Estimation of a Nonlinear Panel Data Model with Semiparametric Individual Effects

Wayne-Roy Gayle[a,*], Soiliou Daw Namoro[b]

[a] *Department of Economics, University of Virginia, Monroe Hall, McCormick Rd, Room 208A, Charlottesville, VA 22903*

[b] *Department of Economics, University of Pittsburgh, 230 S. Bouquet St., Pittsburgh, PA 15260*

Abstract

This paper investigates identification and estimation of a class of nonlinear panel data, single-index models. The model allows for unknown time-specific link functions, and semiparametric specification of the individual-specific effects. We develop an estimator for the parameters of interest, and propose a powerful new kernel-based modified backfitting algorithm to compute the estimator. We derive uniform rates of convergence results for the estimators of the link functions, and show the estimators of the finite-dimensional parameters are root-N consistent with a Gaussian limiting distribution. We study the small sample properties of the estimator via Monte Carlo techniques.

JEL *classification*: C13, C14, C23

Keywords: Semiparametric estimation, modified backfitting, panel data, nonlinear models.

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

*Corresponding author. Tel.: +434-243-4336: fax: +434-982-2904; E-mail: wg4b@virginia.edu

1. Introduction

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

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

where x_{it} is a K-dimensional row vector of random variables, z_i is an L-dimensional row vector of time-constant random variables, ε_{it} is an individual-and time-specific idiosyncratic shock that is assumed to be mean independent of the other explanatory variables, β is a K-dimensional vector of parameters, Φ_t is a strictly increasing and smooth unknown link function, and η is an unknown function. The parameters of primary interest are β, and $\Phi := \{\Phi_t,\ t = 1,\ \cdots,\ T\}$.

We propose a powerful new kernel-based algorithm to compute the estimator for the parameters of interest. The algorithm combines the profile likelihood approach of Severini and Wong (1992) with the backfitting algorithms of Buja et al. (1989), Mammen et al. (1999), and Mammen et al. (2001), and extends them to the present framework. The algorithm fully implements the identification restrictions of the model. We provide sufficient conditions under which the algorithm converges. Also, we derive uniform rates of convergence results for the estimators of the link functions, and show the resulting estimator of β is $\sqrt{N}$-consistent with a Gaussian limiting distribution. Furthermore, estimation of the finite-dimensional parameters is adaptive with respect to estimation of the link functions.

The model presented in equation (1.1) is a panel data version of the generalized partial-linear model (GPLM) with unknown link functions. Other methods can be used to estimate the parameters of the model, including the backfitting estimator developed in Opsomer

(2000), among others, and the methods of series and sieve minimum distance estimation developed in Newey (1994a), Newey and Powell (2003), Ai and Chen (2003), Chen (2007), and Gayle and Viauroux (2007). However, the method developed in this paper has some key advantages over these alternatives.

Opsomer (2000) develops a backfitting procedure to estimate the parameters of additive and partial linear models. This procedure can be modified to the panel data framework. However, it is unclear how to impose shape constraints on the estimators of the infinite-dimensional parameters in an internally consistent way using the method developed in Opsomer (2000). Indeed, a maintained assumption for identification of the parameters of interest in equation (1.1) is that the link functions are strictly monotonic. On the other hand, the modified backfitting algorithm developed in Mammen et al. (1999) and Mammen et al. (2001) accommodates shape restrictions on the infinite-dimensional parameters under the same empirical norm as the one constructed to define estimators of all the parameters. Another key advantage of the algorithm developed in Mammen et al. (1999) and Mammen et al. (2001) is that its convergence is well understood, and does not depend on initial values.

Alternatively, estimating the parameters of interest by implementing the methods of series or sieve estimation developed in Newey (1994a), Newey and Powell (2003), Ai and Chen (2003), Chen (2007), and Gayle and Viauroux (2007) is feasible. However, these methods rely on the choice of smoothers used to compute the estimators of the infinite-dimensional parameters. The estimator developed in this paper can be computed using a wide variety of smoothers.[1] We focus on the case where the smoothers are kernels. To the best of our

[1] See Mammen et al. (2001) for discussions on the implementation of the Nadaraya-Watson smoother and series smoothers.

knowledge, no existing studies investigate kernel-based estimation of panel data GPLM models such as equation (1.1) with shape constraints on the unknown link functions.

The model 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.

The estimator we propose in this paper treats both Φ and η as unknown functions. The models Chamberlain (1980) propose assume the link functions are known, and that η is known up to a set of finite-dimensional parameters. Newey (1994a) extends this framework to allow for η to be an unknown function, while maintaining the parametric specification of the link functions. We extend the model presented in Newey (1994a) to allow for unspecified time-specific link functions. In the discrete-choice framework, Chen (1998) modifies the framework of Newey (1994a) by relaxing the parametric specification of the link functions at the cost of increased restrictions on the finite-dimensional parameters. In this paper, we achieve identification and estimation without these additional restrictions on the finite-dimensional parameters. We restrict identification of the parameters of interest to the static panel data framework. Arellano and Carrasco (2003) develop a panel data discrete-choice model that allows for predetermined explanatory variables. However, the model Arellano and Carrasco (2003) present assume the link functions are known.

Other important developments in the semiparametric panel data literature include Manski (1987) and Honoré and Lewbel (2002). Compared with Manski (1987), this paper imposes stronger restrictions on the joint distribution of the individual effects and observed

regressors, but allows for the errors to be heteroskedastic over time. Honoré and Lewbel (2002) impose a different conditional independence assumption on the distribution of the individual-specific effects and the error term given the observables, which is neither weaker nor stronger than the one we impose in this paper. Also, Honoré and Lewbel (2002) imposes stronger support conditions on the observable and unobservable explanatory variables.

In the next section, we provide an example of how equation (1.1) is derived from the familiar binary-choice, single-index panel data model. However, 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 method we develop in this paper.

We investigate the small sample performance of the proposed estimator in two environments by Monte Carlo analysis. The first exercise examines the performance of the estimator in a static, panel data discrete-choice model, and the second in a static, panel data continuous-outcome model. The simulation exercises show the estimator performs well in small samples.

We organize the rest of paper as follows. Section 2 motivates equation (1.1) by describing how it is derived from various econometric models. Section 3 discusses identification, and section 4 presents the estimator. Section 5 presents the algorithm used to compute the estimator. Section 6 derives the large sample properties of the estimator. Section 7 proposes an estimator for the asymptotic variance of the finite-dimensional parameters. Section 8 is devoted to the Monte Carlo simulations, and section 9 concludes. All proofs and auxiliary lemmas are in the appendix.

2. The Model

In this section, we discuss how equation (1.1) may be derived from more primitive econometric models. Consider the following panel data, single-index model for a unit of observation, i:

$$y_{it} = F_t(x_{it}\beta + \mu_i) + r_{it}, \quad t = 1, \cdots, T, \tag{2.1}$$

where y_{it} is the dependent variable, x_{it} are observable time-varying explanatory variables, μ_i is the time-invariant unobserved effect, F_t is an unknown, strictly increasing function, and r_{it} is the idiosyncratic error where $E[r_{it}|x_{it}, \mu_i] = 0,\ t = 1, \cdots, T$. It is well known that equation (2.1) can be derived from the the following panel data, single-index discrete-choice model:

$$y_{it} = 1\{x_{it}\beta + \mu_i - u_{it} \geq 0\}, \quad t = 1, \cdots, T, \tag{2.2}$$

where x_{it} and μ_i are as described, u_{it} is independent of x_{it} and μ_i with an unknown time-specific distribution function F_t that is absolutely continuous with respect to a Lebesgue measure.

Suppose that for each unit of observation, a vector of observable time-invariant explanatory variables, z_i, exists, and assume the individual-specific effects can be decomposed as follows: $\mu_i = \eta(z_i) + v_i$. Assume that v_i is independent of z_i and $x_i := (x_{i1}, \cdots, x_{iT})$ with distribution that is absolutely continuous with respect to a Lebesgue measure, and has a Radon-Nikodym derivative f_v. Under these assumptions, taking conditional expectations of y_{it} conditional on (x_{it}, z_i, v_i) in equation (2.1) gives

$$E[y_{it}|x_{it}, z_i, v_i] = F_t(x_{it}\beta + \mu_i) = F_t(x_{it}\beta + \eta(z_i) + v_i).$$

Equation (2.1) is therefore obtained by defining $r_{it} = y_{it} - E[y_{it}|x_{it}, z_i, v_i]$. Furthermore, by the law of iterated expectations,

$$E[y_{it}|x_{it}, z_i] = \Phi_t(x_{it}\beta + \eta(z_i)),$$

where $\Phi_t(a) := \int F_t(a+v) f_v(v) dv$. By defining $\varepsilon_{it} = y_{it} - E[y_{it}|x_i, z_i]$, we obtain equation (1.1), where Φ_t inherits the monotonicity constraint on F_t. Because we estimate Φ_t, and not F_t, any predictions made using these estimates should be interpreted as predictions made after integrating out the "pure" random effects component, v_i. Note that F_t is not needed to obtain average partial effects, because these effects can be computed from Φ_t, β, and η. This discussion shows 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.

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

$$P_{it} := P(w_{it}) := E[y_{it}|w_{it}] = \Phi_t(x_{it}\beta + \eta(z_i)), \quad t = 1, \cdots, T. \tag{2.3}$$

The following assumption formalizes the monotonicity constraint on the link function we will maintain in this paper.

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

Define the inverse link function $\varphi_t = \Phi_t^{-1}$, which exists under Assumption 2.1. Equation (2.3) implies

$$\varphi_t(P_{it}) = x_{it}\beta + \eta(z_i), \quad t = 1, \cdots T, \tag{2.4}$$

which in turn implies

$$\Delta\varphi_t(P_{it}) = \Delta x_{it}\beta, \quad t = 2, \cdots T, \tag{2.5}$$

where $\Delta\varphi_t(P_{it}) := \varphi_t(P_{it}) - \varphi_{t-1}(P_{it-1})$ and $\Delta x_{it}\beta := (x_{it}\beta - x_{it-1}\beta)$. The time-invariant function $\eta(z_i)$ is eliminated by the first differencing of equation (2.4). Because $\eta(z_i)$ is not estimated jointly with the other parameters of the model, the computational cost due to the potentially large dimension of z_i is incurred only once in the estimation of P_{it}, $t = 1, \cdots, T$.

3. Identification

Let $\varphi := (\varphi_1, \cdots, \varphi_T)$, and $w_i := (x_i, z_i)$. The parameter vector we are interested in identifying is denoted by $\pi = (\beta', \varphi, \eta)$. Let $\|\cdot\|$ be the Euclidean norm on $\Re^K$ and let $\tau(w_i)$ be a weighting function, which we formally define in section 4. For any J-dimensional vector, $e = (e_1 \cdots, e_J)$, let $e_{-k} := \{e_j : j \neq k\}$ be the $(J-1)$-dimensional vector constructed by deleting the k-th element of e. Define $w_{it,-k} = (x_{it,-k}, z_i)$. Assumption 2.1 and the following assumption are sufficient for identification of the parameters of the model defined in equation (1.1).

Assumption 3.1.

1. *Equation (1.1) holds with $E[\varepsilon_{it}|w_{it}] = 0$, $t = 1, \cdots, T$. For at least one $k \in [1, \cdots, K]$, the conditional distribution of $x_{it,k}$ given $w_{it,-k}$ is absolutely continuous with respect to a Lebesgue measure with nondegenerate Radon-Nikodym density, and $\beta_k \neq 0$. Without loss of generality, let $k = K$.*
2. *rank$(E[(1, \Delta x_{it})'(1, \Delta x_{it})]) = K + 1$, $t = 2, \cdots, T$.*
3. *$\|\beta\| = 1$ and $E[\tau(w_i)\varphi_1(P_{i1})] = 0$.*

Part 1 of Assumption 3.1 states the model (1.1) generates y_{it}, and x_{itK} is a relevant

regressor drawn from a continuous distribution. It is analogous to Assumption A.1 that Honoré and Lewbel (2002) impose on their "special regressor." In our case, this assumption is necessary to identify the link functions. As shown in Newey (1994a), continuity of x_{itK} is not needed for identification of β if the link functions are known.

While Assumption 3.1.1 restricts $x_{iK} := (x_{i1K}, \cdots, x_{iTK})$ from being directly included in z_i, it does allow for dimension reducing functions of this vector to be included in z_i (such as $\sum_{t=1}^{T} x_{itK}$), so long as Assumption 3.1.1 is maintained. This dimension reducing strategy is proposed by Wooldridge (2002), where it is characterized as a Mundlak (1978) version of the assumption in Chamberlain (1980). In the static, panel data discrete-choice framework, the restrictions imposed in this paper are strictly weaker than those proposed in Wooldridge (2005), because $\eta(z_i)$ and the distribution of v_i remain unspecified in $\mu_i = \eta(z_i) + v_i$. Furthermore, as discussed in Altonji and Matzkin (2005), dimension-reducing restrictions other than the sample average may be implemented, and one may choose among different restrictions by comparing the predicted outcomes under each restriction to the predicted outcomes without the restriction.

Part 2 of Assumption 3.1 is a full-rank assumption that is necessary to identify π. It ensures the link functions φ_t are identified up to a common location constant. While Assumption 3.1.2 restricts x_{it} from including time-constant random variables, their effects can be controlled for by including them in z_i. Also, random variables that change by a fixed amount over time, such as age, cannot be included in x_{it}. Note that in contrast to the special regressor of Honoré and Lewbel (2002), this assumption requires that x_{itK} varies over time.

Part 3 of Assumption 3.1 includes the scale and location normalizations necessary for point identification of π. The assumption $\|\beta\| = 1$ fixes the scale π. This normalization is frequent in single-index models (e.g., Manski, 1985, and Manski, 1987). An alternative

normalization (see Horowitz, 1992, and Ichimura, 1993) is $|\beta_K| = 1$. We can also prove identification under this alternative normalization. The assumption that $E[\tau(w_i)\varphi_1(P_{i1})] = 0$ fixes the location of the φ's and η. We choose these particular normalizations because they are easy to implement in the proposed algorithm.

Let P_{it0} be the conditional expectation P_{it} induced by $\pi_0 = (\beta_0', \varphi_0, \eta_0)$, and assume P_{it0} coincides with the population conditional expectation, $E_t[y_{it}|w_{it}]$. Suppose an alternative vector of parameters, $\pi_* = (\beta_*', \varphi_*, \eta_*)$ is observationally equivalent to π_0, in that

$$P_{it0} = \Phi_{t*}\left(x_{it}\beta_* + \eta_*(z_i)\right), \quad t = 1, \cdots, T \tag{3.1}$$

almost surely, with $\varphi_{t*} = \Phi_{t*}^{-1}$. The identification theorem is stated as follows.

Theorem 3.2. *(Identification) Suppose π_0 and π_* satisfy Assumption 2.1 and parts 1 and 2 of Assumption 3.1. Then, for constants c and $a > 0$, $\beta_0 = a\beta_*$, $\eta_0 = a\eta_* + c$, and for $t = 1, \cdots, T$, $\varphi_{t0} = a\varphi_{t*} + c$. Furthermore, if π_0 and π_* also satisfy part 3 of Assumption 3.1, $a = 1$ and $c = 0$.*

Proof. See Appendix A.1 □

4. The Estimator

We define the estimator for (β_0', φ_0) in this section. Let $[(y_{it}, x_{it},\ t = 1, \cdots, T), z_i]$ be a random vector, where $y_{it} \in \mathcal{Y}_t \subseteq \Re$, $x_{it} \in \mathcal{X}_t \subseteq \Re^K$, and $z_i \in \mathcal{Z} \subseteq \Re^L$, so that $y_i := (y_{i1}, \cdots, y_{iT}) \in \mathcal{Y} := \times_{t=1}^T \mathcal{Y}_t \subseteq \Re^T$, $x_i \in \mathcal{X} := \times_{t=1}^T \mathcal{X}_t \subseteq \Re^{KT}$, $w_{it} \in \mathcal{X}_t \times \mathcal{Z} \subseteq \Re^{K+L}$, and $w_i \in \mathcal{X} \times \mathcal{Z} \subseteq \Re^{KT+L}$.

Let f_{w_t}, f_w, and f_{y_t, w_t} be the probability density functions of w_{it}, w_i, and (y_{it}, w_{it}) defined

on $\mathcal{X}_t \times \mathcal{Z}$, $\mathcal{X} \times \mathcal{Z}$, and $\mathcal{Y}_t \times \mathcal{X} \times \mathcal{Z}$ with respect to some dominating measure. Because the predicted outcomes, $P_{it0} = E_t[y_{it}|w_{it}] = \int(\tilde{y}_t f_{y_t,w_t}(\tilde{y}_t, w_{it})\nu(\mathrm{d}\tilde{y}))/f_{w_t}(w_{it})$, have the density f_{w_t} in the denominator, f_{w_t} must be bounded away from zero for $t = 1, \cdots, T$. We therefore impose a fixed-trimming condition by defining the compact subset $\mathcal{W} \subset \mathcal{X} \times \mathcal{Z}$, where f_w is bounded away from zero on $\mathcal{W}$. To this end, define the fixed-trimming function, $\tau_i = \tau(w_i) := 1\{w_i \in \mathcal{W}\}$. As pointed out by Newey (1994b), the fixed-trimming device is convenient because the resulting theory is less complicated than the one resulting from data-depending trimming.[2] Also, the theory resulting from fixed-trimming is roughly equivalent to trimming on a large auxiliary sample, which is often available in practice. This fixed-trimming condition implies a compact connected subset, $\mathcal{K} \subset \Re$, exists in which all the predicted outcomes P_{it0} lie. For elements $P_t \in \mathcal{K}, t = 1, \cdots, T$, define $P = (P_1, \cdots, P_T)'$. Let $\Lambda^2_{c_2}(\mathcal{K}) := \{m \in \mathcal{C}^2(\mathcal{K}) : \|m\|_{s,2} \leq c_2 < \infty\}$, where $\|\cdot\|_{s,j}$ is the supremum Sobolev norm of order $j = 0, 1, 2$, as defined by Newey (1994b), and let $\mathcal{S}_{\mathcal{K}}$ be a compact and convex subset of $\Lambda^2_{c_2}(\mathcal{K})$ composed of increasing functions. Define the operator Δ as $a := (a_1, \ldots, a_T)' \longmapsto \Delta a := (a_2 - a_1, \ldots, a_T - a_{T-1})'$,[3] and let

$$
\begin{aligned}
\tilde{\mathcal{F}} &:= \left\{m(a) = (m_2(a) \cdots, m_T(a))' : \right. \\
&\qquad \left. m_t(a) = m_t(a_1, \cdots, a_T) \in \Re, t = 2, \cdots, T\right\}, \\
\mathcal{F} &:= \left\{m(a) \in \tilde{\mathcal{F}} : \Delta m(a) = (\Delta m_2(a_2), \cdots, \Delta m_T(a_T)), m_t(a_t) \in \Lambda^2_{c_2}(\mathcal{K}), t = 1, \cdots, T\right\}, \\
\mathcal{F}^0 &:= \left\{m(a) \in \mathcal{F} : \int m_1(P_1) f_{P_1}(P_1) dP_1 = 0\right\}, \quad \text{and} \\
\mathcal{F}_c &:= \left\{m(a) \in \mathcal{F}^0 : m_t(a_t) \in \mathcal{S}_{\mathcal{K}}, t = 1, \cdots, T\right\}.
\end{aligned}
$$

[2] see Robinson (1988), and Ai (1997) for examples of theoretical results with data-dependent trimming.

[3] In what is to come, we will also define Δ as follows $\Delta x_{it} := x_{it} - x_{it-1}$, $\Delta\varphi_t = \varphi_t - \varphi_{t-1}$, and so on. This is to conserve on notation. The distinction in the alternating definitions is clear from observing what Δ is operating on.

Assume $\theta_0 := (\beta_0', \varphi_0) \in \Theta := \mathcal{B} \times \mathcal{F}_c$, where $\mathcal{B} \subset \Re^K$ is compact and convex with non-empty interior. We further require the induced density of P, denoted by $f_P(P)$, to be bounded away from zero on $\mathcal{K}$. This requirement holds in general given boundedness conditions on $f_w(w)$ (see Mood et al. (1974), sections 5 and 6 for detailed discussions). Define $\rho(x, P, \theta) = (\Delta\varphi(P) - \Delta x\beta)$ and $Q_i(\theta, P) = \tau_i \rho(x_i, P, \theta)' \Sigma^{-1} \rho(x_i, P, \theta)$, where Σ is a $(T-1)$-dimensional symmetric, positive-definite weighting matrix. Let $\breve{\theta}_0$ minimize the following objective function:

$$Q_0(\theta) := E\left[Q_i(\theta, P_{i0})\right] \tag{4.1}$$

over Θ. Because Θ is a compact and convex set and $Q_0(\theta)$ is continuous in θ, the solution $\breve{\theta}_0$ exists and is typically set valued. However, the identification results of Theorem 3.2 imply the transformation $\theta_0 := (\breve{\beta}_0'/a, \{\breve{\varphi}_{t0}/a,\ t = 1, \cdots, T\})$, where $a := \|\breve{\beta}_0\|$, maps $\breve{\theta}_0$ onto a singleton.

Estimation of θ_0 from the sample analog of equation (4.1) is infeasible because the predicted outcomes are unknown. Replacing P_{it0} with a consistent kernel estimator, $\hat{P}_{it}$, resolves this issue. Define the function

$$K_{\sigma_1}(c - c_i) = \sigma_1^{-d_c} \prod_{k=1}^{d_c} K_1((c_k - c_{ik})/\sigma_1),$$

where d_c is the dimension of c, $K_1 : \Re \to \Re$ is a kernel, and σ_1 is the bandwidth. For $t = 1, \cdots, T$, let $q_{it} := (1, y_{it})$ and define $\hat{\gamma}_t(w_t) = (\hat{\gamma}_{1t}(w_t), \hat{\gamma}_{2t}(w_t))$ by

$$\hat{\gamma}_t(w_t) \quad = \quad N^{-1} \sum_{j=1}^{N} q_{jt} K_{\sigma_1}(w_t - w_{jt}).$$

Then the estimator for P_{it0} is defined by $\hat{P}_{it} = \hat{\gamma}_{2t}(w_{it})/\hat{\gamma}_{1t}(w_{it})$, and $\hat{f}_{w_t}(w_t) := \hat{\gamma}_{1t}(w_t)$ is the estimator for $f_{w_t}(w_t)$. In section 6, we provide sufficient conditions for $\hat{f}_{w_t}(w_t)$, $t = 1, \cdots, T$ to be bounded away from zero uniformly on $\mathcal{W}$ with probability approaching one.

To define the estimator, let

$$
\begin{aligned}
\tilde{\mathcal{F}}^N &:= \{m = (m^i, i = 1, \cdots, N) : m^i \in \tilde{\mathcal{F}}\}, \\
\mathcal{F}^N &:= \{m = (m^i, i = 1, \cdots, N) : m^i \in \mathcal{F}\}, \\
\mathcal{F}_m^N &:= \{m \in \mathcal{F}^N : m^i \text{ does not depend on } i\}, \\
\mathcal{F}_m^{0,N} &:= \left\{ m \in \mathcal{F}_m^N : \int m_1(P_1)\, \hat{f}_P(P_1) dP_1 = 0 \right\}, \quad \text{and} \\
\mathcal{F}_c^N &:= \left\{ m \in \mathcal{F}_m^{0,N} : m \in \mathcal{F}_c \right\}.
\end{aligned}
$$

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

$$
\begin{aligned}
m + g &= (m^i + g^i,\ i = 1, \cdots, n), \text{ for } m, g \in \tilde{\mathcal{F}}, \quad \textit{and} \\
\alpha \cdot m &= (\alpha m^i,\ i = 1, \cdots, n), \text{ for } \alpha \in \mathfrak{R},\ m \in \tilde{\mathcal{F}}.
\end{aligned}
$$

For a fixed P_t, an element of P, define $\hat{\omega}_{it}(P_t) = \sigma_2^{-1} K_2(\sigma_2^{-1}(\hat{P}_{it} - P_t))$, where σ_2 is a positive constant and K_2 is a kernel with $K_2(\cdot) \geq 0$. Then $\hat{f}_{P_t}(P_t) := N^{-1} \sum_{i=1}^N \tau_i \hat{\omega}_{it}(P_t)$ is the estimator for the marginal density $f_{P_t}(P_t)$ of P_t, $\hat{f}_{t,s}(P_t, P_s) := N^{-1} \sum_{i=1}^N \tau_i \hat{\omega}_{it}(P_t) \hat{\omega}_{is}(P_s)$ is the estimator for the joint density $f_{t,s}(P_t, P_s)$ of (P_t, P_s), $t, s \in \{1, \cdots, T\}$, $s \neq t$, and $\hat{f}_P(P) := N^{-1} \sum_{i=1}^N \tau_i \hat{\omega}_i(P)$, where $\hat{\omega}_i(P) := \prod_{t=1}^T \hat{\omega}_{it}(P_t)$, is the estimator for $f_P(P)$, the

joint density of $P = (P_1, \cdots, P_T)$. Define the inner product on $\tilde{\mathcal{F}}^N$ by

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

where W is a positive definite matrix W. This inner product induces the following semi-norm on $\tilde{\mathcal{F}}^N$:

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

Define the sample residual vector $\rho(x_i, P, \theta) = (\Delta\varphi(P) - \Delta x_i \beta)$, and let

$$\hat{Q}_i(P, \theta) := \tau_i \rho(x_i, P, \theta)' \hat{\Sigma}^{-1} \rho(x_i, P, \theta),$$

where $\hat{\Sigma}$ is a consistent estimator of Σ. Define $\breve{\theta}$ to be the minimizer of

$$\hat{Q}(\theta) := \int N^{-1} \sum_{i=1}^{N} \hat{Q}_i(P, \theta) \hat{\omega}_i(P) \mathrm{d}P \tag{4.2}$$

over $\Theta_N := \mathcal{B} \times \mathcal{F}_c^N$. It can be shown that $\mathcal{F}_c^N$ is a compact and convex set. Because Θ_N is compact and convex, and $\hat{Q}(\theta)$ is continuous in θ, the solution $\breve{\theta}$ exists with probability approaching one, and is typically set valued. The feasible semiparametric least squares estimator of θ_0 is given by $\hat{\theta} := (\breve{\beta}'/\hat{a}, \{\breve{\varphi}_t/\hat{a}, t = 1, \cdots, T\})$, where $\hat{a} := \|\breve{\beta}\|$.

Remark 4.1. For the semi-norm defined above to be well-defined, we require that $\hat{\omega}_{it} \geq 0$ and $\hat{\omega}_{it} = 0$ on a set of measure zero. An consequence of this restriction is that higher-order kernels cannot be used to define $\hat{\omega}_{it}$.

5. Computing the Estimator

The approach to computing the estimator of $\theta_0 = (\beta_0', \varphi_0)$ from the objective function in equation (4.2) can be summarized as follows. First, compute the estimator of the infinite-dimensional parameter, φ_0, for any fixed value of the finite-dimensional parameter β. Next, substitute this estimator (a function of β) for φ into the objective function (4.2), and solve for the estimator of β. This approach is indeed the intuition of the profile-likelihood approach of Severini and Wong (1992). To describe the algorithm, further definitions and regularity conditions are necessary. Let $\gamma \mapsto \theta(\gamma) = (\beta(\gamma)', \varphi(\gamma))$ be a smooth mapping from the real hypercube $(a,\ b)^K$ to Θ, and assume the curve may be parameterized to have $\beta \mapsto (\beta', \varphi(\beta))$. For fixed $\beta \in \mathcal{B}$, let $\varphi_0(\beta)$ minimize $Q_0(\beta, \varphi)$ over $\mathcal{F}_c$ so that $\varphi_0(\beta_0) = \varphi_0$. Computation of the estimator for θ_0 proceeds by computing the estimator for $\varphi_0(\beta)$ denoted by $\hat{\varphi}(\beta)$, which minimizes $\hat{Q}(\beta, \varphi)$ over $\mathcal{F}_c^N$ for fixed β, and then minimizing $\hat{Q}(\beta, \hat{\varphi}(\beta))$ over $\mathcal{B}$ to obtain the estimator $\hat{\beta}$ of β_0. The estimator of $\varphi_0(\beta_0)$ is obtained by evaluating $\hat{\varphi}(\beta)$ at $\hat{\beta}$ to have $\hat{\varphi} := \hat{\varphi}(\hat{\beta})$.

5.1. Projection onto $\mathcal{F}_c^N$

Let part (1) of Assumption 6.2 hold. We begin by defining the projection of $\Delta x\beta$ onto $\mathcal{F}_c^N$ for a fixed $\beta \in \mathcal{B}$. This projection is defined as the fixed point to a backfitting algorithm. Proposition 1 of Mammen et al. (2001) implies this projection can be decomposed into four cascading projections, which we detail below. The first is the projection of $\Delta x\beta$ onto the set $\tilde{\mathcal{F}}^N$ to obtain the $(T-1)$-dimensional unconstrained estimator $\check{m}(\beta) := (\check{m}_2(\beta), \cdots, \check{m}_T(\beta))'$. Note that for $t = 2, \cdots, T$, and for fixed P, $\check{m}_t(P, \beta)$ is an estimator for $m_{t0}(P, \beta) := E[\tau_i \Delta x_{it}\beta | P]$, the conditional expectation of x_{it} given P. The second is the projection of $\check{m}(\beta)$ onto the set $\mathcal{F}_m^N$ to obtain the T-dimensional estimator

$\varphi^{**}(\beta) := (\varphi_1^{**}(\beta), \cdots, \varphi_T^{**}(\beta))$. The third is the projection of $\varphi^{**}(\beta)$ onto $\mathcal{F}_m^{0,N}$ to obtain the T-dimensional vector of location-normalized estimator $\varphi^*(\beta) := (\varphi_1^*(\beta), \cdots, \varphi_T^*(\beta))$. The fourth is the projection of $\varphi^*(\beta)$ onto $\mathcal{F}_c^N$ to obtain our constrained estimator $\breve{\varphi}(\beta) := (\breve{\varphi}_1(\beta), \cdots, \breve{\varphi}_T(\beta))'$. By construction, $\mathcal{F}_c^N \subset \mathcal{F}_m^N \subset \mathcal{F}_m^{0,N} \subset \tilde{\mathcal{F}}^N$ so that by the law of iterated projections, these four steps do obtain the projection of $\Delta x\beta$ onto $\mathcal{F}_c^N$. To begin, one needs to choose a grid on $\mathcal{K}$ where the projections are evaluated. We recommend the grid be constructed in the interior of $\mathcal{K}$ to avoid boundary problems. See Mammen et al. (2001) for discussions on the choice of the weight measure.

Projection onto $\tilde{\mathcal{F}}^N$.

The $(T-1)$-dimensional unconstrained estimator, $\breve{m}(\beta) = (\breve{m}_2(\beta), \cdots, \breve{m}_T(\beta))'$, is given by

$$\breve{m}(\beta) = \arg\min_{\tilde{m} \in \tilde{\mathcal{F}}^N} \int \frac{1}{N} \sum_{i=1}^N \tau_i (\tilde{m} - \Delta x_i \beta)' \hat{\Sigma}^{-1} (\tilde{m} - \Delta x_i \beta) \hat{\omega}_i(P) dP.$$

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

$$\breve{m}(P, \beta) := \arg\min_{\tilde{m} \in \tilde{\mathcal{F}}^N} \frac{1}{N} \sum_{i=1}^N \tau_i (\tilde{m} - \Delta x_i \beta)' \hat{\Sigma}^{-1} (\tilde{m} - \Delta x_i \beta) \hat{\omega}_i(P), \tag{5.1}$$

with the solution given by $\breve{m}_t(P, \beta) = N^{-1} \sum_{i=1}^N \tau_i \Delta x_{it} \beta \hat{\omega}_i(P) / \hat{f}_P(P), t = 2, \cdots, T$. Notice $\breve{m}_t(P, \beta)$ has the following alternative representation: $\breve{m}_t(P, \beta) = \Delta \hat{m}_t(P, \beta)$, where $\hat{m}_t(P, \beta) := N^{-1} \sum_{i=1}^N \tau_i x_{it} \beta \hat{\omega}_i(P) / \hat{f}_P(P)$.

Projection onto $\mathcal{F}_m^N$.

We next define the empirical projection of the solution $\breve{m}(\beta)$ onto the set $\mathcal{F}_m^N$. The solution

of this projection $\varphi^{**}(\beta)$ minimizes

$$\begin{aligned}
\|\breve{m}(\beta)-\Delta\tilde{\varphi}\|_T^2 &= \int N^{-1}\sum_{i=1}^{N}\tau_i[\breve{m}(P,\beta)-\Delta\tilde{\varphi}(P)]'\hat{\Sigma}^{-1}[\breve{m}(P,\beta)-\Delta\tilde{\varphi}(P)]\hat{\omega}_i(P)\mathrm{d}P, \\
&= \int [\breve{m}(P,\beta)-\Delta\tilde{\varphi}(P)]'\hat{\Sigma}^{-1}[\breve{m}(P,\beta)-\Delta\tilde{\varphi}(P)]\,\hat{f}(P)\mathrm{d}P, \\
&= \int [\hat{m}(P,\beta)-\tilde{\varphi}(P)]'\Delta'\hat{\Sigma}^{-1}\Delta[\hat{m}(P,\beta)-\tilde{\varphi}(P)]\,\hat{f}(P)\mathrm{d}P, \\
&= \int \sum_{s=1}^{T}\sum_{t=1}^{T}\hat{\sigma}_*^{st}[\hat{m}_s(P,\beta)-\tilde{\varphi}_s(P)][\hat{m}_t(P,\beta)-\tilde{\varphi}_t(P)]\,\hat{f}(P)\mathrm{d}P
\end{aligned}$$

over $\mathcal{F}_m^{0,N}$, where $\hat{\sigma}_*^{st} := [\Delta'\hat{\Sigma}^{-1}\Delta]_{st}$. For $s,t=1,\cdots,T$, define $\hat{f}_{s|t}(P_s|P_t)=\hat{f}_{s,t}(P_s,P_t)/\hat{f}_{P_t}(P_t)$ and $\hat{m}_{st}(P_t,\beta)=\sum_{i=1}^{N}\tau_i x_{is}\beta\hat{\omega}_{it}(P_t)/\hat{f}_{P_t}(P_t)$. Notice $\hat{f}_{s|t}(P_s|P_t)$ is an estimator for the conditional density of P_s, conditioned on P_t, and $\hat{m}_{st}(P_t,\beta)$ is an estimator for $m_{st0}(P_t,\beta) := E[\tau_i x_{is}\beta|P_t]$, the conditional expectation of $x_{is}\beta$ given P_t. Noting $\hat{m}_t(P_t,\beta) := \hat{m}_{tt}(P_t,\beta)$, the solution is characterized by the following system of equations:

$$\begin{aligned}
\varphi_1^{**}(P_1,\beta) &= \hat{m}_1(P_1,\beta)+\textstyle\sum_{s=2}^{T}\frac{\hat{\sigma}_*^{s1}}{\hat{\sigma}_*^{11}}\left[\hat{m}_{s1}(P_1,\beta)-\int\varphi_s^{**}(P_s,\beta)\,\hat{f}_{s|1}(P_s|P_1)dP_s\right], \\
\varphi_2^{**}(P_2,\beta) &= \hat{m}_2(P_2,\beta)+\textstyle\sum_{s\neq 2}\frac{\hat{\sigma}_*^{s2}}{\hat{\sigma}_*^{22}}\left[\hat{m}_{s2}(P_2,\beta)-\int\varphi_s^{**}(P_s,\beta)\,\hat{f}_{s|2}(P_s|P_2)dP_s\right], \\
&\vdots \\
\varphi_T^{**}(P_T,\beta) &= \hat{m}_T(P_T,\beta)+\textstyle\sum_{s=1}^{T-1}\frac{\hat{\sigma}_*^{sT}}{\hat{\sigma}_*^{TT}}\left[\hat{m}_{sT}(P_T,\beta)-\int\varphi_s^{**}(P_s,\beta)\,\hat{f}_{s|T}(P_s|P_T)dP_s\right].
\end{aligned} \tag{5.2}$$

The details of the derivation of this system of equations are similar to those in Mammen et al. (1999). Note that $\varphi_t^{**}(P_t,\beta)$ can equivalently be written as

$$\varphi_t^{**}(P_t,\beta)=\frac{1}{N}\sum_{i=1}^{N}\tau_i\frac{\hat{\omega}_{it}(P_2)}{\hat{f}_{P_t}(P_t)}\left\{x_{it}\beta+\sum_{s\neq t}\frac{\hat{\sigma}_*^{st}}{\hat{\sigma}_*^{22}}\left[x_{is}\beta-\int\varphi_s^{**}(P_s,\beta)\hat{\omega}_{is}(P_s)dP_s\right]\right\},$$

which may have numerical advantages over the representations in equation (5.2).[4] For fixed β, the system of equations (5.2) can be solved by the following backfitting algorithm.

Inner Backfitting Algorithm (IBA)

Step 1. Obtain initial guesses $(\varphi_t^{**[0]}(P_t),\ t = 1, \cdots, T)$.

Step 2. Apply the following loop:

Do for $r \geq 1$

$$\begin{aligned}
\varphi_1^{**[r]}(P_1,\beta) &= \hat{m}_1(P_1,\beta) + \textstyle\sum_{s=2}^{T} \frac{\hat{\sigma}_*^{s1}}{\hat{\sigma}_*^{11}} \left[\hat{m}_{s1}(P_1,\beta) - \int \varphi_s^{**[r-1]}(P_s,\beta)\, \hat{f}_{s|1}(P_s|P_1) dP_s\right], \\
\varphi_t^{**[r]}(P_t,\beta) &= \hat{m}_t(P_t,\beta) + \textstyle\sum_{s<t} \frac{\hat{\sigma}_*^{st}}{\hat{\sigma}_{**}^{tt}} \left[\hat{m}_{st}(P_s,\beta) - \int \varphi_s^{**[r]}(P_s,\beta)\, \hat{f}_{s|t}(P_s|P_t) dP_s\right], \\
&+ \textstyle\sum_{s>t} \frac{\hat{\sigma}_*^{st}}{\hat{\sigma}_*^{tt}} \left[\hat{m}_{st}(P_s,\beta) - \int \varphi_s^{**[r-1]}(P_s,\beta)\, \hat{f}_{s|t}(P_s|P_t) dP_s\right], \\
&\quad \text{for } t = 2, \cdots, T,
\end{aligned} \tag{5.3}$$

until convergence in $\varphi^{**}(\beta)$ is reached.

Projection onto $\mathcal{F}_m^{0,N}$.

The projection of $\varphi^{**}(\beta)$ onto $\mathcal{F}_m^{0,N}$ yields

$$\varphi_t^*(\beta) = \varphi_t^{**}(\beta) - \int \varphi_1^{**}(P_1,\beta)\, \hat{f}_{P_1}(P_1) dP_1, t = 1, \cdots, T. \tag{5.4}$$

Note that subtracting the same constant from each of the functions $\varphi^{**}(P_t,\beta)$ leaves the objective function unchanged.

Projection onto $\mathcal{F}_c^N$.

[4] We thank an anonymous referee for pointing out this equivalent representation of $\varphi_t^{**}(P_t,\beta)$.

We next define the empirical projection of $\varphi^*(\beta)$ onto $\mathcal{F}_c^N$. The solution $\breve{\beta}$ minimizes

$$\|\Delta\varphi^*(\beta) - \Delta\tilde{\varphi}\|_T^2 = \int \sum_{t=1}^{T}\sum_{s=1}^{T} \hat{\sigma}_*^{st}[\varphi_s^*(P_s,\beta) - \tilde{\varphi}_s(P_s)][\varphi_t^*(P_t,\beta) - \tilde{\varphi}_t(P_t)]\,\hat{f}_{s,t}(P_s,P_t)dP_s dP_t \tag{5.5}$$

over $\mathcal{F}_c^N$. At first, computing $\breve{\varphi}(\beta)$ seems to involve another round of backfitting, which would then prove to be computationally costly in practice. However, the next theorem states the projection from $\mathcal{F}_m^{0,N}$ onto $\mathcal{F}_c^N$ can be done for each function $\varphi_t(\beta)$ separately. In other words, the "off-diagonal" terms in the objective function have no effect on the minimization of $\|\Delta\varphi^*(\beta) - \Delta\tilde{\varphi}\|_T^2$ over $\tilde{\varphi} \in \mathcal{F}_c^N$.

Theorem 5.1. *For $t = 1,\cdots,T$, let $\breve{\varphi}_t(\beta)$ minimize $\int[\varphi_t^*(P_t,\beta) - \tilde{\varphi}_t(P_t)]^2\hat{f}_{P_t}(P_t)dP_t$ over $\tilde{\varphi}_t \in \mathcal{S}_{\mathcal{K}}$. Then $\breve{\varphi}(\beta) = (\breve{\varphi}_1(\beta),\cdots,\breve{\varphi}_T(\beta))$ minimizes $\|\Delta\varphi^*(\beta) - \Delta\tilde{\varphi}\|_T^2$ over $\tilde{\varphi} \in \mathcal{F}_c^N$.*

Proof. See Appendix A.2 □

The results of Barlow et al. (1972), Robertson et al. (1988), and Mammen et al. (2001) then imply $\breve{\varphi}_t(P_t)$, $t = 1,\cdots,T$ are given by

$$\breve{\varphi}_t(P_t,\beta) = \inf_{v \geq P_t}\sup_{u \leq P_t} \frac{\int_u^v \varphi_t^*(\tilde{P}_t,\beta)\,\hat{f}_{P_t}(\tilde{P}_t)\mathrm{d}\tilde{P}_t}{\int_u^v \hat{f}_{P_t}(\tilde{P}_t)\mathrm{d}\tilde{P}_t}. \tag{5.6}$$

5.2. Projection onto $x\mathcal{B}$

Given the estimator for $\breve{\varphi}(\beta)$, the next step is to project this vector (an element of $\mathcal{F}_c^N$) onto $x\mathcal{B}$. This projection minimizes $\hat{Q}(\beta,\breve{\varphi}(\beta))$ over $\mathcal{B}$. Let $\hat{Q}_\beta(\beta,\breve{\varphi}(\beta))$ be the partial derivative of $\hat{Q}(\beta,\breve{\varphi}(\beta))$ with respect to β holding the effect of β on $\breve{\varphi}$ constant (referred to as the direct effect). Define $\hat{Q}_{\varphi_t}(\beta,\breve{\varphi}(\beta))$ to be the derivative of $\hat{Q}(\beta,\tilde{\varphi}_t,\breve{\varphi}_{-t}(\beta))$ with respect to $\tilde{\varphi}_t$ and evaluated at $\breve{\varphi}_t(\beta)$. The derivative of $\hat{Q}_\beta(\beta,\breve{\varphi})$ with respect to β is given

by

$$\nabla_\beta \hat{Q}(\beta, \breve{\varphi}(\beta)) \quad = \quad \hat{Q}_\beta(\beta, \breve{\varphi}) + \sum_{t=1}^{T} \hat{Q}_{\varphi_t}(\beta, \breve{\varphi}(\beta))\,(\partial \breve{\varphi}_t(\beta)/\partial \beta)\,. \tag{5.7}$$

Computation of the solution $\breve{\beta}$ of equation (5.7) would be greatly simplified if the envelope theorem (or, equivalently, Proposition 2 of Newey (1994a)) holds so that $\hat{Q}_{\varphi_t}(\beta, \breve{\varphi}(\beta)) = 0$ identically in β. If this condition holds, the solution $\breve{\beta}$ does not depend on the derivative of $\breve{\varphi}(\beta)$ with respect to β. Because $\breve{\varphi}(\beta)$ minimizes a convex function over a convex set, we typically only know that $\hat{Q}_{\varphi_t}(\beta, \breve{\varphi}(\beta))[\breve{\varphi}_t - \tilde{\varphi}_t] \geq 0$ for any $\tilde{\varphi}_t \in \mathcal{S}_{\mathcal{K}}$. As a result, it would seem that the envelope theorem does not apply to the current framework. However, the law of iterated projections obtains

$$\|\Delta x\beta - \Delta\breve{\varphi}(\beta)\|_T^2 \quad = \quad \|\Delta x\beta - \breve{m}(\beta)\|_T^2 + \|\Delta\hat{m}(\beta) - \Delta\varphi^*(\beta)\|_T^2 + \|\Delta\varphi^*(\beta) - \Delta\breve{\varphi}(\beta)\|_T^2$$

(see Section 6.3 or equation (4.4) of Mammen et al. (2001), and the references therein), so that

$$\hat{Q}_{\varphi_t}(\beta, \breve{\varphi}(\beta)) = \frac{\partial}{\partial \tilde{\varphi}_t}\|\Delta x\beta - \Delta\breve{\varphi}(\beta)\|_T^2 = \frac{\partial}{\partial \tilde{\varphi}_t}\|\Delta\varphi^*(\beta) - \Delta\breve{\varphi}(\beta)\|_T^2 = 0$$

identically in β, where the second equality is implied by Theorem 5.1 and the third by Theorem 1.3.2 of Robertson et al. (1988), which is obtained by noting the projection is onto a convex cone. Hence, the envelope theorem does apply to our framework and equation (5.7) obtains the implicit solution $\breve{\beta}$ as follows:

$$\breve{\beta} = \left[\sum_{i=1}^{N} \tau_i x_i' \Delta' \hat{\Sigma}^{-1} \Delta x_i\right]^{-1} \left[\sum_{i=1}^{N} \tau_i x_i' \Delta' \hat{\Sigma}^{-1} \Delta \int \breve{\varphi}(P, \breve{\beta})\hat{\omega}_i(P)dP\right], \tag{5.8}$$

or equivalently,

$$\breve{\beta} = \left[\sum_{i=1}^{N} \tau_i x_i' \Delta' \hat{\Sigma}^{-1} \Delta x_i\right]^{-1} \left[\int \sum_{t=1}^{T} \left(\sum_{s=1}^{T} \hat{\sigma}_*^{st} \hat{m}_{st}(P_t, \breve{\beta})\right)' \breve{\varphi}_t(P_t, \breve{\beta}) \hat{f}_t(P_t) dP_t\right].$$

Therefore, $\breve{\beta}$ can be solved by implementing the following outer backfitting algorithm.

Outer Backfitting Algorithm (OBA)

Step 1. Obtain initial guesses $\breve{\beta}^{[1]}$ and $\breve{\varphi}^{[0]}(\breve{\beta}^{[0]})$.

Step 2. Apply the following loop.

Do for $s \geq 1$

• Compute the updated estimates $\varphi^{**[s]}(\breve{\beta}^{[s]})$ by implementing the IBA initialized by $\breve{\varphi}^{[s-1]}(\breve{\beta}^{[s-1]})$ and β fixed at $\breve{\beta}^{[s]}$.

• Update $\varphi_t^*,\ t = 1, \cdots, T$ by

$$\varphi_t^{*[s]}(P_t, \beta^{[s]}) = \varphi^{**[s]}(P_t, \beta^{[s]}) - \int \varphi_1^{**[s]}(P_1, \beta^{[s]}) \hat{f}_{P_1}(P_1) dP_1.$$

• Update $\breve{\varphi}_t,\ t = 1, \cdots, T$ by

$$\breve{\varphi}_t^{[s]}(P_t, \breve{\beta}^{[s]}) = \inf_{v \geq P_t} \sup_{u \leq P_t} \frac{\int_u^v \varphi_t^{*[s]}(P_t, \breve{\beta}^{[s]}) \hat{f}_{P_t}(P_t) \mathrm{d}P_t}{\int_u^v \hat{f}_{P_t}(P_t) \mathrm{d}P_t}.$$

• Update β using equation (5.8), by

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

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 $\hat{\varphi}_t = \breve{\varphi}_t/\hat{a},\ t = 1, \cdots, T$.

Remark 5.2. As in conventional panel data models, if $\Phi_t = \Phi$ for $t = 1, \cdots, T$, then the method developed in this section can conveniently accommodate this additional restriction by letting

$$\bar{\mathcal{F}} := \left\{ m(a) \in \mathcal{F}^0 : m_t(a_t) = m(a_t),\ t = 1, \cdots, T \right\},$$

and projecting $\varphi^*(\beta)$ onto $\bar{\mathcal{F}}$. For the vector $P = (P_1, \cdots, P_1)$, it can be shown the solution is a weighted average of $\varphi_t^*(P_1, \beta)$, $t = 1, \cdots, T$, where the weights are given by $\hat{\sigma}_*^{tt} \hat{f}_{P_t}(P_1) / \sum_{s=1}^{T} \hat{\sigma}_*^{ss} \hat{f}_{P_s}(P_1)$. This restriction would then be added between the second and third bullet points in step 2 of the OBA. Also, if we assume the link functions are not time dependent, then Assumption 3.1.2 may be relaxed to: $\text{rank}(E[(\Delta x_{it})' \Delta x_{it}]) = K,\ t = 2, \cdots, T$. Time dummies may then be included in x_{it} to control for aggregate effects.

6. Convergence of the algorithm and asymptotic properties of the estimator

To derive the asymptotic properties of the estimator, some regularity conditions are needed. We use the following notations in all assumptions, theorems and proofs: $\sup_{w_t} = \sup_{w_t \in \mathcal{W}}$, $\sup_{P_t} = \sup_{P_t \in \mathcal{K}}$, $\sup_{\varphi_t} = \sup_{\varphi_t \in \mathcal{S}_{\mathcal{K}}}$, and $\sup_{\beta} = \sup_{\beta \in \mathcal{B}}$. Also, we adopt the short-hand notations: $\sup_{P_t, \beta}$ to mean $\sup_{P_t} \sup_{\beta}$, $\sup_{P_t, P_s}$ to mean $\sup_{P_t} \sup_{P_s}$, and so on. Let $F_{y,w}$ denote the population distribution of $[(y_{it}, x_{it},\ t = 1, \cdots, T), z_i]$.

Assumption 6.1. *A sample of N independent realizations is drawn from $F_{y,w}$. For each $i = 1, \cdots, N$, $[(y_{it}, x_{it},\ t = 1, \cdots, T), z_i]$ is observed.*

Define $a_{it} := \hat{P}_{it} - P_{it0}$, $a_i := (a_{i1}, \cdots, a_{iT})$, and let $f_{x,P,a}$ be the joint density of (x_i, P_i, a_i).

Assumption 6.2. *1. $K_2(u)$ is differentiable of order $d_2 \geq 2$; the derivatives of order d_2 are bounded; $K_2(u)$ is zero outside a bounded set; $K_2(u) \geq 0$; $\int K_2(u)du = 1$; and $\int uK_2(u)du = 0$. 2. For $t = 1, \cdots, T$, a version of $\varphi_{0t}(P_t, \beta)$ exists, say $\tilde{\varphi}_{0t}(P_t, \beta)$, such that for each (P_t, β) in an open set containing $\mathcal{K} \times \mathcal{B}$ and for all $k, r = 0, 1, 2$, $k + r \leq 3$,*

$$\left| \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r} \tilde{\varphi}_{0t}(P_t, \beta) \right|$$

exists. 3. For $t = 1, \cdots, T$, a version of $f_{P_t}(P_t)$ exists that is continuously differentiable to order d_2 with bounded derivatives on an open set containing $\mathcal{K}$. 4. For $t = 1, \cdots, T$, a version of the joint density $f_{x,P_t,a}$ exists, say $\tilde{f}_{x,P_t,a}$ that is continuously differentiable in P_t to order d_2 on an open set containing $\mathcal{K}$ with $\sup_{\tilde{x},\tilde{P},\tilde{a}} |\tau \partial^r \tilde{f}_{x,P,a}(\tilde{x}, \tilde{P}_t, \tilde{a}) / \partial P_t^r| < \infty$ for r=0,1,2. 5. The bandwidth $\sigma_2 = \sigma_2(N)$ satisfies $N\sigma_2 / \ln N \longrightarrow \infty$.

If the P_{i0} and Σ were known, Assumption 6.2 along with additional conditions on the bandwidth σ_2 would be enough to ensure uniform convergence of $\hat{\varphi}_t(P_t, \beta)$ to $\varphi_{0t}(P_t, \beta)$. However, because they are not known and are estimated, additional regularity conditions are necessary to ensure uniform convergence of plug-in estimators $\hat{\Sigma}$ and $\hat{P}_{it}$. For $t = 1, \cdots, T$, let $\gamma_{t0}(w_t) := (\gamma_{1t0}(w_t), \gamma_{2t0}(w_t))$, where $\gamma_{1t0}(w_t) := f_{w_t}(w_t)$ and $\gamma_{2t0}(w_t) := f_{w_t}(w_t) E[y_{it}|w_t]$. Note that $P_{it0} = \gamma_{20}(w_{it}) / \gamma_{10}(w_{it})$. Define $\gamma_0(w) = (\gamma_{10}(w_1), \cdots, \gamma_{T0}(w_T))$.

Assumption 6.3. *1. $K_1(u)$ is differentiable of order $d_1 \geq 2$; the derivatives of order d_1 are bounded; $K_1(u)$ is zero outside a bounded set; $\int K_1(u)du = 1$; and a positive integer*

m_2 exists such that for all $j < m_2$, $\int K_1(u)u^j du = 0$. 2. A version of $\gamma_0(w)$ exists that is continuously differentiable to order d_1 with bounded derivatives on an open set containing $\mathcal{W}$. 3. There is $p \geq 4$ such that $E[\|y\|^p] < \infty$ and $E[\|y\|^p|w]f_0(w)$ is bounded. 4. The bandwidth $\sigma_1 = \sigma_1(N)$ satisfies $N^{1-(2/p)}\sigma_1^{K+L}/\ln N \longrightarrow \infty$.

Assumption 6.4. *The weighting matrix $\hat{\Sigma}$ is symmetric and positive definite.*

Let $\hat{Q}(P_t,\theta) := \int N^{-1}\sum_{i=1}^N \hat{Q}_i(P,\theta)\hat{\omega}_i(P)dP_{-t}$, and $Q_0(P_t,\theta) := E[Q_i(P_i,\theta)|P_t]f_{P_t}(P_t)$. Define $\hat{Q}^{(r)}(P_t,\theta)$ and $Q_0^{(r)}(P_t,\theta)$ to be the r-th derivative of $\hat{Q}(P_t,\theta)$ and $Q_0(P_t,\theta)$ with respect to P_t. Define $\hat{f}_{P_t}^{(r)}(P_t)$, $f_{P_t}^{(r)}(P_t)$, $\hat{f}_{t,s}^{(r)}(P_t,P_s)$, and $f_{t,s}^{(r)}(P_t,P_s)$ analogously. The following lemma provides uniform boundedness in probability results, which are useful for obtaining consistency and rates-of-convergence results of the nuisance parameters, as well as $\sqrt{N}$-convergence of the estimator $\hat{\beta}$. Its proof is in the online appendix (available at *http://people.virginia.edu/ wg4b/*).

Lemma 6.5. *Suppose (i) Assumptions 2.1 and 3.1 hold, (ii) Assumptions 6.1- 6.4 hold.*

Then for $j, r = 0, 1$, $k = 0, 1, 2$, $k + r \leq 1$, *and* $t = 1, \cdots, T$,

$$\sup_{P_t, \beta, \varphi} \left\| \frac{\partial^{k+j}}{\partial \beta^k \partial \varphi_t^j} \hat{Q}^{(r)}(P_t, \theta) - \frac{\partial^{k+j}}{\partial \beta^k \partial \varphi_t^j} Q_0^{(r)}(P_t, \theta) \right\| =$$
$$O_p \Big(\ln(N)^{1/2} (N\sigma_2^{2r+1})^{-1/2} + \sigma_2^2 + \ln(N)^{1/2} (N\sigma_1^{K+L})^{-1/2} + \sigma_1^m$$
$$+ \ln(N)/(N\sigma_2^{2r+3}\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^{2r+3} + \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\| \Big).$$

$$\sup_{\tilde{P}_t} \left\| \hat{f}_{P_t}(\tilde{P}_t) - f_{P_t}(\tilde{P}_t) \right\| =$$
$$O_p \Big(\ln(N)^{1/2} (N\sigma_2)^{-1/2} + \sigma_2^2 + \ln(N)^{1/2} (N\sigma_1^{K+L})^{-1/2} + \sigma_1^m$$
$$+ \ln(N)/(N\sigma_2^3\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^3 \Big), \quad \textit{and}$$

$$\sup_{\tilde{P}_t, \tilde{P}_s} \left\| \hat{f}_{t,s}(\tilde{P}_t, \tilde{P}_s) - f_{t,s}(\tilde{P}_t, \tilde{P}_s) \right\| =$$
$$O_p \Big(\ln(N)^{1/2} (N\sigma_2^2)^{-1/2} + \sigma_2^2 + \ln(N)^{1/2} (N\sigma_1^{K+L})^{-1/2} + \sigma_1^m$$
$$+ \ln(N)/(N\sigma_2^4\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^4 \Big).$$

The next two theorems present the convergence results for the algorithm developed in section 5. Convergence of the IBA is stated in the following theorem.

Theorem 6.6. *(Convergence of IBA) Suppose the conditions of Lemma 6.5 hold. Suppose* $\ln(N)^{1/2}(N\sigma_2^3)^{-1/2} \to 0$, $\ln(N)^{1/2}(N\sigma_1^{K+L})^{-1/2} \to 0$, $\ln(N)/(N\sigma_2^4\sigma_1^{K+L}) \to 0$, $\sigma_1^m \to 0$, $\sigma_2^2 \to 0$, *and* $\sigma_1^{2m}/\sigma_2^4 \to 0$. *Then for fixed* $\beta \in \mathcal{B}$, *a solution of the IBA,* $\varphi^*(\beta)$, *exists that is unique with probability tending to one on* $\mathcal{W}$.

Proof. See Appendix A.3. □

As discussed in the introduction, one attractive advantage of the modified backfitting algorithm is that the solution does not depend on the starting values.

Theorem 6.7. *(Convergence of the OBA) Suppose the conditions of Theorem 6.6 hold. Then a solution of the OBA,* $(\hat{\beta}', \hat{\varphi})$*, exists that is unique with probability tending to one on* $\mathcal{B} \times \mathcal{W}$*.*

Proof. See Appendix A.4. □

The key to showing the OBA converges is to note it defines a series of alternating projections between the two compact and convex sets, $x\mathcal{B}$ and $\mathcal{F}_c^N$. Define the distance, **d**, on Θ as follows:

$$\mathbf{d}[(\beta^1, \varphi^1), (\beta^2, \varphi^2)] := \|\beta^1 - \beta^2\| + \sum_{t=1}^{T} \|\varphi_t^1 - \varphi_t^2\|_{s,0}.$$

The next two theorems provide consistency results for $\hat{\theta}$ and uniform convergence results for $\hat{\varphi}$.

Theorem 6.8. *Suppose the conditions of Theorem 6.6 hold. Then* $\mathbf{d}[(\hat{\beta}, \hat{\varphi}), (\beta_0, \varphi_0)] = o_p(1)$*.*

Proof. See Appendix A.5. □

Theorem 6.9. *Suppose the conditions of Lemma 6.5 hold. Then, for* $k = 0, 1, 2$*,* $r = 0, 1$*,*

$k+r \leq 1$, *and* $t = 1, \cdots, T$,

$$\sup_{\beta} \sup_{P_t} \left| \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r} \hat{\varphi}_t(P_t, \beta) - \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r} \varphi_{t0}(P_t, \beta) \right| =$$
$$O_p \Big(\ln(N)^{1/2} (N\sigma_2^{2r+1})^{-1/2} + \sigma_1^m + \ln(N)^{1/2} (N\sigma_1^{K+L})^{-1/2} + \sigma_2^2$$
$$+ \ln(N)/(N\sigma_2^{2r+3}\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^{2r+3} + \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\| \Big).$$

Proof. See Appendix A.6. □

To present the asymptotic distribution of $\hat{\beta}$, additional notations are needed. Let $h_{i0}(P_i) := \partial \Delta \varphi_0(P_i, \beta_0)/\partial \beta - \Delta x_i$, $h_{i0} := h_{i0}(P_{i0})$, $\varphi'_{0t}(P_{it}) := \partial \varphi_{0t}(P_{it}, \beta_0)/\partial P_{it}$, $R(P_i) :=$diag$[(-\varphi'_0(P_{it-1})$ $, \varphi'_0(P_{it})), t = 2, \cdots, T]$, and $R_i := R(P_{0i})$. Recall that $\varepsilon_{it} := y_{it} - P_{it0}$ and $\varepsilon_i := (\varepsilon_{i1}, \cdots, \varepsilon_{iT})'$. The proof of the following theorem is available in the online appendix.

Theorem 6.10. *Suppose the conditions of Theorem 6.6 hold. Let* $\sqrt{N}\ln(N)/(N\sigma_2) \to 0$, $\sqrt{N}\ln(N)/(N\sigma_1^{K+L}) \to 0$, $\sqrt{N}(\ln(N)/(N\sigma_2^3\sigma_1^{K+L})^2 \to 0$, $\sqrt{N}\sigma_1^m \to 0$, $\sqrt{N}\sigma_2^4 \to 0$, $\sqrt{N}(\sigma_1^{2m}/\sigma_2^3)^2 \to 0$, *and* $\|\hat{\Sigma} - \Sigma\| = o_p(N^{-1/4})$. *Then*

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

where

$$V := E[\tau_i h'_{i0} \Sigma^{-1} h_{i0}]^+ E[\tau_i h'_{i0} \Sigma^{-1} R_i \varepsilon_i \varepsilon'_i R'_i \Sigma^{-1} h_{i0}] E[\tau_i h'_{i0} \Sigma^{-1} h_{i0}]^+,$$

and A^+ a the generalized inverse of A (such as the Moore-Penrose generalized inverse), which we use because of the singularity of $E[\tau_i h'_{i0} \Sigma^{-1} h_{i0}]$ that results from the scale normalization of β.

Remark 6.11. The algorithm and asymptotic results of this section are presented under

the assumption that all elements of x_{it} and z_i are continuously distributed. However, with regards to rates of convergence, these results should be treated as worst-case scenarios. Indeed, identification and estimation of the parameters only require at least one of the elements of x_{it} to be continuous. Suppose x_{it} (z_i) contains k_1 (l_1) continuous regressors and $k_2 = K - k_1$ ($l_2 = L - l_1$) discrete regressors. Then, to estimate the predicted outcomes one may construct the multivariate kernel as the product of $k_1 + l_1$ univariate kernels for the continuous regressors and indicator functions for the discrete regressors. The convergence rates obtained in this section would depend on $k_1 + l_1$ instead of $K + L$, and σ_1^{K+L} would be replaced with $\sigma_1^{k_1+l_1}$ in all places where it appears.

Remark 6.12. Obtaining an estimate of the function η may be of interest. Note that under the assumptions of Theorem 3.2, $\eta_0(z_i) = \sum_{t=1}^{T}(\varphi_{t0}(P_{it0}) - x_{it}\beta_0)/T$. Define $\hat{\eta}_i(w_i) = \sum_{t=1}^{T}(\hat{\varphi}_t(\hat{P}_{it}) - x_{it}\hat{\beta})/T$. Then an estimator for $\eta_0(z)$ is defined as follows:

$$\hat{\eta}(z) = \frac{\sum_{i=1}^{N} \tau_i \hat{\eta}_i(w_i) K_{\sigma_1}(z - z_i)}{\sum_{i=1}^{N} \tau_i K_{\sigma_1}(z - z_i)},$$

where K_{σ_1} is defined in section 4. The asymptotic properties of this estimator is beyond the scope of this paper, and is left for future work.

Remark 6.13. As is typical in nonlinear models, the estimator of β_0 itself may not of final interest. Instead, studying $\Phi'_{t0}(x_{it}\beta_0 + \eta_0(z_i))\beta_{k0}$, or $E[\tau_i \Phi'_{t0}(x_{it}\beta_0 + \eta_0(z_i))]\beta_{k0}$, $k = 1, \cdots, K$, where Φ'_t is the derivative of Φ_t may be of more interest. One approach to calculating these marginal effects is to estimate $\eta_0(z_i)$ as in Remark 6.12, invert $\hat{\varphi}_t$ to obtain $\hat{\Phi}_t$, numerically take the derivative of $\hat{\Phi}_t$, and combine these components to compute these effects.

Remark 6.14. As is typical in conventional panel data, the rate of convergence for the estimators does not improve with the imposition of the restriction in 5.2 that $\Phi_t = \Phi$, $t =$

$1, \cdots, T$, because we impose no restrictions on the covariance structure of ε_t, $t = 1, \cdots, T$. However, this restriction may result in better finite sample performance of the estimator due to the weighted averaging of the link functions over time.

7. Estimating the asymptotic variance

To obtain a consistent estimator for the asymptotic variance, V, the obvious plug-in estimator suffices. Define $\hat{h}_i = \partial\Delta\hat{\varphi}(\hat{P}_i, \hat{\beta})/\partial\beta - \Delta x_i$, $\hat{R}_i = \text{diag}[(-\hat{\varphi}'_{t-1}(\hat{P}_{it-1}, \hat{\beta}), \hat{\varphi}'_t(\hat{P}_{it}, \hat{\beta})), t = 2, \cdots, T]$, $\hat{\varepsilon}_{it} = y_{it} - \hat{P}_{it}$, and $\hat{\varepsilon}_i = (\hat{\varepsilon}_{i1}, \cdots, \hat{\varepsilon}_{iT})'$. Define $V_1 = E[\tau_i h'_{i0}\Sigma^{-1}h_{i0}]$ and $V_2 = E[\tau_i h'_{i0}\Sigma^{-1}R_i\varepsilon_i\varepsilon'_i R'_i\Sigma^{-1}h_{i0}]$ so that $V = V_1^+ V_2 V_1^+$. Define the sample analogs $\hat{V}_1 := \sum_{i=1}^N \tau_i \hat{h}'_i \hat{\Sigma}^{-1}\hat{h}_i/N$, $\hat{V}_2 := \sum_{i=1}^N \tau_i \hat{h}'_i \hat{\Sigma}^{-1}\hat{R}_i\hat{\varepsilon}_i\hat{\varepsilon}'_i\hat{R}'_i\hat{\Sigma}^{-1}\hat{h}_i/N$, and $\hat{V} := \hat{V}_1^+\hat{V}_2\hat{V}_1^+$.

Theorem 7.1. *Suppose the conditions of Theorem 6.10 hold. Suppose* $\ln(N)/(N\sigma_2^3) \to 0$, $\ln(N)/(N\sigma_2^5\sigma_1^{K+L}) \to 0$, *and* $\sigma_1^{2m}/\sigma_2^5 \to 0$. *Then* $\hat{V} \xrightarrow{p} V$.

Proof. See Appendix A.7. □

Remark 7.2. Expressions for $\partial\Delta\hat{\varphi}(\hat{P}_i, \hat{\beta})/\partial\beta$ and $\hat{\varphi}'(\hat{P}_i, \hat{\beta})$ are needed to compute the $\hat{V}$. $\partial\Delta\hat{\varphi}(P, \hat{\beta})/\partial\beta$ can be computed numerically by solving the IBA for φ^* on the grid of P, at nearby values of β. Likewise $\hat{\varphi}'(P, \hat{\beta})$ can be computed numerically on the grid of P. For $\hat{P}_i$ off the grid of P, $\partial\Delta\hat{\varphi}(\hat{P}_i, \hat{\beta})/\partial\beta$ and $\hat{\varphi}'(\hat{P}_i, \hat{\beta})$ can be computed by numerical interpolation.

7.1. Optimal weighting matrix

If it is assumed that $E[\tau_i R_i\varepsilon_i\varepsilon'_i R'_i|w] = E[\tau_i R_i\varepsilon_i\varepsilon'_i R'_i]$; then the choice of the weighting matrix that minimizes the asymptotic variance V is $\Sigma_0 = E[\tau_i R_i\varepsilon_i\varepsilon'_i R'_i]$. The asymptotic

variance then becomes $V = V_1^+ = E[\tau_i h_{i0}' \Sigma_0^{-1} h_{i0}]^+$, and as shown in the proof of Theorem 7.1, $\hat{V}_1 \xrightarrow{p} V_1$ under the assumptions of Theorem 7.1.

Theorem 6.10 requires that any consistent estimator for Σ converges at the rate $N^{1/4}$. Let $\tilde{\beta}$ be an initial $\sqrt{N}$-consistent estimator of β_0, obtained with an initial weighting matrix, such as the $(T-1)$-dimensional identity matrix. Also, let $\tilde{\varphi}_t'(\hat{P}_{it}, \tilde{\beta})$ be the estimated derivative of the link functions with respect to P_t, evaluated at $(\tilde{\beta}, \hat{P}_{it})$, and computed as suggested in Remark 7.2. Let $\tilde{R}_i := \text{diag}[(-\tilde{\varphi}_{t-1}'(\hat{P}_{it-1}, \tilde{\beta}), \tilde{\varphi}_t'(\hat{P}_{it}, \tilde{\beta})),\ t = 2, \cdots, T]$. Then the proposed estimator for Σ is defined as $\hat{\Sigma} = \sum_{i=1}^N \tau_i \tilde{R}_i \hat{\varepsilon}_i \hat{\varepsilon}_i' \tilde{R}_i' / N$. The following theorem outlines the conditions under which $\hat{\Sigma}$ obtains the required rate of convergence.

Theorem 7.3. *Suppose (i)* $\|\tilde{\beta} - \beta_0\| = O_p(N^{-1/2})$, *(ii) the conditions of Theorem 7.1 hold, and (iii)* $\sqrt{N}\ln(N)/(N\sigma_2^3) \to 0$, $N^{1/4}\ln(N)/(N\sigma_2^5\sigma_1^{K+L}) \to 0$, *and* $N^{1/4}\sigma_1^{2m}/\sigma_2^5 \to 0$. *Then*

$$\left\|\hat{\Sigma} - \Sigma\right\| = o_p(N^{-1/4}).$$

Proof. See Appendix A.8. □

The conditions imposed on the bandwidths and kernels in Theorem 7.3 are stronger than the conditions imposed in Theorem 6.10. In particular, the bandwidth and kernel conditions imposed in Theorem 7.3 imply greater under-smoothing and higher order kernels than the conditions imposed in Theorem 6.10. Using the same bandwidths to compute the estimators for β_0 and Σ may result in poor finite sample performance of the estimator for β_0. In this case, one may use different bandwidths and kernels to compute $\hat{\beta}$ and $\hat{\Sigma}$.

We now have all the components to propose a two-step procedure similar to that of the two-step efficient GMM estimator. The first stage replaces $\hat{\Sigma}$ in equation (4.2) with the identity matrix to obtain an initial consistent estimator for β_0, denoted by $\tilde{\beta}$. Given $\tilde{\beta}$,

compute $\tilde{\varphi}(P,\tilde{\beta})$ by implementing the IBA. Compute the derivatives $\tilde{\varphi}'(\hat{P}_{it},\tilde{\beta})$ as suggested in Remark 7.2, and construct $\tilde{R}_i$. Next, use $\tilde{R}_i$ and $\hat{\varepsilon}_i$ so construct $\hat{\Sigma}$. The second stage uses $\hat{\Sigma}$ equation (4.2) to compute the optimally-weighted estimator $\hat{\theta} = (\hat{\beta}', \hat{\varphi})$.

8. Experimental Evidence

In this section, we examine the small sample properties of the estimator via Monte Carlo experiments. We investigate two model specifications. The first (Design 1) examines the performance of the estimator for a static, panel data discrete-choice model with heteroskedasticity over time. The second (Design 2) considers a static, panel data model with continuous outcomes and nonlinear link functions, again with heteroskedasticity over time. In both experiments, we perform 100 Monte Carlo replications with three sample sizes: 200, 400, and 800. For each sample size, we calculate the mean bias and the root mean squared error (RMSE) for the estimator with the weighting matrix equal to the identity, and also with the optimally-weighted estimator, denoted by KLS and OKLS respectively. For comparison, we also report the mean bias and the RMSE for the estimator proposed in Newey (1994a), where the link functions are assumed to be known (denoted by Newey).[5] For the OKLS estimator, we further report the simulated standard errors (Sim. se), and the average of the estimated asymptotic standard errors (E-A se.) of the estimates of β_0. Additionally, for the OKLS estimator, and for each $\beta_{k0}, k = 1, \cdots, K$, we report the simulated probability of committing a type one error (Size), which is defined as the simulated average of the indicator function equal to one if β_{k0} lies outside the 95% confidence interval constructed using the estimated coefficient and its corresponding estimated asymptotic

[5]Because of its poor performance, the results of the optimally-weighted version of this estimator are not reported.

standard error. Figures 1 and 2 show the average behavior of the OLKS estimator of the link functions for designs on and two respectively. They show the mean (in a — line), median (in a – – line), 25th and 75th percentiles (in – · – lines), for T=3, and N= 200, 400, and 800, respectively when T=3.

In both simulation exercises, we set $K = 2$, $L = 1$, and $\beta_0 = (0.6, 0.8)'$. We use the optimal sixth-order multiplicative Epanechnikov kernel to compute the estimates of the predicted outcomes with bandwidth equal to $c_1 N^{-1/13}$. To compute the trimming function, we independently sample from the data-generating process $\bar{N} = 2000$ realizations of the regressors, and construct the estimated kernel density using the Gaussian kernel with the Silverman rule-of-thumb bandwidth. We then construct the trimming function to trim away observations that fall within the lowest five percent of the empirical pdf. This strategy is consistent with the suggestion of Newey (1994b). We also use this auxiliary sample to compute a sample of predicted outcomes, $\bar{P}_{it}$. To compute the estimates of the link functions, we select a grid of $ng = 100$ equidistant points between $[L_t, U_t]$, $t = 1, \cdots, T$ where

$$L_t = \min_{i \leq \bar{N}} \{\bar{P}_{it} : \tau_i \neq 0,\ \bar{F}_{P_t}(\bar{P}_{it}) \geq 0.025\},$$

$$U_t = \max_{i \leq \bar{N}} \{\bar{P}_{it} : \tau_i \neq 0,\ \bar{F}_{P_t}(\bar{P}_{it}) \leq 0.975\},$$

and $\bar{F}_{P_t}$ is the empirical CDF of $\bar{P}_{it}$. In practice, the investigator may not have a large auxiliary sample to perform this trimming strategy. In this case, one may implement the standard method of trimming away a predetermined percentage of the observations from the tails of each regressor. Preliminary analysis of this method using 2.5% trimming suggests the performance of the estimators under these two alternative trimming methods are

comparable.

To compute the estimates of the link functions, we use the Gaussian kernel truncated on $[-2,2]$ with bandwidth equal to $c_2 N^{-1/6}$. We choose the scaling factor c_1 to be equal to the standard deviations of the regressors using the auxiliary sample, and choose c_2 to be equal to the standard deviations of the corresponding predicted probabilities. The convergence criterion of the IBA is

$$\max_{t\leq T}\max_{g\leq ng}\frac{\left|\varphi_t^{*[s]}(P_{gt})-\varphi_t^{*[s-1]}(P_{gt})\right|}{1+\left|\varphi_t^{*[s-1]}(P_{gt})\right|}<\xi_1,$$

where $\xi_1 = 1E-5$ and the convergence criterion for the OBA is

$$\max_{k\leq K}\frac{\left|\breve{\beta}_k^{[s]}-\breve{\beta}_k^{[s-1]}\right|}{1+\left|\breve{\beta}_k^{[s-1]}\right|}<\xi_2,$$

where $\xi_2 = 1E-6$. As discussed in section 7.1, the estimates, $\tilde{\varphi}_t'(\hat{P}_{it},\tilde{\beta})$, are needed to compute the estimate of the optimal weighting matrix. To obtain these estimates, we compute $\tilde{\varphi}_t(P_t,\tilde{\beta})$ using the first state estimates of β_0, the twelfth-order multiplicative Epanechnikov kernel with bandwidth $c_1 N^{-0.039}$ and c_1 is computed as above, and the Gaussian kernel truncated on $[-2,2]$ with bandwidth equal to $c_2 N^{-0.1255}$ and c_2 is also computed as above. We then compute the numerical derivative of $\tilde{\varphi}_t(\hat{P}_{it},\tilde{\beta})$ on the grid for P, and compute the derivatives at $\hat{P}_{it}$ of off the grid by linear interpolation. The code is written in FORTRAN 90 and executed on a UNIX workstation. On average, estimation of both the first stage and optimally weighted parameters take 1.5 seconds (0.75 seconds per estimator) for N=200, 1.9 seconds (0.95 seconds per estimator) for N=400, and 3.5 seconds (1.75 seconds per estimator) for N=800.

8.1. Design 1: Static panel data discrete choice

For the first simulation exercise, consider the following model:

$$y_{it} = 1\{\beta_1 x_{it1} + \beta_2 x_{it2} + \eta(z_i) + u_{it} > 0\}, \quad i = 1, \cdots, N, \quad t = 1, 2, 3,$$

where $z_i = 0.6z_{i1} + 0.4(x_{i12} + x_{i22} + x_{i32})/3$. Here, x_{it1}, x_{it2}, z_{i1}, and u_i are mutually independent with x_{it1}, x_{it2}, and z_{i1} distributed $N(0,1)$, and u_{it} distributed $N(0, 0.4 + 0.3(t-1))$. The individual-specific function is given by

$$\eta(z_i) = \frac{\exp(z_i)}{1 + \exp(z_i)} - 0.5.$$

To simulate the model where the link function is known to be the normal CDF, additional trimming is necessary to ensure the predicted probabilities remain in the unit interval. In this exercise, we restrict the sample so that $\hat{P}_{it} \in [0.001, 0.999]$.

8.2. Design 2: Static panel data continuous-outcome model

Consider the following data-generating process:

$$y_{it} = \Phi_t(x_{1it}\beta_1 + x_{2it}\beta_2 + \eta(z_i)) + \varepsilon_{it}, \quad i = 1, \cdots, N, \quad t = 1, 2, 3,$$

where, once more, $z_i = 0.6z_{i1} + 0.4(x_{i12} + x_{i22} + x_{i32})/3$. In this exercise, x_{it1}, x_{it2}, z_{i1} and ε_{it} are mutually independent with x_{it1}, x_{it2}, and z_{i1} distributed $U[-10,\ 10]$, and ε_{it} distributed $N(0,\ 1)$. The link functions are given by

$$\Phi_t(x) \quad = \quad (\sqrt{t} \cdot x + 3 \cdot x^2 + 4\sqrt{t} \cdot x^3)/1000.$$

Table 1: Simulation results for the static panel data discrete-choice model.

	β_1			β_2		
	N=200					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	0.0027	-0.0047	-0.0094	-0.0051	-0.0000	0.0027
RMSE	0.0560	0.0607	0.0668	0.0417	0.0445	0.0499
Sim. se			0.0605			0.0661
E-A se			0.0231			0.0228
Size			0.2100			0.4400
	N=400					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	0.0059	-0.0027	-0.0037	-0.0058	-0.0041	-0.0005
RMSE	0.0384	0.0461	0.0477	0.0300	0.0349	0.0365
Sim. se			0.0477			0.0365
E-A se			0.0190			0.0188
Size			0.2800			0.4000
	N=800					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	0.0074	0.0038	-0.0079	-0.0063	-0.0040	0.0050
RMSE	0.0263	0.0335	0.0318	0.0202	0.0251	0.0230
Sim. se			0.0308			0.0230
E-A se			0.0099			0.0098
Size			0.1600			0.2100

Table 2: Simulation results for the static continuous-outcome panel data model.

	β_1			β_2		
	N=200					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	-0.0151	0.0197	0.0187	0.0090	-0.0167	-0.0157
RMSE	0.0500	0.0431	0.0411	0.0354	0.0344	0.0329
Sim. se			0.0367			0.0289
E-A se			0.0025			0.0025
Size			0.0600			0.0700
	N=400					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	-0.0064	0.0184	0.0175	0.0036	-0.0152	-0.0142
RMSE	0.0351	0.0372	0.0321	0.0255	0.0295	0.0254
Sim. se			0.0268			0.0210
E-A se			0.0024			0.0025
Size			0.0500			0.0700
	N=800					
	Newey	KLS	OKLS	Newey	KLS	OKLS
Mean Bias	0.0074	-0.0155	0.0150	-0.0063	-0.0123	-0.0118
RMSE	0.0263	0.0261	0.0226	0.0202	0.0206	0.0176
Sim. se			0.0168			0.0131
E-A se			0.0024			0.0024
Size			0.0700			0.0900

The individual-specific function is given by

$$\eta(z_i) = 6\left(\frac{\exp(z_i)}{1+\exp(z_i)} - 0.5\right).$$

The simulation exercises imply the following about the finite sample performance of our estimators. The estimators performs well in small samples in terms of recovering the finite dimensional parameters. The estimators of the finite-dimensional parameters are indeed $\sqrt{N}$-consistent. Estimation of the finite-dimensional parameters is adaptive with respect to estimation of the link functions in that the RMSE of $\hat{\beta}$ approaches the RMSE of the estimator for β_0 when link functions are known. The OKLS estimator produces

the correct size corresponding to the test that $\beta_{0k},\ k = 1,2$ is true in continuous outcome models. However, while they converge to the correct size, the probability of incorrectly rejecting the null that β_0 is true is relatively large in discrete choice models. Finally, the OKLS estimator does a better job recovering the link functions in continuous outcome models than in discrete choice models.

9. Conclusion

This paper investigates identification and estimation of a class of panel data single-index models with semiparametric individual-specific effects, which includes a semiparametric discrete-choice panel model with unconditional heteroskedastic errors. We develop a general estimator of the finite- and infinite-dimensional parameters of interest. This estimator combines and extends the semiparametric profile likelihood method of Severini and Wong (1992) with the modified backfitting algorithm of Mammen et al. (1999) and Mammen et al. (2001) to the panel data framework. The full algorithm is composed of an inner backfitting and an outer backfitting algorithm. We provide sufficient conditions for convergence of the algorithm.

We derive uniform rates of convergence results for the estimators for the unknown link functions, and show the estimators of the finite-dimensional parameters are $\sqrt{N}$-consistent with a Gaussian limiting distribution. We propose a consistent estimator for the asymptotic variance of the finite-dimensional parameters and an estimator for the optimal weighting matrix that converges at the rate required for $\sqrt{N}$-consistency of the optimally weighted estimator. Because the finite-dimensional parameters may not be of final interest, we propose estimators of the partial effects and the average partial effects of the explanatory variables. Derivation of the asymptotic properties of these estimators is involved and therefore left for future work.

We investigate the finite sample performance of the estimator and find it performs satisfactorily in terms of speed and accuracy.

Appendix A. LEMMA AND THEOREMS

Appendix A.1. Proof of Theorem 3.2

Proof. Equations (2.3) and (3.1) imply that

$$\begin{aligned} \varphi_{t0}^{-1}(x_{it}\beta_0+\eta_0(z_i)) &= \varphi_{t*}^{-1}(x_{it}\beta_*+\eta_*(z_i)) \Leftrightarrow \\ x_{it}\beta_0+\eta_0(z_i) &= \varphi_{t0}(\varphi_{t*}^{-1}(x_{it}\beta_*+\eta_*(z_i))). \end{aligned} \tag{A.1}$$

Strict monotonicity of the link functions imply they are differentiable almost everywhere. Differentiating equation (A.1) with respect to the continuous regressor x_{itK} gives

$$a := \frac{\beta_{0K}}{\beta_{*K}} = \frac{\varphi'_{t0}(\varphi_{t*}^{-1}(x_{it}\beta_*+\eta_*(z_i)))}{\varphi'_{t*}(\varphi_{t*}^{-1}(x_{it}\beta_*+\eta_*(z_i)))} > 0, \tag{A.2}$$

where the positive sign follows from the assumption the link functions are strictly increasing. We have from equation (A.2) that $\varphi'_{t0}(P_{it0}) = a\varphi'_{t*}(P_{it0})$, which implies

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

Noting equation (2.4) also holds for π_*, and taking the first difference of equation (A.3) obtains the following for $t = 2, \cdots, T$:

$$\begin{aligned} \Delta x_{it}\beta_0 &= a\Delta x_{it}\beta_* + \Delta c_t \Rightarrow \\ (1,\Delta x_{it})\gamma &= 0 \Rightarrow \qquad \text{(A.4)} \\ E[(1,\Delta x_{it})'(1,\Delta x_{it})]\gamma &= 0, \qquad \text{(A.5)} \end{aligned}$$

where $\gamma := (\Delta c_t, (\beta_0 - a\beta_*)')'$. Under Assumption 3.1.2, equation (A.5) implies $\gamma = 0$ so that $\beta_0 = a\beta_*$ and $\Delta c_t = 0,\ t = 2, \cdots, T$, the latter implying $c_t = c,\ t = 1, \cdots, T$.

Substituting these equalities into equation (A.3) gives

$$x_{it}\beta_0 + \eta_0(z_i) = x_{it}(a\beta_*) + a\eta_*(z_i) + c \Rightarrow \tag{A.6}$$

$$\eta_0(z_i) = a\eta_*(z_i) + c, \tag{A.7}$$

which proves the second part of the theorem. The assumption that $\|\beta_0\| = \|\beta_*\| = 1$ implies from $\beta_0 = a\beta_*$ that $|a| = 1$. But $a > 0$, which implies $a = 1$. Also, equation (A.3) and the assumption $E[\tau_i\varphi_{10}(P_{i10})] = E[\tau_i\varphi_{1*}(P_{i10})] = 0$ imply $c = 0$. □

Appendix A.2. Proof of Theorem 5.1

Proof. Using theorem 1.3.2 of Robertson et al. (1988) we can show that if $\bar{\varphi}(\beta)$ minimizes $\|\Delta\varphi^*(\beta) - \Delta\tilde{\varphi}\|_T^2$ over $\tilde{\varphi} \in \mathcal{F}_c^N$, then for $t = 1, \cdots, T$,

$$\int \bar{\varphi}_t(P_t, \beta)[\varphi_t^*(P_t, \beta) - \bar{\varphi}_t(P_t, \beta)]\,\hat{f}_{P_t}(P_t)dP_t + \sum_{s \neq t} \frac{\hat{\sigma}_*^{st}}{\hat{\sigma}_*^{tt}} \int [\varphi_s^*(P_s, \beta) - \bar{\varphi}_s(P_t, \beta)]\,\hat{f}_{P_s}(P_s)dP_s = 0. \tag{A.8}$$

For $t = 1, \cdots, T$, theorem 1.3.2 of Robertson et al. (1988) implies $\int \breve{\varphi}_t(P_t, \beta)[\varphi_t^*(P_t, \beta) - \breve{\varphi}_t(P_t, \beta)]\,\hat{f}_{P_t}(P_t)dP_t = 0$, and theorem 1.3.3 of Robertson et al. (1988) implies $\int [\varphi_t^*(P_t, \beta) - \breve{\varphi}_t(P_t, \beta)]\,\hat{f}_{P_t}(P_t)dP_t = 0$. Substituting $\breve{\varphi}$ for $\bar{\varphi}$ in the left-hand side of equation (A.8) and using these conditions on $\breve{\varphi}$ show that $\breve{\varphi}$ also minimizes $\|\Delta\varphi^*(\beta) - \Delta\tilde{\varphi}\|_T^2$ over $\tilde{\varphi} \in \mathcal{F}_c^N$. □

Appendix A.3. Proof of Theorem 6.6

Proof. Under the conditions of the theorem, along with the use of the product kernel, Assumptions (A1) and (A2) of Mammen et al. (1999) are verified. Also, under the conditions of the theorem, and by the same arguments in the proof of Lemma 6.5, it can be shown that for $s,t = 1,\cdots,T$,

$$\sup_{P_t,\beta} \|\hat{m}_{st}(P_t,\beta) - m_{st0}(P_t,\beta)\| = o_p(1).$$

Also, $E[\|m_{st0}(P_t,\beta)\|^2] < \infty,\ s,t = 1,\cdots,T$ so that Assumption (A3) of Mammen et al. (1999) is verified. Then, on the set $\mathcal{W}$, result follows from the first part of Theorem 1 in Mammen et al. (1999). □

Appendix A.4. Proof of Theorem 6.7

Proof. The OBA defines a series of alternating projections between two sets $x\mathcal{B}$ and $\mathcal{F}_c^N$. For the projection of $\Delta x\beta$ onto the set $\mathcal{F}_c^N$, denote the corresponding projector as $\mathcal{T}_{\mathcal{F}_c^N}$. For the projection of $\Delta\varphi$ onto $x\mathcal{B}$, denote the corresponding projector as $\mathcal{T}_{x\mathcal{B}}$. This notation shows 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 that $\mathcal{F}_c^N$ and $\mathcal{B}$ are compact and convex sets, and given the scale normalization and Given a solution of the IBA, $\varphi^*(\beta)$, exists that is unique with probability tending to one on $\mathcal{W}$, Theorem 4 of Cheney and Goldstein (1959) shows the sequence $\mathcal{T}^n b$ converges in probability to a fixed point on $\mathcal{W}$ as n tends to infinity. □

Appendix A.5. Proof of Theorem 6.8

Proof. From the conditions on the bandwidths in the theorem and from Theorem 6.5, we have

$$\sup_{\beta,\varphi} |\hat{Q}(\beta,\varphi) - Q_0(\beta,\varphi)| \leq \sup_{P_t,\beta,\varphi} |\hat{Q}(P_t,\beta,\varphi) - Q_0(P_t,\beta,\varphi)| = o_p(1). \tag{A.9}$$

Because $\hat{\theta}$ is the minimizer of $\hat{Q}(\theta)$ and $Q_0(\theta_0) = 0$, equation (A.9) implies

$$\begin{aligned} 0 &\leq \hat{Q}(\hat{\theta}) \leq \hat{Q}(\theta_0) - Q_0(\theta_0) \\ &\leq \sup_{\beta,\varphi} |\hat{Q}(\beta,\varphi) - Q_0(\beta,\varphi)| = o_p(1). \end{aligned} \tag{A.10}$$

Also, equations (A.9) and (A.10) imply

$$\begin{aligned} 0 &\leq Q_0(\hat{\theta}) = Q_0(\hat{\theta}) - \hat{Q}(\hat{\theta}) + \hat{Q}(\hat{\theta}) \\ &\leq \sup_{P_t,\beta,\varphi} |\hat{Q}(P_t,\beta,\varphi) - Q_0(P_t,\beta,\varphi)| + \hat{Q}(\hat{\theta}) = o_p(1). \end{aligned} \tag{A.11}$$

Because the model is identified, for all $\delta > 0$, $\varepsilon > 0$ exists such that $\mathbf{d}[(\beta,\varphi),(\beta_0,\varphi_0)] > \delta \Rightarrow Q_0(\beta,\varphi) > \varepsilon$, which implies $\Pr\{\mathbf{d}[(\hat{\beta},\hat{\varphi}),(\beta_0,\varphi_0)] > \delta\} \leq \Pr\{(Q_0(\hat{\beta},\hat{\varphi}) > \varepsilon\} \longrightarrow 0$. Hence $\mathbf{d}[(\hat{\beta},\hat{\varphi}),(\beta_0,\varphi_0)] \xrightarrow{p} 0$ □

Appendix A.6. Proof of Theorem 6.9

Proof. For $t = 1, \cdots, T$, define $\hat{Q}_{\varphi_t}(P_t, \beta, \varphi(\beta)) = \partial \hat{Q}(P_t, \beta, \varphi(\beta))/\partial \varphi_t$. Define $Q_{\varphi_t,0}(P_t, \beta, \varphi(\beta))$ analogously. Because $\hat{Q}_{\varphi_t}(P_t, \beta, \varphi^*(\beta)) = Q_{\varphi_t,0}(P_t, \beta, \varphi_0(\beta)) = 0$, $t = 1, \cdots, T$,

$$\begin{aligned} 0 &= Q_{\varphi_t,0}(P_t, \beta, \varphi_0(\beta)) - \hat{Q}_{\varphi_t}(P_t, \beta, \varphi^*(\beta)) \\ &= Q_{\varphi_t,0}(P_t, \beta, \varphi^*(\beta)) - \hat{Q}_{\varphi_t}(\beta, \varphi^*(\beta)) + Q_{\varphi_t\varphi_t,0}(\beta, \overline{\varphi}(\beta))((\varphi_{t0}(\beta) - \varphi_t^*(\beta)), \end{aligned}$$

where $Q_{\varphi_t\varphi_s,0}(\beta, \varphi(\beta)) := \partial^2 Q_0(\beta, \varphi(\beta))/\partial \varphi_t \partial \varphi_s$, $t, s = 1, \cdots, T$. By noting that $Q_{\varphi_t\varphi_t,0}(\beta, \overline{\varphi}(\beta)) = 2\sigma^{tt} f_{P_t}(P_t)$, and $\inf_{P_t} f_{P_t}(P_t) > 0$, we have

$$\begin{aligned} &\sup_{\beta, P_t} \left\| \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r}((\varphi_{t0}(P_t, \beta) - \varphi_t^*(P_t, \beta)) \right\| \leq C \sup_{\beta, P_t \varphi} \left\| \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r}(Q_{\varphi_t,0}(P_t, \beta, \varphi) - \hat{Q}_{\varphi_t}(\beta, \varphi)) \right\|, \\ &= O_p \Big(\ln(N)^{1/2}(N\sigma_2^{2r+1})^{-1/2} + \sigma_1^m + \ln(N)^{1/2}(N\sigma_1^{K+L})^{-1/2} + \sigma_2^2 \\ &+ \ln(N)/(N\sigma_2^{2r+3}\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^{2r+3} + \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\| \Big). \end{aligned} \tag{A.12}$$

Because projection onto $\mathcal{S}_{\mathcal{K}}$ is a distance reducing operator and $\varphi_{t0} \in \mathcal{S}_{\mathcal{K}}$, Theorem 6.8, equation (A.12), and the triangular inequality obtain

$$\begin{aligned} &\sup_{\beta, P_t} \left\| \frac{\partial^{k+r}}{\partial \beta^k \partial P_t^r}((\breve{\varphi}_t(P_t, \beta) - \varphi_{t0}(P_t, \beta)) \right\| \\ &= O_p \Big(\ln(N)^{1/2}(N\sigma_2^{2r+1})^{-1/2} + \sigma_1^m + \ln(N)^{1/2}(N\sigma_1^{K+L})^{-1/2} + \sigma_2^2 \\ &+ \ln(N)/(N\sigma_2^{2r+3}\sigma_1^{K+L}) + \sigma_1^{2m}/\sigma_2^{2r+3} + \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\| \Big). \end{aligned}$$

Thus

$$\sup_{\beta,P_t}\left\|\frac{\partial^{k+r}}{\partial\beta^k\partial P_t^r}((\hat{\varphi}_t(P_t,\beta)-\varphi_{t0}(P_t,\beta))\right\|$$
$$=O_p\left(\ln(N)^{1/2}(N\sigma_2^{2r+1})^{-1/2}+\sigma_1^m+\ln(N)^{1/2}(N\sigma_1^{K+L})^{-1/2}+\sigma_2^2\right.$$
$$\left.+\ln(N)/(N\sigma_2^{2r+3}\sigma_1^{K+L})+\sigma_1^{2m}/\sigma_2^{2r+3}+\left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\|\right).$$

□

Appendix A.7. Proof of Theorem 7.1

Proof. Define $u_i := h_{0i}'\Sigma^{-1}R_i\varepsilon_i$ and $\hat{u}_i := \hat{h}_i'\hat{\Sigma}^{-1}\hat{R}_i\hat{\varepsilon}_i$ so that

$$\|\hat{V}_2-V_2\|\leq\left\|\frac{1}{N}\sum_{i=1}^N\tau_i[\hat{u}_i\hat{u}_i'-u_iu_i']\right\|+\left\|\frac{1}{N}\sum_{i=1}^N\tau_iu_iu_i'-E[\tau_iu_iu_i']\right\|.$$

We use some key results in the proof of the theorem. In what follows, the supremum is taken for (P,β) in the $o(1)$ neighborhood of (P_0,β_0). First,

$$\begin{aligned}\|\hat{h}_i-h_{0i}\| &\leq \sup_{P,\beta}\left\|\frac{\partial}{\partial\beta}\hat{\varphi}(P,\beta)-\frac{\partial}{\partial\beta}\varphi_0(P,\beta)\right\|+\sup_{P,\beta}\left\|\frac{\partial^2}{\partial\beta^2}\varphi_0(P,\beta)\right\|\|\hat{\beta}-\beta_0\|\\ &+ \sup_{P,\beta}\left\|\frac{\partial^2}{\partial\beta\partial P}\varphi_0(P,\beta)\right\|\|\hat{P}-P_0\|_{s,0}=o_p(N^{-1/4}).\end{aligned}$$

Second, under the conditions of the theorem, $\sup_{P,\beta}\|\hat{\varphi}'(P,\beta)-\varphi_0'(P,\beta)\|=o_p(1)$, so that

$$\begin{aligned}\|\hat{R}_i-R_i\| &\leq C\|\hat{\varphi}'(\hat{P}_i,\hat{\beta})-\varphi_0'(P_{i0},\beta_0)\|\\ &\leq C_1\sup_{P,\beta}\|\hat{\varphi}'(P,\beta)-\varphi_0'(P,\beta)\|+\sup_{P,\beta}\left\|\frac{\partial^2}{\partial P\partial\beta}\varphi_0'(P,\beta)\right\|\|\hat{\beta}-\beta_0\|\\ &+ \sup_{P,\beta}\left\|\frac{\partial^2}{\partial P^2}\varphi_0(P,\beta)\right\|\|\hat{P}-P_0\|_{s,0}=o_p(1).\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\hat{u}_i - u_i\| &\leq \|\hat{h}_i - h_{0i}\|\|\hat{\Sigma}^{-1}\|\|\hat{R}_i\|\|\hat{\varepsilon}_i\| + \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|\|h_{0i}\|\|\hat{R}_i\|\|\hat{\varepsilon}_i\| \\
&+ \|\hat{R}_i - R_i\|\|h_{0i}\|\|\Sigma^{-1}\|\|\hat{\varepsilon}_i\| + \|\hat{\varepsilon}_i - \varepsilon_i\|\|h_{0i}\|\|\Sigma^{-1}\|\|R_i\| \\
&\leq b(y_i, w_i)\left(\|\hat{h}_i - h_{0i}\| + \|\hat{\Sigma}^{-1} - \Sigma^{-1}\| + \|\hat{R}_i - R_i\| + \|\hat{P}_i - P_{i0}\|\right),
\end{aligned}$$

where $b(y_i, w_i) = \|\hat{\Sigma}^{-1}\|\|\hat{R}_i\|\|\hat{\varepsilon}_i\| + \|h_i\|\|\hat{R}_i\|\|\hat{\varepsilon}_i\| + \|h_i\|\|\Sigma^{-1}\|\|\hat{\varepsilon}_i\| + \|h_i\|\|\Sigma^{-1}\|\|R_i\|$. All estimators included in $b(y_i, w_i)$ converge to their population counterparts faster than N on $\mathcal{W}$. This result, boundedness of the true functions on $\mathcal{W}$, and the moment condition on y_i obtains $E[\tau_i b(y_i, w_i)^2] = O_p(1)$, so that $\tau_i\|\hat{u}_i - u_i\|^2 = o_p(1)$. Thus, using these results, and that $E[\tau_i\|u_i\|^2] < \infty$, we have

$$\begin{aligned}
\left\|\frac{1}{N}\sum_{i=1}^{N}\tau_i[\hat{u}_i\hat{u}_i' - u_i u_i']\right\| &\leq \frac{1}{N}\sum_{i=1}^{N}\tau_i\left\|\hat{u}_i\hat{u}_i' - u_i u_i'\right\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\tau_i\|\hat{u}_i - u_i\|^2 + 2\left(\frac{1}{N}\sum_{i=1}^{N}\tau_i\|u_i\|^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\tau_i\|\hat{u}_i - u_i\|^2\right)^{1/2} \\
&= o_p(1),
\end{aligned}$$

and by the WLLN, $\|\sum_{i=1}^{N}(\tau_i u_i u_i' - E[\tau_i u_i u_i'])/N\| = o_p(1)$, obtaining $\|\hat{V}_2 - V_2\| = o_p(1)$. Similar calculations show $\|\hat{V}_1 - V_1\| = o_p(1)$, so by the continuous mapping theorem and the triangular inequality, $\hat{V} = V + o_p(1)$. □

Appendix A.8. Proof of Theorem 7.3

Proof. Note,

$$\|\tilde{\varphi}_t'(\hat{P}_{it},\tilde{\beta})-\varphi_0'(P_{it0},\beta_0)\| \leq \sup_{P_t,\beta}\|\tilde{\varphi}'(P_t,\beta)-\varphi_0'(P_t,\beta)\|+\sup_{P_t,\beta}\left\|\frac{\partial^2}{\partial P_t\partial\beta}\varphi_0'(P_t,\beta)\right\|\|\tilde{\beta}-\beta_0\|$$
$$+\sup_{P_t,\beta}\left\|\frac{\partial^2}{\partial P^2}\varphi_0(P_t,\beta)\right\|\|\hat{P}_t-P_{t0}\|_{s,0},$$

so that, under the conditions of the theorem, $|\tilde{\varphi}_t'(\hat{P}_{it},\tilde{\beta})-\varphi_{t0}'(P_{it0})| = o_p(n^{-1/4})$. Also, because $\sum_{i=1}^N \tau_i\varphi_{t0}'(P_{it0})\varepsilon_{it}\varphi_{s0}'(\hat{P}_{is}-P_{is0})/N$ and $\sum_{i=1}^N \tau_i\varphi_{t0}'(P_{it0})\varepsilon_{it}\varepsilon_{is}(\tilde{\varphi}_s'(\hat{P}_{is},\tilde{\beta})-\varphi_{s0}'(P_{is0}))/N$, $t,s=1,\cdots,T$ are the leading terms in

$$\sum_{i=1}^N \tau_i[\tilde{\varphi}_t'(\hat{P}_{it},\tilde{\beta})\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}\tilde{\varphi}_s'(\hat{P}_{is},\tilde{\beta})-\varphi_{t0}'(P_{it0})\varepsilon_{it}\varepsilon_{is}\varphi_{s0}'(P_{is0})]/N,$$

under the conditions of the theorem, it can be shown that

$$\left|\frac{1}{N}\sum_{i=1}^N \tau_i[\tilde{\varphi}_t'(\hat{P}_{it},\tilde{\beta})\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}\tilde{\varphi}_s'(\hat{P}_{is},\tilde{\beta})-\varphi_{t0}'(P_{it0})\varepsilon_{it}\varepsilon_{is}\varphi_{s0}'(P_{is0})]\right| = o_p(n^{-1/4}).$$

From this result, we conclude $\|\sum_{i=1}^N \tau_i(\tilde{R}_i\hat{\varepsilon}_i\hat{\varepsilon}_i'\tilde{R}_i'-R_i\varepsilon_i\varepsilon_i'R_i')/N\| = o_p(n^{-1/4})$.

Also, by $E[\tau_i\|R_i\varepsilon_i\varepsilon_i'R_i'\|^2] \leq E[\tau_i\|R_i\|^4\|\varepsilon_i\|^4] \leq C_1+C_2E[\tau_i\|y_i\|^4] < \infty$, by iid data, and the Lindgerg-Levy CLT, we have $\left\|\sum_{i=1}^N(\tau_iR_i\varepsilon_i\varepsilon_i'R_i'-E[\tau_iR_i\varepsilon_i\varepsilon_i'R_i'])\right\|/\sqrt{N} = O_p(1)$ so that $\left\|\sum_{i=1}^N \tau_iR_i\varepsilon_i\varepsilon_i'R_i'/N-\Sigma_0\right\| = o_p(N^{-1/4})$. The result then follows from the triangular inequality. □

References

Ai, C., 1997. A semiparametric maximum likelihood estimator. Econometrica 65 (4), 933–963.

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71 (6), 1975–1843.

Altonji, J., Matzkin, R., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73 (4), 1053–1102.

Arellano, M., Carrasco, R., 2003. Binary choice panel data models with predetermined variables. Journal of Econometrics 115, 125–157.

Barlow, R., Barthlomew, D., Bremner, J., Brunk, H., 1972. Statistical Inference under Order Restrictions. John Wiley and Sons.

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. The Annals of Statistics 17 (2), 453–555.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economics Studies XLVII, 225–238.

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., Goldstein, A., 1959. Proximitymaps for convex sets. Proceedings of the American Mathematical Society 10, 448–450.

Gayle, G., Viauroux, C., 2007. Root-n consistent semiparametric estimators of a dynamic panel data sample selection model. Journal of Econometrics 141, 179–212.

Honoré, B., Lewbel, A., 2002. Semiparametric binary choice panel data models without strictly exogeneous regressors. Econometrica 70 (5), 2053–2063.

Horowitz, J., 1992. A smoothed maximum score estimator for the binary response model. Econometrica 60 (3), 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., Linton, O., Nielsen, J., 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. The Annals of Statistics 27 (5), 1443–1490.

Mammen, E., Marron, J., Turlach, B., Wand, M., 2001. a general projection framework for constrained smoothing. Statistical Science 16 (3), 232–248.

Manski, C., 1985. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. Journal of Econometrics 27, 313–33.

Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55 (2), 357–362.

Mood, A., Graybill, F., Boes, D., 1974. Introduction to the Theory of Statistics. McGraw Hill.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46 (1), 69–85.

Newey, W. K., 1994a. Asymptotic variance of semiparametric estimators. Econometrica 62 (6), 1349–1382.

Newey, W. K., 1994b. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10 (2), 233–253.

Newey, W. K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Elsevier Science Publishers.

Newey, W. K., Powell, J. L., 2003. Instrumental variable estimation of nonparametric models. Econometrica 71 (5), 1565–1578.

Opsomer, J., 2000. Asymptotic properties of backfitting estimators. Journal of Multivariate Analysis 73, 166–179.

Robertson, T., Wright, F., Dykstra, R., 1988. Order Restricted Statistical Inference. John Wiley and Sons.

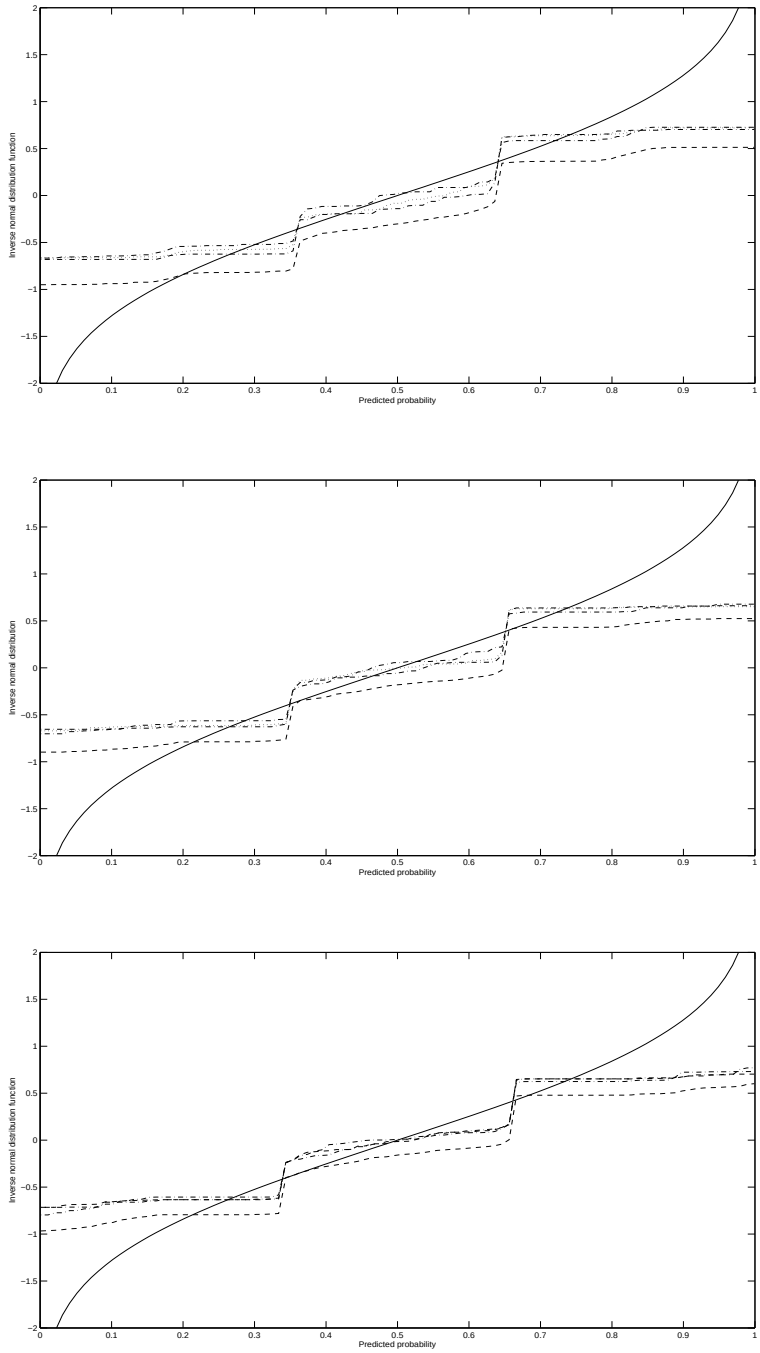
Robinson, P., 1988. Root-n-consistent semiparametric regression. Econometrica 56 (4), 931–954.

Severini, T. A., Wong, W. H., 1992. Profile likelihood and conditionally parametric models. The Annals of Statistics 20 (4), 1768–1802.

Wooldridge, J., 2002. Eonometrics Analysis of Cross Section and Panel Data. MIT Press.

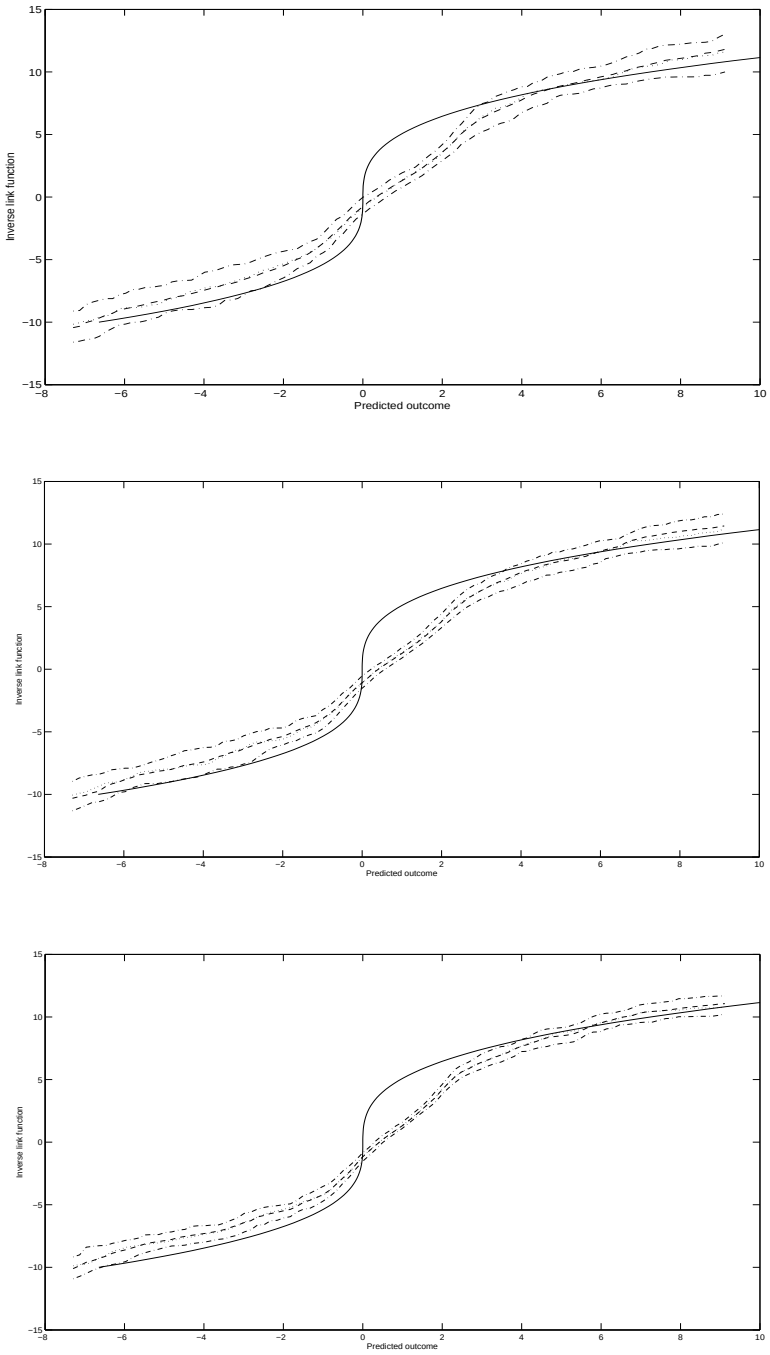
Wooldridge, J., May 2005. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. Review of Economics and Statistics 87 (2), 385–390.

Figure 1: Behavior of the OLKS estimator of the inverse normal distribution function for design 1.[1]



[1]The top to bottom panels present the mean (in a — line), median (in a – – line), 25th and 75th percentiles (in – · – lines), for T=3, and N= 200, 400, and 800, respectively.

Figure 2: Behavior of the OLKS estimator of the inverse link function for design 2.[1]



[1]The top to bottom panels present the mean (in a — line), median (in a – – line), 25th and 75th percentiles (in – · – lines), for T=3, and N= 200, 400, and 800, respectively.