convergence of both the KMD and the EKMD. The convergence of the algorithm is slightly slower for this exercise, with the IBA converging typically in 2 to 4 iterations and the OBA converging in 3 to 6 iterations.

As expected all three estimators perform uniformly worse in estimating the dynamic binary choice model than estimating the static model of the first Monte Carlo exercise. However, the results show that the model does perform well in recovering the finite dimensional parameters of interest. These exercises also highlight the potential severity of incorrectly specifying the index function.

8 Conclusion

This paper investigates identification and estimation of a class of single-index panel data models with semiparametric individual-specific effects. This class includes the semiparametric discrete-choice panel models with heteroskedastic errors. The model allows for the inclusion of predetermined variables, as well as lagged dependent variables. A stronger restriction on the individual-specific effects is needed in the latter case. We develop a general minimum distance estimator of the finite and infinite parameters of interest. This estimator extends the minimum distance estimator of Mammen et al. (2001) to the panel data framework and has the advantage that the estimator can be computed with either a sieve or a kernel smoother. In the case where a kernel smoother is chosen, this paper provides a new algorithm to compute the estimators that fully implements the restrictions of the model. The algorithm is an extension of the backfitting algorithm proposed in Buja et al. (1989), Mammen et al. (1999) and Mammen et al. (2001). The full algorithm is composed of an inner backfitting algorithm and an outer backfitting algorithm. Convergence of this algorithm is proved and our experience shows that both the inner and outer backfitting algorithms typically converge within 2-5 iterations. We show that the estimators of the finite dimensional

parameters are $\sqrt{N}$-consistent and asymptotically normal. We show also that the estimators of the infinite dimensional parameters are consistent. We derive the semiparametric efficiency bound for this class of models and show that our estimator indeed achieves this bound. Identification of the model does require that the individual-specific effect be independent of one of the continuous explanatory variables given the other covariates. It may be possible to relax this assumption under alternative restrictions of the form of the individual specific effect. The paper provides a small Monte Carlo exercise that shows that the estimator performs well is small samples. The simulation results verify $\sqrt{N}$-convergence of the finite dimensional parameters and show that the model outperforms other models that miss-specify the index function.

A LEMMA AND THEOREMS

A.1 Proof of Theorem 3.2

Proof. Equations (2.6) and (3.1) imply that

$$\begin{aligned} \varphi_{t0}^{-1}(x_{it}\beta_0 + f_0(z_i)) &= \varphi_{t1}^{-1}(x_{it}\beta_1 + f_1(z_i)) \Leftrightarrow \\ x_{it}\beta_0 + f_0(z_i) &= \varphi_{t0}(\varphi_{t1}^{-1}(x_{it}\beta_1 + f_1(z_i))), \end{aligned} \tag{A.1}$$

Strict monotonicity of the index function implies that it is differentiable almost everywhere. Differentiating equation (A.1) with respect to the continuous regressor x_{itK} gives:

$$a := \frac{\beta_{0k}}{\beta_{1k}} = \frac{\varphi'_{t0}(\varphi_{t1}^{-1}(x_{it}\beta_1 + f_1(z_i)))}{\varphi'_{t1}(\varphi_{t1}^{-1}(x_{it}\beta_1 + f_1(z_i)))} > 0, \tag{A.2}$$

where the positive sign follows from the assumption that the index function is strictly increasing. We have from equation (A.2) that $\varphi'_{t0}(P_{it0}) = a\varphi'_{t1}(P_{it0})$ which implies that:

$$\varphi_{t0}(P_{it0}) = a\varphi_{t1}(P_{it0}) + c_t. \tag{A.3}$$

By taking expectations of (A.3), Assumption 3.1.3 implies that $c_t = 0$. Taking first difference of equations (A.3) gives:

$$\Delta[\varphi_{t0}(P_{it0})] \quad = \quad a\Delta[\varphi_{t1}(P_{it0})]. \tag{A.4}$$

Noting that equation (2.7) also holds for π_1, equation (A.4) implies

$$\begin{aligned} a\Delta[\varphi_{t1}(P_{it0})] &= \Delta x_{it}\beta_0 \\ a\Delta[\varphi_{t1}(P_{it0})] &= a\Delta x_{it}\beta_1. \end{aligned} \tag{A.5}$$

Equating the RHS of the equations in (A.5), pre-multiplying by $\Delta x_{it}'$ and taking expectations gives:

$$E[\Delta x_{it}'\Delta x_{it}]\beta_0 = aE[\Delta x_{it}'\Delta x_{it}]\beta_1. \tag{A.6}$$

Then by the invertibility of $E[\Delta x_{it}'\Delta x_{it}]$ we have

$$\beta_0 = a\beta_1. \tag{A.7}$$

Equation (A.3) gives

$$x_{it}\beta_0 + f_0(z_i) = x_{it}(a\beta_1) + a f_1(z_i). \tag{A.8}$$

Substituting equation (A.7) into equation (A.8) gives

$$f_0(z_i) = a f_1(z_i). \tag{A.9}$$

The assumption that $\|\beta_0\| = \|\beta_1\| = 1$ implies from equation (A.7) that $|a| = 1$. But $a > 0$, which implies that $a = 1$. □

We first state and prove some lemmas that are needed for the existence and uniqueness of the proposed estimator (4.2).

Lemma A.1. *(i) The cartesian product $\mathcal{S}_{\mathcal{K}}^T := \bigotimes_{j=1}^T \mathcal{S}_{\mathcal{K}}$ is compact in the sup-norm topology. (ii) The spaces $\mathcal{F}_c = \{\Delta\varphi \mid \varphi \in \mathcal{S}_{\mathcal{K}}^T\}$ and $\mathcal{F}_c^N = \{(m,\cdots,m)' \mid m \in \mathcal{F}_c\}$, where the vector in the last set has N components, are compact in their respective sup-norm topologies.*

Proof. Given that $\mathcal{S}_{\mathcal{K}}$ is compact, claim (i) follows from Tychonov theorem on the compactness of product spaces. Given that $\mathcal{S}_{\mathcal{K}}^T$ is compact, and the operator Δ is (linear and) continuous in the sup-norm, $\mathcal{F}_c$ is compact and, again by Tychonov theorem, $\mathcal{F}_c^N$ is also compact. □

Lemma A.2. *The objective function in the minimization problem (4.2), $\hat{Q}_N(\theta)$, is a uniformly continuous and strictly convex function of* θ.

Proof. The strict convexity follows from the observations that the function is strictly convex in ρ and ρ is linear in θ. If θ converges uniformly to θ^*, then ρ converges uniformly to ρ^*, where ρ^* is obtained by substituting θ^* for θ in ρ. Hence, the objective function converges uniformly to $\hat{Q}_N(\theta^*)$. □

A.2 Proof of Lemma 4.2

Proof. By Lemma A.1 the set Θ^N is compact in the sup-norm topology. Since the functional $Q_N(\theta)$ is continuous (Lemma A.2), by Weierstrass theorem, it has a maximum and a minimum. Since it is also strictly convex (Lemma A.2), the minimum is unique. □

A.3 Proof of Theorem 5.2

Proof. First note that the projection of an element h of the set $\tilde{\varphi}_2 := \{\varphi_2 \,|\, \varphi_2 \in \mathcal{S}_{\mathcal{K}}\}$ onto the set $a + \tilde{\varphi}_1 := \{a + \varphi_1 \,|\, \varphi_1 \in \mathcal{S}_{\mathcal{K}}\}$ for fixed a is equal to a plus the projection of $h - a$ onto the set $\tilde{\varphi}_1 := \{\varphi_1 \,|\, \varphi_1 \in \mathcal{S}_{\mathcal{K}}\}$. Hence, the backfitting algorithm is indeed a sequence of alternating projections under the norm $\|\cdot\|_2$. Let $\mathcal{T}_{a+\tilde{\varphi}_1}$ and $\mathcal{T}_{\tilde{\varphi}_2}$ denote the projectors onto $a + \tilde{\varphi}_1$ and $\tilde{\varphi}_2$ respectively, as defined by equations (5.2) to (5.4). The restrictions on the kernel K_2 in Assumption 5.1 and the monotonization step (equation (5.4)), along with Proposition 1 of Mammen et al. (2001) ensure that the resulting projections do lie in their respective sets. Then for an arbitrary $f_0 \in a + \tilde{\varphi}_1$, the sequence of alternating projections is given by $Q^n f_0 := \left(\mathcal{T}_{a+\tilde{\varphi}_1} \mathcal{T}_{\tilde{\varphi}_2}\right)^n f_0$. Finally, given the compactness results of Lemma A.1, Theorem 4 of Cheney and Goldstein (1959) shows that the sequence $Q^n f_0$ converges to a fixed point when n tends to infinity. The theorem is reproduced here for convenience.

Theorem 4. *Let K_1 and K_2 be two closed convex sets in Hilbert space and Q the composition $P_l P_2$ of their proximity maps. Convergence of $Q^n x$ to a fixed point of Q is assured when either (a) one set is compact, or (b) one set is finite dimensional and the distance between the sets is attained.* □

A.4 Proof of Theorem 5.3

Proof. Given β and φ_{t-1}, the projection the set $\tilde{\varphi}_t := \{\varphi_t \,|\, \varphi_t \in \mathcal{S}_{\mathcal{K}}\}$ is closed form and is implemented by equation (5.6). Hence, steps 1 and 2 of OBA are projections of $\Delta x\beta$ onto the set $\mathcal{F}_c^N$. Denote the corresponding projector as $\mathcal{T}_{\mathcal{F}_c^N}$. Step 3 of the OBA is clearly a projection of $\Delta\varphi$ onto $x\mathcal{B}$ under its respective norm. Denote the corresponding projector as $\mathcal{T}_{x\mathcal{B}}$. This notation shows that the OBA is indeed sequences of alternating projections under the norm $\|\cdot\|_T$. For an arbitrary $b \in \mathcal{B}$, the sequence of alternating projections is given by $\mathcal{T}^n b := \left(\mathcal{T}_{x\mathcal{B}}\mathcal{T}_{\mathcal{F}_c^N}\right)^n b$. Given compactness of $\mathcal{F}_c^N$ (Lemma A.1) and of $\mathcal{B}$, Theorem 4 of Cheney and Goldstein (1959) shows that the sequence $\mathcal{T}^n b$ converges to a fixed point as n tends to infinity. □

A.5 Proof of Theorem 6.3

In order to prove consistency of the first and second stage estimators, we first state and prove the following auxiliary lemma.

Lemma A.3. *Let W be a positive definite, symmetric with $\|W\| < \infty$. Let $\hat{W} \xrightarrow{p} W$ as $N \to \infty$. Let $\breve{\theta} := (\breve{\beta}, \breve{\varphi})$ minimize the objective function*

$$\tilde{Q}_N := \int N^{-1} \sum_{i=1}^{N} \tau_i \rho(w_i, P, \theta)' \hat{W}^{-1} \rho(w_i, P, \theta) \omega_i(P) dP$$

over the set Θ, and θ_0 minimize $\tilde{Q}_0 := E\left[\tau_i \rho(w_i,\theta)' W^{-1} \rho(w_i,\theta)\right]$ over Θ. Let the assumptions 2.1, 3.1, 5.1, 6.1, and 6.2 be satisfied. Then

$$\breve{\beta} \xrightarrow{p} \beta_0$$
$$\|\breve{\varphi}(P) - \varphi_0(P)\|_{s,2} \xrightarrow{p} 0$$

Proof. To begin, define $\tilde{\omega}_{it}(P_t) := \sigma_2^{-1} K_2(\sigma_2^{-1}(P_{it} - P_t))$, $\tilde{\omega}_i(P) := \prod_{t=1}^{T} \tilde{\omega}_{it}(P_t)$, $\tilde{f}_P(P) := N^{-1}\sum_{i=1}^{N} \tilde{\omega}_i$, $\hat{m}_t(P,\beta) := N^{-1}\sum_{i=1}^{N} \tau_i \Delta x_{it} \beta \hat{\omega}_i(P)/\hat{f}_P(P)$, and $\tilde{m}_t(P,\beta) := N^{-1}\sum_{i=1}^{N} \tau_i \Delta x_{it} \beta \tilde{\omega}_i(P)/\tilde{f}_P(P)$. Define also $\hat{m}(P,\beta) := (\hat{m}_2(P,\beta), \cdots, \hat{m}_T(P,\beta))'$. Define $m_0(P,\beta)$ and $\tilde{m}(P,\beta)$ analogously.

The law of iterated projections (Mammen et al. (2001)) implies that $\breve{\theta}$ also minimizes $Q_N(\theta, \hat{P}) := \int (\hat{m}(P,\beta) - \Delta\varphi(P))' \hat{W}^{-1} (\hat{m}(P,\beta) - \Delta\varphi(P)) \hat{f}(P) \mathrm{d}P$, and that θ_0 also minimizes $Q_0(\theta) := E[(m_0(P,\beta) - \Delta\varphi(P))' W^{-1} (m_0(P,\beta) - \Delta\varphi(P))]$. Define $Q_N(\theta, P) := \int (\tilde{m}(P,\beta) - \Delta\varphi(P))' W^{-1} (\tilde{m}(P,\beta) - \Delta\varphi(P)) \tilde{f}(P) \mathrm{d}P$,

and note that $Q_0(\theta_0) = 0$. We make the following claims:

$$\sup_\Theta |Q_N(\theta, P) - Q_0(\theta)| \xrightarrow{P} 0 \tag{A.10}$$

$$\sup_\Theta |Q_N(\theta, \hat{P}) - Q_N(\theta, P)| \xrightarrow{P} 0 \tag{A.11}$$

$$\breve{\varphi}_t \in S_{\mathcal{K}} \text{ wpa1}, \quad t = 1, \cdots, T \tag{A.12}$$

Proof of claim A.10: Note that for each $\theta \in \Theta$,

$$\begin{aligned} Q_N(\theta, P) &= \int (\tilde{m}(P,\beta) - m_0(P,\beta))' W^{-1} (\tilde{m}(P,\beta) - m_0(P,\beta)) \tilde{f}(P) \mathrm{d}P \\ &+ \int (m_0(P,\beta) - \Delta\varphi(P))' W^{-1} (m_0(P,\beta) - \Delta\varphi(P)) \tilde{f}(P) \mathrm{d}P \\ &+ 2 \int (\tilde{m}(P,\beta) - m_0(P,\beta)) W^{-1} (m_0(P,\beta) - \Delta\varphi(P)) \tilde{f}(P) \mathrm{d}P. \end{aligned} \tag{A.13}$$

Under assumptions 5.1 and 6.2, $\tilde{m}(P,\beta) - m_0(P,\beta)$ and $\tilde{f}(P)$ converges in probability to 0 and $f(P)$ respective, and are both bounded. Thus application of the Lebesgue dominated convergence theorem, the first and third terms of the RHS of equation (A.13) converges in probability to 0, and the second term converges to $Q_0(\theta)$. Since Θ is compact, the convergence is uniform.

Proof of claim A.11: For each $\theta \in \Theta$ we have that

$$\begin{aligned} Q_N(\hat{P}, \theta) - Q_N(P, \theta) &= \int (\hat{m}(P,\beta) - \Delta\varphi(P))' (\hat{W}^{-1} - W^{-1}) (\hat{m}(P,\beta) - \Delta\varphi(P)) \hat{f}(P) \mathrm{d}P \\ &+ \int (\hat{m}(P,\beta) - \tilde{m}(P,\beta))' W^{-1} (\hat{m}(P,\beta) - \tilde{m}(P,\beta)) \hat{f}(P) \mathrm{d}P \\ &+ \int (\tilde{m}(P,\beta) - \Delta\varphi(P))' W^{-1} (\tilde{m}(P,\beta) - \Delta\varphi(P)) (\hat{f}(P) - \tilde{f}(P)) \mathrm{d}P \end{aligned} \tag{A.14}$$

Define $q_{it} := (1 \ \Delta x_{it}\beta)'$. Under assumptions 5.1 , 6.1 and 6.2 we have that $|\frac{1}{N}\sum_i q_{it}(\omega_i(P) - \tilde{\omega}_i(P))| \leq (\frac{1}{N}\sum_i \|q_{it}\|^2)^{1/2} (\frac{1}{N}\sum_i |\omega_i(P) - \tilde{\omega}_i(P)|^2)^{1/2} \leq C(\frac{1}{N}\sum_i \|q_{it}\|^2)^{1/2} (N\sigma_2^{T+1})^{-1} (\sqrt{N}\|\hat{P}_i - P_i\|_{s,2}^2)^{1/2}$ $\to 0$, which implies that $|\hat{m}(P,\beta) - \tilde{m}(P,\beta)|$ and $|\hat{f}(P) - \tilde{f}(P)|$ converge to zero in probability. This, assumptions 5.1 , 6.1 and 6.2 and the Lebesgue dominated convergence theorem imply that the second and third terms of equation (A.14) converge to 0 in probability. Similarly, the first term converges to zero in probability by assumptions 5.1 , and 6.2, by the Lebesgue dominated convergence theorem, and by the consistency of $\hat{W}$ for W. Thus we have that $|Q_N(\hat{P}, \theta) - Q_N(P, \theta)|$ converges to zero in probability for any $\theta \in \Theta$. Since Θ is compact, the convergence is uniform over Θ.

Proof of claim A.12: To prove this claim, it is sufficient to consider the isotonic kernel smoother $\breve{\varphi}_2$. One obtains this monotone function from the unconstrained estimate by replacing parts of the

unconstrained smoother with finite constant pieces (Mammen et al. (2001)). These pieces clearly satisfy the restrictions of $S_{\mathcal{K}}$. Outside these intervals, assumptions 5.1, and 6.2 ensure that the unconstrained smoother satisfies the restrictions of $S_{\mathcal{K}}$ wpa1.

Since $\check{\theta}$ is the minimizer of $Q_N(\theta,\hat{P})$, we have that

$$\begin{aligned} 0 &\leq Q_N(\hat{\theta},\hat{P}) \leq Q_N(\theta_0,\hat{P}) \\ &\leq |Q_N(\theta_0,\hat{P}) - Q_N(\theta_0,\hat{P})| + |Q_N(\theta_0,\hat{P}) - Q_0(\theta_0)| + Q_0(\theta_0) \\ &\leq \sup_{\Theta} |Q_N(\theta_0,\hat{P}) - Q_N(\theta_0,\hat{P})| + \sup_{\Theta} |Q_N(\theta_0,\hat{P}) - Q_0(\theta_0)| + Q_0(\theta_0) \xrightarrow{P} 0, \end{aligned} \tag{A.15}$$

by equations (A.10) and (A.11). Also,

$$\begin{aligned} 0 &\leq Q_0(\check{\theta}) \\ &= Q_N(\check{\theta},P) - Q_N(\check{\theta},\hat{P}) + Q_0(\check{\theta}) - Q_N(\check{\theta},P) + Q_N(\check{\theta},\hat{P}) \\ &\leq \sup_{\Theta} |Q_N(\theta,P) - Q_N(\theta,\hat{P})| + \sup_{\Theta} |Q_0(\theta,P) - Q_N(\theta,P)| + Q_N(\check{\theta},\hat{P}) \xrightarrow{P} 0, \end{aligned} \tag{A.16}$$

by claims (A.10), (A.11), (A.12) and equation (A.15). Since the model is identified, for all $\delta > 0$ there exists $\varepsilon > 0$ such that $d[(\beta,\varphi),(\beta_0,\varphi_0)] > \delta \Rightarrow Q_0(\beta,\varphi) > \varepsilon$, which implies that that $\Pr\{d[(\check{\beta},\check{\varphi}),(\beta_0,\varphi_0)] > \delta\} \leq \Pr\{(Q_0(\check{\beta},\check{\varphi}) > \varepsilon\} \longrightarrow 0$. □

We are now in a position to prove Theorem 6.3.

Proof. By Lemma A.3, and setting $\hat{W} = W = I_{T-1}$, where I_{T-1} is the $T-1$-dimensional identity matrix, we obtain the desired result. □

A.6 Proof of Lemma 6.4

Proof. We have that $\|\Sigma\| \leq E[\|R\varepsilon\varepsilon' R'\|] \leq E\left[\|R\|^2\|\varepsilon\|^2\right] \leq E\left[\|R\|^4\right]^{1/2} E\left[\|\varepsilon\|^4\right]^{1/2} < \infty$, where the last inequality comes from the uniform boundedness of Θ and Assumption 6.2. Defining $\hat{u}_i := \hat{R}(\hat{P}_i)\hat{\varepsilon}_i$ and $u_i := R_i\varepsilon_i$, we have that $\|\sum_{i=1}^N \hat{u}_i\hat{u}_i'/N - E[u_iu_i']\| \leq \|\sum_{i=1}^N \hat{u}_i\hat{u}_i'/N - \sum_{i=1}^N u_iu_i'/N\| + \|\sum_{i=1}^N u_iu_i'/N - E[u_iu_i']\|$. The last term is $o_P(1)$ by the LLN. Also, we have that $\|\sum_{i=1}^N(\hat{u}_i\hat{u}_i' - u_iu_i')/N\| \leq \sum_{i=1}^N \|\hat{u}_i\hat{u}_i' - u_iu_i'\|/N \leq \sum_{i=1}^N \|\hat{u}_i - u_i\|^2/N + 2\sum_{i=1}^N \|u_i\|\|\hat{u}_i - u_i\|/N \leq \sum_{i=1}^N \|\hat{u}_i - u_i\|^2/N + 2\left(\sum_{i=1}^N \|u_i\|^2/N\right)^{1/2}\left(\sum_{i=1}^N \|\hat{u}_i - u_i\|^2/N\right)^{1/2}$. By adding and subtracting $\hat{R}(\hat{P}_i)\varepsilon_i$, and $R(\hat{P}_i)\varepsilon_i$, and by application of the triangle and Cauchy-Schwartz inequalities, we have that $\sum_{i=1}^N \|\hat{u}_i - u_i\|^2/N \leq \sum_{i=1}^N(\|\hat{R}(\hat{P}_i)\|\|\hat{P}_i - P_i\| + \|\hat{R}(\hat{P}_i) - R(\hat{P}_i)\|\|\varepsilon_i\| + \|R(\hat{P}_i) - R(P_i)\|\|\varepsilon_i\|)^2/N \leq C_1\sum_{i=1}^N \|\hat{P}_i - P_i\|^2/N +$

$\sup_P \|\hat{R}(P) - R(P)\| \sum_{i=1}^N \|\varepsilon_i\|^2/N + C_2(\sum_{i=1}^N \|\hat{P}_i - P_i\|^4/N)^{1/2}(\sum_{i=1}^N \|\varepsilon_i\|^4/N)^{1/2} + 2C_1 \sup_P \|\hat{R}(P) - R(P)\|(\sum_{i=1}^N (\|\hat{P}_i - P_i\|^2/N)^{1/2}(\sum_{i=1}^N \|\varepsilon_i\|^2/N)^{1/2} + C_1C_2(\sum_{i=1}^N (\|\hat{P}_i - P_i\|^4/N)^{1/2}\ (\sum_{i=1}^N \|\varepsilon_i\|^2/N)^{1/2}$, where the constant C_1 comes from the uniform boundedness of $\hat{R}$ and C_2 comes from the uniform Lipschitz condition. Assumptions 5.1, 6.1.1, and Theorem 6.3 imply that all the terms on the RHS of the last inequality converge in probability to zero. Thus $\|\sum_{i=1}^N \hat{u}_i\hat{u}_i'/N - \sum_{i=1}^N u_iu_i'/N\| = o_P(1)$. Furthermore, $\sum_{i=1}^N \|u_i\|^2/N \leq C\sum_{i=1}^N \|\varepsilon_i\|^2/N = O_p(1)$ by Assumption 6.1.1. Thus $\|\sum_{i=1}^N \hat{u}_i\hat{u}_i'/N - E[u_iu_i']\| = o_P(1)$. □

A.7 Proof of Theorem 6.5

Proof. Set $\hat{W} = \hat{\Sigma}$ and $W = \Sigma$. Then application of Lemmas 6.4 and A.3 obtains the desired result. □

Lemma A.4. *Consider the problem of Lemma A.3. Then under the conditions of Lemma A.3*

$$\sqrt{N}(\breve{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where

$$V := \left[E[\tau_i h_{i0}' W^{-1} h_{i0}]\right]^{-1} E\left[\tau_i h_{i0}' W^{-1}\Sigma W^{-1} h_{i0}\right]\left[E[\tau_i h_{i0}' W^{-1} h_{i0}]\right]^{-1}.$$

Proof. Note that the backfitting algorithm works by iteratively solving for φ given a fixed β, and then solving for β. Thus we have that

$$g(w_i, \beta, \varphi) = \frac{\partial}{\partial \beta} Q_0(w_i, \beta, \varphi)$$

and

$$\varphi = \arg \max_{m \in \mathcal{F}_c(\beta)} E[Q(x_i, \beta, m)].$$

The notation $\mathcal{F}_c(\beta)$ makes it explicit that the resulting estimator $\varphi = \varphi(\cdot; \beta)$ is a function of beta. Proposition 2 of Newey (1994a) therefore implies that the estimation of φ can be ignored in calculating the asymptotic distribution of $\breve{\beta}$. Therefore, in what follows, we ignore the estimation of φ in the calculation of the asymptotic distribution of $\breve{\beta}$.

Define $h_{i0}(\hat{P}_i; \beta) := \partial \Delta\varphi_0(\hat{P}_i; \beta)/\partial\beta - \Delta x_i$. Theorem 3.2 implies that for any solution β to (4.2), $a\beta$ is also a solution, including where $a = \|\beta\|^{-1}$. By construction, $\breve{\beta} = \beta^*/\|\beta^*\| \in \text{int}(\mathcal{B}_1)$. Taking

a mean value expansion of $g(w_i, \breve{\beta}, \varphi_0(\hat{P}_i; \breve{\beta}))$ obtains

$$\begin{aligned}\sqrt{N}(\breve{\beta} - \beta_0) &= -\left[\hat{M}_1(\bar{\beta}, W) + \hat{M}_2(\bar{\beta}, W)\right]^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau_i h_{i0}(\hat{P}_i)' W (\Delta x_i' \beta_0 - \Delta[\varphi_0(\hat{P}_i)])\right], \\ &= \left[\hat{M}_1(\bar{\beta}, W) + \hat{M}_2(\bar{\beta}, W)\right]^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau_i h_{i0}(\hat{P}_i)' W (\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})])\right],\end{aligned}$$

where

$$\begin{aligned}\hat{M}_1(\bar{\beta}, W) &:= \frac{1}{N} \sum_{i=1}^{N} \tau_i h_{i0}(\hat{P}_i; \bar{\beta})' W h_{i0}(\hat{P}_i; \bar{\beta}), \\ \hat{M}_2(\bar{\beta}, W) &:= \frac{1}{N} \sum_{i=1}^{N} \tau_i \left[(\Delta[\varphi_0(\hat{P}_i; \bar{\beta})] - \Delta x_i \bar{\beta}) W^{-1} \otimes I_{T-1}\right] \left[\frac{\partial}{\partial \beta} h_0(\hat{P}_i; \bar{\beta})\right],\end{aligned}$$

$\bar{\beta} \in (\beta_0, \breve{\beta})$, and $\otimes$ denotes the Kroneker product. The inverse term on the RHS exists with probability one because W is positive definite, and Δx and $\frac{\partial}{\partial \beta} h_0(\hat{P}_i; \beta)$ has full rank. Define $g(x_i, \hat{\gamma}_i) := \tau_i h_{i0}(\hat{P}_i)' W (\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})])$. The rest of this section of the proof involves checking conditions (i)-(iv) of Theorem 8.11 of Newey and McFadden (1994). Notice that $g(x_i, P_{i0}) = 0$ implying that $E[g(x_i, \gamma_{i0})] = 0$ and $E[\|g(x_i, \gamma_{i0})\|^2] = 0$. Linearizing $g(x_i, \hat{\gamma}_i)$ around (γ_{i0}) gives $D(w_i, \hat{\gamma} - \gamma_0) := \tau_i h_{i0}(P_{i0}) W R_i \underline{f}^{-1}(w_i) G_i [\hat{\gamma}(w_i) - \gamma_0(w_i)]$, where

$$\begin{aligned}\underline{f}^{-1}(w_i) &:= \operatorname{diag}\left(f^{-1}(w_{it}), t = 1, \cdots, T\right) \\ G_i &:= \operatorname{diag}((-P_{it0}\ 1), t = 1, \cdots, T) \\ \gamma_0(w_i) &:= (\gamma_{10}(w_{i1}), \gamma_{20}(w_{i1}), \cdots, \gamma_{10}(w_{iT}), \gamma_{20}(w_{iT})) \\ \hat{\gamma}(w_i) &:= (\hat{\gamma}_1(w_{i1}), \hat{\gamma}_2(w_{i1}), \cdots, \hat{\gamma}_1(w_{iT}), \hat{\gamma}_2(w_{iT}))'.\end{aligned}$$

Conditions (i) and (ii) of Theorem 8.11 ofNewey and McFadden (1994) are satisfied by noting that boundedness of Δx_i, of γ_0 and its first two derivatives of $\mathcal{K}$, and of W gives $\|g(x_i, \hat{\gamma}_i) - D(w_i, \hat{\gamma} - \gamma_0)\| \leq b(w) \|\hat{\gamma}(w_i) - \gamma_0(w_i)\|^2$, with $E[b(w)] < \infty$, and $D(w, \gamma) = \tau_i h_{i0}(P_{i0})' W R_i \underline{f}^{-1}(w_i) G_i \gamma \leq c(w) \|\gamma\|$ with $E[c(w)^2] \leq \infty$. Condition (iii) is also immediately satisfied by observing that $\int D(w, \gamma) f_w(w) \mathrm{d}w = \int \nu(w) \gamma(w) \mathrm{d}w$, where $\nu(w_i) := \tau_i h_{i0}(P_{i0})' W R_i G_i$. Given continuity of $\nu(w)$ on $\mathcal{W}$ and assumption 6.1.1, verification of conditions (iv) of Theorem 8.11 of Newey and McFadden (1994) is given in the proof thereof. Therefore, by Theorem 8.11 of Newey and McFadden (1994)

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} g(w_i, \hat{\gamma}) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = \text{Var}(\delta(w))$, and $\delta(w) = \nu(w)q - E[\nu(w)q]$. Application of the law of iterated expectations show that $E[\nu(w)q] = 0$. Also, straightforward calculations show that $\nu(w_i)q_i = \tau_i h_{i0}(P_{i0})I_i W R_i \varepsilon_i$, where $\varepsilon_i = (y_i - P_{i0})$. Therefore, $\Omega = E[\tau_i h_{i0}(P_{i0})' W R_i \varepsilon_i \varepsilon_i' R_i' W h_{i0}(P_{i0})]$. By assumption $E[\tau_i R_i \varepsilon_i \varepsilon_i' R_i' | w] = \Sigma(w) = \Sigma$. By Applying the law of iterated expectations, we have that $\Omega = E[\tau_i h_{i0}(P_{i0})' W \Sigma W h_{i0}(P_{i0})]$. Defining $M(W) := E[h_{i0}' W h_{i0}]$, straightforward calculations show that

$$\begin{aligned}
\|\hat{M}_1(\bar{\beta}, W) - M(W)\| \quad \leq \quad & \|\tfrac{1}{N}\textstyle\sum_i \tau_i h_{i0}' W h_{i0} - E[h_{i0}' W h_{i0}]\| \\
+ \quad & \tfrac{2}{N}\textstyle\sum_i \|h_{i0}\| \|W\| \|h_{i0}(\hat{P}_i; \bar{\beta}) - h_{i0}(\hat{P}_i; \beta_0)\| \\
+ \quad & \tfrac{2}{N}\textstyle\sum_i \|h_{i0}\| \|W\| \|h_{i0}(\hat{P}_i; \beta_0)) - h_{i0}\| \\
+ \quad & \tfrac{1}{N}\textstyle\sum_i \|h_{i0}\| \|W\| \|h_{i0}(\hat{P}_i; \bar{\beta}) - h_{i0}(\hat{P}_i; \beta_0)\| \|h_{i0}(\hat{P}_i; \beta_0)) - h_{i0}\| \\
+ \quad & \tfrac{2}{N}\textstyle\sum_i \|h_{i0}\| \|W\| \|h_{i0}(\hat{P}_i; \bar{\beta}) - h_{i0}(\hat{P}_i; \beta_0)\|^2 \\
+ \quad & \tfrac{2}{N}\textstyle\sum_i \|h_{i0}\| \|W\| \|h_{i0}(\hat{P}_i; \beta_0)) - h_{i0}\|^2
\end{aligned} \tag{A.17}$$

By the LLN, the first term on the RHS of equation (A.17) is $o_P(1)$. Note that

$$\|h_{i0}(\hat{P}_i; \bar{\beta}) - h_{i0}(\hat{P}_i; \beta_0)\| = \|\frac{\partial}{\partial \beta}\varphi_0(\hat{P}_i; \bar{\beta}) - \frac{\partial}{\partial \beta}\varphi_0(\hat{P}_i; \beta_0)\|,$$

which is $o_P(1)$ by the continuous mapping theorem and by the consistency of $\bar{\beta}$ for β_0. This and the boundedness conditions on h_{i0} and W imply that the second and fifth terms on the RHS of equation (A.17) are $o_P(1)$. Furthermore,

$$\|h_{i0}(\hat{P}_i; \beta_0)) - h_{i0}\| = \|\frac{\partial}{\partial \beta}\varphi_0(\hat{P}_i) - \frac{\partial}{\partial \beta}\varphi_0(P_{i0})\|,$$

which is $o_P(1)$ by the same conditions. This, along with the above convergence and boundedness conditions imply that the third, fourth, and sixth terms on the RHS of equation (A.17) are all $o_P(1)$. We thus have that

$$\hat{M}_1(\bar{\beta}, W) = E\tau_i[h_{i0}' W h_{i0}] + o_P(1).$$

Note also that

$$\hat{M}_2(\bar{\beta}, W) \leq \frac{1}{N}\sum_i \tau_i \|\frac{\partial}{\partial \beta}\varphi_0(\hat{P}_i; \bar{\beta})\| \|W\| \|I\| \|\Delta\varphi_0(\hat{P}_i; \bar{\beta}) - \Delta x_i \bar{\beta}\| = o_P(1)$$

by the consistency theorem, and the boundedness conditions on W and $\frac{\partial}{\partial \beta}\varphi_0(\hat{P}_i; \bar{\beta})$. Thus we have that $\hat{M}_1(\bar{\beta}, W) + \hat{M}_2(\bar{\beta}, W) = M(W) + o_P(1)$.

Since $\sum_{i=1}^{N} g(w_i,\hat{\gamma})/\sqrt{N} = O_P(1)$, the Slutsky theorem gives

$$\sqrt{N}(\check{\beta}-\beta_0) \xrightarrow{d} N(0, M(W)^{-1}\Omega M(W)^{-1}).$$

□

A.8 Proof of Theorem 6.6

Proof. Setting $W = I$ in Lemma A.4 obtains

$$\sqrt{N}(\bar{\beta}-\beta_0) \xrightarrow{d} N(0,V_1),$$

where $V_1 := [E[\tau_i h_{i0}' h_{i0}]]^{-1} E[\tau_i h_{i0}(P_{i0})'\Sigma^{-1} h_{i0}(P_{i0})] [E[\tau_i h_{i0}' h_{i0}]]^{-1}$. Setting $W = \Sigma^{-1}$ and expanding around β_0 obtains

$$\sqrt{N}(\tilde{\beta}-\beta_0) = \left[\hat{M}_1(\bar{\beta},\Sigma^{-1}) + \hat{M}_2(\bar{\beta},\Sigma^{-1})\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tau_i h_{i0}(\hat{P}_i)'\Sigma^{-1}(\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})])\right],$$

with $\sqrt{N}(\tilde{\beta}-\beta_0) \xrightarrow{d} N(0,V_2)$ by Lemma A.4). Setting $W = \hat{\Sigma}^{-1}$ and expanding around β_0 obtains

$$\sqrt{N}(\hat{\beta}-\beta_0) = \left[\hat{M}_1(\bar{\beta},\hat{\Sigma}^{-1}) + \hat{M}_2(\bar{\beta},\hat{\Sigma}^{-1})\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tau_i h_{i0}(\hat{P}_i)'\hat{\Sigma}^{-1}(\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})])\right].$$

Note that $\hat{M}_1(\bar{\beta},\hat{\Sigma}^{-1}) - \hat{M}_1(\bar{\beta},\Sigma^{-1}) \leq N^{-1}\sum_{i=1}^{N}\tau_i \|h_{i0}(\hat{P}_i,\bar{\beta})\|^2 \|\hat{\Sigma}^{-1}-\Sigma^{-1}\| \leq N^{-1}\sum_{i=1}^{N} C\|\hat{\Sigma}^{-1} - \Sigma^{-1}\| = o_P(1)$, and $\hat{M}_2(\bar{\beta},\hat{\Sigma}^{-1}) - \hat{M}_2(\bar{\beta},\Sigma^{-1}) \leq N^{-1}\sum_i^N \tau_i \|\frac{\partial}{\partial\beta}\varphi_0(\hat{P}_i;\bar{\beta})\| \|I\| \|\Delta\varphi_0(\hat{P}_i;\bar{\beta}) - \Delta x_i\bar{\beta}\| \left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\| \leq C\|\hat{\Sigma}^{-1}-\Sigma^{-1}\| = o_P(1)$. Also,

$$\begin{gathered}\operatorname{vec}\left(N^{-1/2}\textstyle\sum_{i=1}^{N}\tau_i h_{i0}(\hat{P}_i)'(\hat{\Sigma}^{-1}-\Sigma^{-1})(\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})])\right) = \\ \left(N^{-1/2}\textstyle\sum_{i=1}^{N}\tau_i[(\Delta[\varphi_0(\hat{P}_i)] - \Delta[\varphi_0(P_{i0})]) \otimes h_{i0}(\hat{P}_i)]'\right)\left(\operatorname{vec}[\hat{\Sigma}^{-1}-\Sigma^{-1}]\right).\end{gathered}$$

Asymptotic normality of the first stage estimator implies that the first term on the RHS of the equality in parenthesis is $O_p(1)$. The second term on the RHS of the equality in parenthesis is $o_P(1)$ by Lemma 6.4. Thus Slutsky's theorem implies that $\sqrt{N}(\hat{\beta}-\tilde{\beta}) = o_P(1)$, which obtains

$$\sqrt{N}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V_2).$$

□

A.9 Proof of Theorem 6.7

Proof. The proof of efficiency uses the results developed in Newey (1994a). To proceed, we first set up the environment so that the results are directly applicable.

As noted in section 5, the objective function in (4.1) can be written by concentrating out φ, and writing φ as a function of β, $\varphi(P_{i0};\beta)$. The derivative of (4.1) with respect to β is then given by

$$E_Q\left[\tau_i h_i(P_{i0};\beta)[\Sigma(w)]^{-1}[\Delta[\varphi_0(P_{i0};\beta)]-\Delta x_i\beta_0\right]=E_Q[m(x_i,\beta,\varphi,P_{i0})]=0,$$

where $h_i := \frac{\partial}{\partial\beta}\Delta\varphi(P_{i0};\beta)-x_i$. Furthermore, the limit of our estimate $\hat{\varphi}$ maximizes $E_Q[S(x_i,\beta,\varphi,P_i)]$. Thus by proposition 2 of Newey (1994a), the estimation of φ can be ignored in calculating the asymptotic variance. So we work only with $\varphi(P;\beta)=\varphi_0(P;\beta)$.

Let the distribution Q belong to a general family of distributions $\mathcal{Q}$. Define the parametric submodel $\mathcal{Q}(\eta) := \{Q_\eta : Q_\eta \in \mathcal{Q},\ Q_\eta = Q_0 \text{ at } \eta = 0\}$. We assume f_η to be a probability density relative to a fixed measure ν, the map $\eta \mapsto \sqrt{f_\eta(w)}$ is continuously differentiable in a neighborhood of 0, and $\eta \mapsto \int\left[(\partial f_\eta/\partial\eta)^2/f_\eta\right]d\nu$ is finite and continuous in this neighborhood. Then by Lemma 1.9 of van der Vaart (1998), $\eta \mapsto Q_\eta$ is a differentiable path. We use this differentiable path to induce parametric submodels for the parameters that $\hat{\beta}$ and $\hat{P}_i$ are estimating. That is, we define $\mu(\eta)=\mu(Q_\eta) := \text{plim}\,\hat{\beta}$ and $P_i(\eta)=P_i(Q_\eta) := \text{plim}\,\hat{P}_i$, where $\mu(Q_\eta)$ satisfies:

$$E_\eta[m(x,\mu,P(\eta))]=0 \tag{A.18}$$

The rest of the proof involves finding the pathwise derivative $d(w)$ satisfying $\frac{\partial\mu(\eta)}{\partial\eta}=E[d(w)g(w)]$, where $g(w) := \frac{\partial}{\partial\eta|_{\eta=0}}\ln f_\eta(w)$ is the corresponding score. Then the variance bound for the estimation of $\mu(\eta)$ is $\text{Var}(d(w))$. Differentiating equation (A.18) with respect to η and solving for $\frac{\partial\mu(\eta)}{\partial\eta}$ gives

$$\frac{\partial\mu(\eta)}{\partial\eta} = -M^{-1}\left\{E\left[\frac{\partial}{\partial P}\tilde{m}(x,\beta_0,P(\eta))\frac{\partial P(w,\eta)}{\partial\eta}\right]+\frac{\partial}{\partial\eta}E_\eta\left[\tilde{m}(x,\beta_0,P_0)\right]\right\}, \tag{A.19}$$

where $M := \frac{\partial}{\partial\beta}E[m(x,\beta_0P_0)]=E[h'_{i0}\{\Sigma(w)\}^{-1}h_{i0}]$, which is invertible by assumption (3.1.3). From equation (A.18), the last term on the RHS of equation (A.19) is zero. Defining $\delta(x) := \frac{\partial}{\partial P}m(x,\beta_0,P(\eta))$

and applying the law of iterated expectations to $P(w,\eta) = E[y|w]$ gives

$$\begin{aligned} \frac{\partial \mu(\eta)}{\partial \eta} &= -M^{-1}\left\{\frac{\partial}{\partial \eta} E_\eta[\delta(w)(y - P_0(w))]\right\} \\ &= [-(M^{-1}\delta(w)(y - P_0))S(w)] \end{aligned} \tag{A.20}$$

Thus giving $d(w) = -M^{-1}\delta(w)(y - P_0)$. Noting that $\delta(w_i) = \tau_i h'_{i0}\{\Sigma(w)\}^{-1}R_i$, we have that

$$\text{Var}(d(w)) = E\left[\tau_i h'_{i0}\{\Sigma(w)\}^{-1}h_{i0}\right]^{-1} E\left[\tau_i h_{i0}\{\Sigma(w)\}^{-1}R\Omega R'\{\Sigma(w)\}^{-1}h_{i0}\right] E\left[\tau_i h'_{i0}\{\Sigma(w)\}^{-1}h_{i0}\right]^{-1},$$

where $\Omega = E[(y - P_0)(y - P_0)'|w]$. Note that $R\Omega R' = E[R(y - P_0)(y - P_0)'R'|w] = \Sigma(w)$. This gives

$$\text{Var}(d(w)) = E\left[\tau_i h'_{i0}\{\Sigma(w)\}^{-1}h_{i0}\right]^{-1}.$$

Finally, the assumption that $\Sigma(w) = \Sigma$ obtains the asymptotic variance of $\hat{\beta}$ derived in theorem 6.6. □

References

AI, C. AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1975–1843.

ALTUG, S. AND R. A. MILLER (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.

——— (1998): "Effect of Work Experience of Female Wages and Labour Supply," *The Review of Economic Studies*, 65, 45–85.

ARELLANO, M. AND R. CARRASCO (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics*, 115, 125–157.

BAUSCHKE, H. AND J. BORWEIN (1996): "On Optimal Algorithms for Solving Convex Feasibility Problems," *SIAM Review*, 38, 367–426.

BRUNK, H. D. (1958): "On the Estimation of Parameters Restricted by Inequalities," *Annals of Mathematical Statistics*, 29, 437–454.

BUJA, A., T. HASTIE, AND R. TIBSHIRANI (1989): "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–555.

CHAMBERLAIN, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economics Studies*, XLVII, 225–238.

——— (1993): "Feedback in Panel Data Models," .

CHEN, H. (1995): "Asymptotically Efficient Estimation in Semiparametric Generalized Linear Models," *The Annals of Statistics*, 23, 1102–1129.

CHEN, S. (1998): "Root-N Consistent Estimation of a Panel Data Sample Selection Model," *MIMEO: The Hong Kong University of Science and Technology*.

CHEN, X. (2007): *Large Sample Sieve Estimation of Semi-Nonparametric Models*, Elsevier Science Publishers B.V.

CHENEY, W. AND A. GOLDSTEIN (1959): "Proximitymaps for convex sets," *Proceedings of the American Mathematical Society*, 10, 448–450.

DELFOUR, M. AND J.-P. SOLESIO (1987): "Design and control sensitivity analysis via min max differentiability," *Proceedings of the 26th Conference on Design and Control*, 26, 987–991.

DEUTSCH, F. (2001): *Best Approximation in Inner Product Spaces*, Springer.

GAYLE, G. AND C. VIAUROUX (2007): "Root-N Consistent Semiparametric Estimators of a Dynamic Panel Data Sample Selection Model," *Journal of Econometrics*, 141, 179–212.

GAYLE, W. (2008): "Root-N Consistent Estimation of a Dynamic Panel Data Model," *MIMEO: Univeristy of Virginia*.

HARDLE, W., M. MULLER, S. SPERLICH, AND A. WERWATZ (2004): *Nonparametric and Semiparametric Models*, Springer.

HASTIE, T. AND R. TIBSHIRANI (1986): "Generalized Additive Models," *Statistical Science*, 1, 297–318.

HONORÉ, B. AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

HONORÉ, B. AND A. LEWBELL (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063.

HONORE, B. AND E. TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 661–629.

HOROWITZ, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

MAMMEN, E., O. LINTON, AND J. NIELSEN (1999): "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *The Annals of Statistics*, 27, 1443–1490.

MAMMEN, E., J. MARRON, B. TURLACH, AND M. WAND (2001): "A General Projection Framework for Constrained Smoothing," *Statistical Science*, 16, 232–248.

MANSKI, C. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–33.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

MOOD, A., F. GRAYBILL, AND D. BOES (1974): *Introduction to the Theory of Statistics*, McGraw Hill.

MUNDLAK, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

NEWEY, W. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1994b): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

NEWEY, W. K. AND D. MCFADDEN (1994): *Large Sample Estimation and Hypothesis Testing*, Elsevier Science Publishers.

NEWEY, W. K. AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578, available at http://ideas.repec.org/a/ecm/emetrp/v71y2003i5p1565-1578.html.

PAGAN, A. AND A. ULLAH (1999): "Review: Nonparametric Econometrics by A. Pagan and A. Ullah, Cambridge University Press: Cambridge, UK. xviii+424 pages. 1999," .

VAN DER VAART, A. (1998): "Semiparametric Statistics," *Lecture notes*.