Journal of Economic Theory and Econometrics, Vol. 34, No. 1, Mar. 2023, 1–25

Efficient Estimation of Binary Choice Models with Panel Data[*]

Sungwon Lee[†]

Abstract This paper considers binary choice models with panel data. We extend the correlated random effects binary choice models for panel data in Chamberlain (1980) to semiparametric models in which the conditional expectation projection of the unobserved time-invariant heterogeneity onto the space of functions of time-varying covariates for all time periods is nonparametrically specified. This class of models is tractable for identification and estimation of the model parameters with short panel data. We provide a set of mild conditions under which the parameters are identified. We propose to use the penalized sieve minimum distance (PSMD) estimation and develop the asymptotic theory. The PSMD estimators of finite dimensional parameters are shown to be semiparametrically efficient when the weighting matrix is the optimal one. We also show the bootstrap validity. The Monte Carlo simulation results confirm that the proposed estimator performs well in finite samples.

Keywords Binary choice models, correlated random effects, sieve estimation, semiparametric efficiency, bootstrap

JEL Classification C13, C14, C31

[*] The author is grateful to the editor and two anonymous referees for their helpful comments.

[†] 35 Baekbeom-ro Mapo-gu, Sogang University GN 710, Seoul, Korea 04107; +82-2-705-8510.

Submitted December 17, 2022, Accepted February 16, 2023

1. INTRODUCTION

The use of panel data in empirical economics has become popular. One advantage of using panel data is that we can allow for time-invariant unobserved individual characteristics to be present in a model. Such time-invariant unobserved individual characteristics can be correlated with time-varying regressors in an arbitrary way, and the model becomes a fixed effects model. While the unobserved time-invariant factors may cause omitted variable bias in standard ordinary least squares estimators without controlling them, panel data allow us to circumvent the issue about these omitted variables by using time variation in regressors. It is very attractive to resolve the issue without requiring an excluded variable, but this approach may not work in many nonlinear models.

This paper considers identification and estimation of a class of binary choice models with panel data. Specifically, we focus on the following threshold crossing equation model: for each time period $t = 1,2,...,T$,

$$Y_{it} = \mathbf{1}\left(X_{it}'\theta + c_i \geq U_{it}\right), \tag{1}$$

where i indexes individuals, $\mathbf{1}(\cdot)$ is an indicator function, $X_{it} \in \mathbb{R}^{d_x}$ is a vector of time-varying regressors, c_i is time-invariant unobserved heterogeneity, and U_{it} is a latent error term.

The time-invariant unobserved heterogeneity c_i may be correlated with X_{it}, and one can employ the fixed effects approach, such as within transformation and first-difference, to consistently estimate θ if the model is linear. However, due to the nature of the dependent variable being binary, such approaches are not applicable to model (1). One may wish to control for the time-invariant heterogeneity by including individual-specific dummy variables into the model, but this approach may cause an incidental parameters problem (Neyman and Scott, 1948). For this reason, it is usual to impose a specific distributional assumption on U_{it} to eliminate c_i. When the distribution of U_{it} is the standard logistic distribution, the model becomes a fixed effects logit model (Wooldridge, 2010). While fixed effects logit models do not suffer from the incidental parameters problem, the resulting estimator of θ is not efficient as we cannot use the whole observations in estimation. Another solution to the incidental parameters problem is to use large-T panel data. There are several approaches to dealing with the incidental parameters problem caused by the time-invariant heterogeneity when the number of time periods is large relative to the number of individuals, but they are not applicable to microdata.[1]

[1] Fernández-Val and Weidner (2018) provide an excellent review on large-T panel data mod-

One way to circumvent the incidental parameters problem in nonlinear panel data models is to employ the correlated random effects (CRE) approach, which was pioneered by Mundlak (1978) and Chamberlain (1980), to deal with the time-invariant heterogeneity in (1). The main idea of Mundlak (1978) and Chamberlain (1980) is to specify the conditional expectation projection of c onto X_t's. The CRE approach has been used by many other studies in the literature, including Wooldridge (1995), Wooldridge (2005), Abrevaya and Dahl (2008), Bester and Hansen (2009), and Arellano and Bonhomme (2016), just to name a few. The CRE approach provides flexible and tractable models, while allowing the dependence between the unobserved time-invariant heterogeneity and time-varying regressors. Among these studies, Chamberlain (1980) considers estimation of CRE binary choice models with panel data and a parametric form of the conditional expectation function of c on X_t's. The parametric specification is tractable, and it is easy to estimate the model parameters in practice. However, it is vulnerable to model misspecification, which may result in inconsistency of estimators.

In this paper, we extend the binary choice model with panel data considered by Chamberlain (1980) to a class of semiparametric models. Based on the CRE binary choice model in Chamberlain (1980), we consider a semiparametric specification in which the conditional expectation function of c given X_t's is nonparametrically specified, and the finite dimensional parameter of interest is the coefficient on X_t, θ. In doing so, our semiparametric models may alleviate the issue about model misspecification. On the other hand, we impose a distributional assumption on the latent error term, as in Chamberlain (1980). The distributional assumption on the latent error term increases the possibility of model misspecification, but it provides much more of tractability of the model. In addition, one can consider a wider class of distributions for the latent error term, while allowing for time-invariant unobserved heterogeneity. As a result, the class of models considered in this paper is expected to be useful in empirical analysis.

We then provide a set of conditions under which the model parameters are identified. The key identifying assumption is that X_t should have enough time variation. This assumption is also required for standard linear panel data models with individual fixed effects.

We propose to use a sieve approach to estimating the model parameters (Chen, 2007). The methods of sieve are very flexible and easy to implement in practice. We employ the penalized sieve minimum distance (PSMD) approach developed by Chen and Pouzo (2009). We verify the high-level conditions provided by Chen and Pouzo (2009) and establish the asymptotic theory for the

els, and one can refer to them for a detailed discussion on such models.

estimator. Our focus is on the finite dimensional parameter θ. We develop the asymptotic theory for the PSMD estimator, including consistency, convergence rates, and asymptotic normality for θ. The asymptotic theory relies on large N, and thus, the proposed estimator is applicable to short panel data. The asymptotic variance of the PSMD estimator of θ may be hard to be estimated. For this reason, we show the validity of weighted bootstrap that allows us to avoid estimating the asymptotic variance. The proposed PSMD estimator is shown to be semiparametrically efficient when the weighting matrix is appropriately chosen. The optimal weighting matrix is unknown, but it can be easily estimated by using standard nonparametric approaches. The Monte Carlo simulation study in this paper confirms that our PSMD estimator performs well in finite samples.

While it is common to use maximum likelihood (ML) estimation for binary choice models, we point out that there are several advantages of the PSMD estimators over semiparametric ML estimators, especially sieve ML estimators considered by Chen et al. (2006), Chen (2007), Bierens (2014), or Chen and Liao (2014), in our setting. First, our PSMD procedure does not require to know about the joint distribution of $(Y_1, ..., Y_T)$. Although we impose a distributional assumption on the model, the assumption only restricts the marginal distribution of Y_t for each $t = 1, 2, ..., T$. The PSMD procedure in this paper does not rely on the knowledge on the joint distribution of the dependent variable, whereas the ML estimation requires to specify the joint distribution. Related to this point, the PSMD estimator of θ can achieve the semiparametric efficiency bound (Newey, 1990) by appropriately choosing the weighting matrix, whereas it is required to identify the joint distribution of endogenous variables to obtain efficient estimators of finite dimensional parameters in the sieve ML framework (Chen et al., 2006; Chen, 2007). Lastly, we can use the bootstrap to consistently estimate the asymptotic variance of the PSMD estimator of the finite dimensional parameter, but there is no existing result on the bootstrap validity for sieve ML estimator.

The rest of this paper is organized as follows. Section 2 describes the model and the CRE approach, and considers identification of parameters. Section 3 explains the PSMD estimation, and Section 4 provides the asymptotic theory. Section 5 reports the Monte Carlo simulation results, and Section 6 concludes.

Notation Before proceeding, we introduce some notation. For a generic random variable A, the support of A is denoted by $Supp(A)$. $\mathbb{E}[\cdot]$ is the expectation operator. For a generic (random) vector x, $||x||_E$ denotes the Euclidean norm of x. For a set of d_x-dimensional random vectors $X_1, X_2, ..., X_T$, $\mathbf{X} \equiv (X_1, X_2, ..., X_T)'$ denotes a $T \times d_x$ random matrix (i.e., the vector that collects all X_t's).

2. MODEL AND IDENTIFICATION

We consider the following threshold crossing equation model: for each $i \in \{1,2,...,n\}$ and $t \in \{1,2,...,T\}$,

$$Y_{it} = \mathbf{1}\left(X_{it}^{'}\theta + c_i \geq U_{it}\right), \tag{2}$$

where $\mathbf{1}(\cdot)$ is an indicator function, $X_{it} \in \mathbb{R}^{d_x}$ is a vector of time-varying reressors, $c_i \in \mathbb{R}$ is unobserved individual heterogeneity, and U_{it} is a time-varying latent error term. We can only observe $\left(Y_{it}, X_{it}^{'}\right)^{'}$ from the data for all $i \in \{1,2,...,n\}$ and $t \in \{1,2,...,T\}$.

Unlike the linear panel data models, it is challenging to deal with the unobserved individual heterogeneity c_i in both identification and estimation when it is correlated with some time-varying regressor. In addition, when we include individual dummy variables to incorporate c_i, it is well known that the estimators suffer from an incidental parameter problem (Neyman and Scott, 1948). One can use a logistic specification for U_{it} to eliminate c_i, but this approach requires that we use not all observations in the data, which may lead to a significant loss of efficiency of the estimator. More importantly, the logistic assumption is vulnerable to model misspecification.

In this paper, we adopt a CRE approach that was pioneered by Mundlak (1978) and Chamberlain (1980). Specifically, we consider the following specification for c that allows arbitrary correlations between c and X_t's:

$$\begin{aligned} c_i &= \mathbb{E}[c_i|\mathbf{X}_i] + V_i \\ &\equiv h(\mathbf{X}_i) + V_i, \end{aligned}$$

where $\mathbf{X}_i$ is the vector that collects X_{it}'s across all time periods, and V_i is the conditional expectation error. Then, model (2) can be written as

$$Y_{it} = \mathbf{1}\left(X_{it}^{'}\theta + h(\mathbf{X}_i) \geq U_{it} - V_i\right).$$

The model is a semiparametric extension of the model in Chamberlain (1980), who considers the case where $h(\mathbf{X}_i) = \sum_{t=1}^{T} X_{it}^{'}\gamma_t$.

We provide a set of conditions under which the model parameters are identified. Let $\varepsilon_{it} \equiv U_{it} - V_i$. Before proceeding, we introduce some notation. For a generic random variable X_{it}, $\Delta X_t \equiv X_{it} - X_{it-1}$ denotes the first-difference, and $\perp$ indicates statistical independence. Since we consider panel data, it is implicitly assumed that $T \geq 2$.

Assumption 1. *The following conditions hold:*

(i) $\mathbb{E}\left[\Delta X_{it}\Delta X_{it}^{'}\right]$ *is of full rank for all* $t = 2,3,...,T$*;*

(ii) For all $i \in \{1,2,...,n\}$ *and* $t = 1,2,...,T$, $\varepsilon_{it}|\mathbf{X}_i \overset{d}{=} \varepsilon_{it} \sim F_\varepsilon$*, where* F_ε *is a known strictly increasing function over* $\mathbb{R}$*.*

Condition (i) of Assumption 1 requires that X_{it} be time-varying. As a result, one cannot include a constant term or time-invariant regressors to X_{it}, which is the same for the standard linear panel data models. We note that time-invariant regressors, including a constant term, can be included in $h(\cdot)$. We also point out that one can use a different rank condition than Assumption 1(i). For example, it is possible to consider a condition that $\mathbb{E}\left[\left(X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}\right)\left(X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}\right)^{'}\right]$ is of full rank for all t. The role of Assumption 1(i) is to guarantee that X_{it} has sufficient time variation. Condition (ii) of Assumption 1 specifies the distribution of ε_{it} and assumes that $\varepsilon_{it} \perp \mathbf{X}_i$. This condition implicitly imposes a strict exogeneity of $\mathbf{X}_i$.

The following theorem shows that the model parameters are identified under Assumption 1:

Theorem 1. *Suppose that Assumption 1 hold. Then,* θ *is identified. Moreover,* $h(\cdot)$ *is identified over the support of* $\mathbf{X}$.

3. ESTIMATION

The model we consider in this paper contains both finite and infinite dimensional objects, and thus it is a semiparametric model. To estimate the model parameters, we adopt the penalized sieve minimum distance (PSMD) estimation approach proposed by Chen and Pouzo (2009). The methods of sieves provide a flexible and tractable way to estimate semiparametric and nonparametric models and are widely used in the literature (e.g., Song, 2015; Lee, 2022).

Let θ_0 and h_0 denote the true parameter value for θ and h, respectively. We start with observing that the identification result in Theorem 1 implies that for each i and t,

$$\mathbb{E}\left[Y_{it} - F_\varepsilon\left(X_{it}^{'}\theta + h(\mathbf{X}_i)\right)|\mathbf{X}_i\right] = 0 \text{ almost surely} \tag{3}$$

if and only if $\theta = \theta_0$ and $h = h_0$. This leads us to considering a PSMD approach to estimating θ and h, based on the conditional moment restriction in (3). Let Θ and $\mathscr{H}$ be the parameter spaces for θ and h, respectively. We denote the parameter as $\alpha \equiv \left(\theta^{'}, h\right)^{'}$ and let $\mathscr{A}$ be the Cartesian product of Θ and $\mathscr{H}$.

Let $W_{it} \equiv \left(Y_{it}, X_{it}^{'}\right)^{'}$ and $\rho_t(\mathbf{W}_i;\alpha) \equiv Y_{it} - F_{\varepsilon}\left(X_{it}^{'}\theta + h(\mathbf{X}_i)\right)$ for each $t = 1,2,...,T$. Then, we have T conditional moment restrictions and denote $\rho(\mathbf{W}_i) \equiv [\rho_1(\mathbf{W}_i),...,\rho_T(\mathbf{W}_i)]^{'}$. We also define $m(\mathbf{X}_i;\alpha) \equiv \mathbb{E}[\rho(\mathbf{W}_i;\alpha)|\mathbf{X}_i]$. Let $\mathscr{A}_n$ be a sieve space for the parameter space $\mathscr{A}$. The PSMD estimator of α_0, $\hat{\alpha}_n$, is defined as

$$\hat{\alpha}_n \equiv \arg\inf_{\alpha\in\mathscr{A}_n}\left\{\frac{1}{n}\sum_{i}^{n}\hat{m}_n(\mathbf{X}_i;\alpha)^{'}\left[\hat{\Sigma}_n(\mathbf{X}_i)\right]^{-1}\hat{m}_n(\mathbf{X}_i;\alpha)+\lambda_n\hat{P}_n(h)\right\}, \qquad (4)$$

where $\hat{m}_n(\mathbf{x};\alpha)$ is a consistent estimator of $m(\mathbf{x};\alpha)$, $\hat{\Sigma}_n(\mathbf{x})$ is a consistent estimator of positive definite matrix $\Sigma(\mathbf{x})$, $\hat{P}_n(h) \geq 0$ is a possibly random penalty function, and λ_n is a positive real sequence such that $\lambda_n \downarrow 0$.

To compute the PSMD estimator $\hat{\alpha}_n$, it is required to obtain a consistent estimator of $m(\mathbf{x})$. In this paper, we use a series estimator of $m(\mathbf{X};\alpha)$, $\hat{m}_n(\mathbf{X};\alpha)$.[2] Specifically, for each $t = 1,2,...,T$, define

$$\hat{m}_{t,n}(\mathbf{X};\alpha) \equiv p^{J_n}(\mathbf{X})^{'}(P^{'}P)^{-}\sum_{i=1}^{n}p^{J_n}(\mathbf{X}_i)\rho_t(\mathbf{W}_i;\alpha), \qquad (5)$$

where $\{p_j(\cdot)\}_{j=1}^{\infty}$ is a sequence of some basis functions,

$$p^{J_n}(\mathbf{x}) \equiv (p_1(\mathbf{x}), p_2(\mathbf{x}),...,p_{J_n}(\mathbf{x}))^{'},$$

$$P \equiv \left[p^{J_n}(\mathbf{X}), p^{J_n}(\mathbf{X}),...,p^{J_n}(\mathbf{X})\right]^{'},$$

and $\left(P^{'}P\right)^{-}$ is the generalized inverse matrix of $P^{'}P$. Then,

$$\hat{m}_n(\mathbf{X};\alpha) \equiv [\hat{m}_{1,n}(\mathbf{X};\alpha),...,\hat{m}_{T,n}(\mathbf{X};\alpha)]^{'}.$$

To define the parameter space for h, we introduce a class of functions. Let $f : \mathbb{D} \to \mathbb{R}$ where $\mathbb{D} \subseteq \mathbb{R}^{d_x}$ for some integer $d_x \geq 1$. Let $\omega = (\omega_1,...,\omega_{d_x})$ be a d_x-tuple of nonnegative integers, and define the differential operator as $\nabla^{\omega} f \equiv \frac{\partial^{|\omega|}}{\partial x_1^{\omega_1}\partial x_2^{\omega_2}\cdots\partial x_{d_x}^{\omega_{d_x}}}f(x)$, where $x = (x_1, x_2,...,x_{d_x}) \in \mathbb{D}$ and $|\omega| \equiv \sum_{i=1}^{d_x}\omega_i$. Let $[p]$ be the integer part of $p \in \mathbb{R}_+$, then a function $f : \mathscr{X} \to \mathbb{R}$ is called p-smooth if it is $[p]$ times continuously differentiable on $\mathscr{X}$ and for all ω such that $|\omega| = [p]$ and for some $\nu \in (0,1]$ and constant $c > 0$, $|\nabla^{\omega} f(x) - \nabla^{\omega} f(y)| \leq c \cdot ||x-y||_E^{\nu}$ for all $x, y \in \mathscr{X}$, where $||\cdot||_E$ is the Euclidean norm. Let $\mathscr{C}^{[p]}(\mathscr{X})$ denote the space of

[2] One can refer to, for example, Li and Racine (2007) for details on series estimation.

all $[p]$ times continuously differentiable real-valued functions on $\mathscr{X}$. A Hölder ball with smoothness p is defined as follows:

$$\Lambda_C^p(\mathscr{X}) \equiv \{f \in \mathscr{C}^{[p]}(\mathscr{X}) : \sup_{|\omega|\leq[p]} \sup_{x\in\mathscr{X}} |\nabla^{\omega} f(x)| \leq C,$$
$$\sup_{|\omega|=[p]} \sup_{x,y\in\mathscr{X},x\neq y} \frac{|\nabla^{\omega} f(x) - \nabla^{\omega} f(y)|}{||x-y||_E^{\nu}} \leq C\},$$

where C is a positive finite constant.

For a random variable X, let x_{min} and x_{max} denote the minimum and maximum values of X, respectively. For a given positive integer l, let $t_0, t_1, ..., t_l$ be real numbers such that $x_{min} = t_0 < t_1 < \cdots < t_{l+1} = x_{max}$. Let $\text{Spl}(r,l)$ denote the space of polynomial splines with order r and l interior knots:

$$\text{Spl}(r,l) \equiv \left\{ \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{l} b_j \left[\max\left\{x - t_j, 0\right\}\right]^{r-1}, x \in [x_{min}, x_{max}] : a_k, b_j \in \mathbb{R} \right\}.$$

Then, the complexity of $\text{Spl}(r,l)$ is determined by $k_n \equiv r + l$.

4. ASYMPTOTIC THEORY

We develop the asymptotic theory for the PSMD estimator $\hat{\alpha}_n$. The asymptotic theory presented in this paper relies on $n \to \infty$ with $T < \infty$. Let $||\cdot||_\infty$ be the supremum norm on $\mathscr{A}$ that is defined as $||\alpha||_\infty \equiv ||\theta||_E + ||h||_\infty$, where $||h||_\infty \equiv \sup_{\mathbf{x}\in Supp(\mathbf{X})} |h(\mathbf{x})|$. We first show that the PSMD estimator is consistent for α_0 with respect to norm $||\cdot||_\infty$. The L_2 norm on $\mathscr{A}$ is defined as $||\alpha||_2 \equiv ||\theta||_E + ||h||_2$, where $||h||_2^2 \equiv \int h(\mathbf{x})^2 dF_{\mathbf{X}}(\mathbf{x})$. For a set A, let $int(A)$ denote the interior of A.

4.1. CONSISTENCY

To establish the consistency of the PSMD estimator $\hat{\alpha}_n$, we impose the following conditions.

Assumption 2. *(i) The data* $\{\mathbf{W}_i : i = 1, 2, ...n\}$ *are i.i.d; (ii) The conditional distribution of* $\mathbf{Y}$ *on* $\mathbf{X}$ *admits its conditional density function* $f_{\mathbf{Y}|\mathbf{X}}(y|\mathbf{x})$ *that is continuous in* $(\mathbf{y}, \mathbf{x})$ *and* $\sup_{\mathbf{y}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) < \infty$ *for all* $\mathbf{x} \in Supp(\mathbf{X})$*; (iii) Supp*$(\mathbf{X})$ *is a compact subset of* $\mathbb{R}^{T\cdot d_x}$ *with Lipschitz continuous boundary; (iv) the density* *function of* $\mathbf{X}$*,* $f_{\mathbf{X}}(\cdot)$*, is bounded and bounded away from zero over* $Supp(\mathbf{X})$*; (v) there exists* $\varepsilon_0 > 0$ *such that for all* $t = 1, 2, ..., T$ *and for all* $\alpha \in \mathscr{A}$*,*

$$F_\varepsilon\left(X_t^{'}\theta + h(\mathbf{X})\right) \in [\varepsilon_0, 1-\varepsilon_0]$$

almost surely; (vi) F_ε *is absolutely continuous with respect to the Lebesgue measure; (vii) The density function* f_ε *is continuous and uniformly bounded over* $\mathbb{R}$*; (viii)* $\mathbb{E}\left[\max_{t\leq T}||X_t||_E^2\right] < \infty$.

Assumption 3. *(i)* $\theta_0 \in int(\Theta)$*, where* Θ *is a compact subset of* $\mathbb{R}^{d_x}$*; (ii)* $h_0 \in \mathscr{H} \equiv \Lambda_{c_h}^{p_h}(Supp(\mathbf{X}))$ *with* $p_h > \frac{Td_x}{2}$*; (iii) all first-order partial derivatives of* h_0 *are uniformly bounded.*

Assumption 4. *(i)* $(p_j(\cdot))_{j=1}^{\infty}$ *is a sequence of polynomial spline (P-spline) functions; (ii) the sieve space for* $\mathscr{H}$ *is*

$$\mathscr{H}_n \equiv \left\{h_n(\mathbf{x}) = p^{k_n}(\mathbf{x})^{'}\beta_n : ||h_n||_\infty \leq c_h\right\},$$

where k_n *is a positive non-decreasing integer sequence such that* $k_n \to \infty$ *and* $k_n = o(n)$*; (iii) The eigenvalues of* $\mathbb{Q}_n \equiv \mathbb{E}\left[p^{k_n}(\mathbf{X})p^{k_n}(\mathbf{X})^{'}\right]$ *are bounded above and away from zero uniformly over all n.*

Assumption 5. *For each* $\alpha \in \mathscr{A}$*,* $m(\cdot;\alpha) \in \Lambda_{c_m}^{p_m}(Supp(\mathbf{X}))$ *with* $p_m > \frac{Td_x}{2}$.

Assumption 6. *(i)* $J_n \geq k_n + d_\theta$ *and* $J_n \log(J_n) = o(n)$*;*
(ii) $\max_{j\leq J_n}\mathbb{E}\left[||p_j(\mathbf{X})||_E^2\right] < C < \infty$ *for some constant C.*

Assumption 7. $\hat{P}_n(h) = 0$ *for all n and* $h \in \mathscr{H}$*.*

Assumption 8. $\Pr\left(\Sigma(\mathbf{X}) = \hat{\Sigma}_n(\mathbf{X}) = I_T\right) = 1$ *for all n.*

Assumption 2 is standard. It is worth mentioning that condition (i) of Assumption 2 allows serial correlation of variables within individuals (i.e., it means that $Cov(W_{it}, W_{is}) \neq 0$ for $t, s \in \{1, 2, ..., T\}$). Condition (v) of Assumption 2 implies that for each time period $t = 1, 2, ..., T$, there are a group of individuals with $Y_t = 1$ and a group of individuals with $Y_t = 0$.

Assumption 3 defines the parameter spaces for θ and h_0. It also imposes some smoothness of h_0. Assumption 4 defines the sieve space for the parameter space for h_0, $\mathscr{H}$. Since $\mathscr{H}$ is a Hölder ball and the support of $\mathbf{X}$ is compact, one can approximate an element of $\mathscr{H}$ by using polynomial, trigonometric, or spline sieve spaces.[3] Since $\mathbf{X} \in \mathbb{R}^{Td_x}$, the P-spline sieve space can be constructed by a

[3] A detailed discussion on the choice of sieve spaces can be found in Chen (2007).

tensor product of univariate P-spline sieve spaces. Condition (iii) of Assumption 4 is standard in the literature on series or sieve estimation.

Assumptions 5 and 6 define the parameter space for m and impose some restriction on the sieve space that is used for approximating $m(\mathbf{X};\alpha)$. Condition (i) of Assumption 6 can be interpreted as an order condition. Condition (ii) of Assumption 6 is standard.

Assumption 7 implies that we do not use penalization. Under this assumption, the PSMD estimator in (4) becomes a SMD estimator considered by Ai and Chen (2003). Although we do not consider penalization, there are some popular non-trivial penalty functions, such as $||\nabla h||_2^2$ or $||\nabla^2 h||_2^2$. One can refer to, for example, Chen and Pouzo (2009) and Chen and Pouzo (2012) for details on non-trivial penalty functions for the PSMD estimation.

Assumption 8 specifies the weighting matrix. Using the identity matrix as a weighting matrix, the PSMD estimator is not optimally weighted. We discuss how to obtain (semiparametrically) efficient estimators of θ_0 at the end of this section.

The next theorem demonstrates that the PSMD estimator $\hat{\alpha}_n$ is consistent with respect to the supremum norm:

Theorem 2. *Suppose that Assumption 1 holds. If Assumptions 2–8 are satisfied, then*

$$||\hat{\alpha}_n - \alpha_0||_\infty = o_p(1).$$

4.2. CONVERGENCE RATES

Given that the sieve estimator $\hat{\alpha}_n$ is consistent for α_0 with respect to $||\cdot||_{\mathscr{A},\infty}$, we consider a shrinking $||\cdot||_{\mathscr{A},\infty}$ neighborhood around α_0. For given small $\varepsilon > 0$ and large $M > 0$, we define

$$\mathscr{A}_{os} \equiv \left\{\alpha \in \mathscr{A} : ||\alpha - \alpha_0||_{\mathscr{A},\infty} \leq \varepsilon, ||\alpha||_{\mathscr{A},\infty} \leq M\right\},$$
$$\mathscr{A}_{osn} \equiv \mathscr{A}_{os} \cap \mathscr{A}_n.$$

Define

$$\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \frac{d\mathbb{E}\left[\rho(\mathbf{W};(1-t)\alpha_0 + t\alpha|\mathbf{X}\right]}{dt}\bigg|_{t=0}$$

as the pathwise derivative of m in the direction $[\alpha - \alpha_0]$ evaluated at α_0. Let $||\cdot||$ denote a pseudo metric on $\mathscr{A}_{os}$, where for any $\alpha_1, \alpha_2 \in \mathscr{A}_{os}$,

$$||\alpha_1 - \alpha_2|| \equiv \sqrt{\mathbb{E}\left[\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)'(\Sigma(\mathbf{X}))^{-1}\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)\right]}.$$

For any positive real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ means that there exist a finite constant $C > 0$ and $N \in \mathbb{N}$ such that $a_n \leq Cb_n$ for all $n \geq N$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$, it is denoted by $a_n \asymp b_n$.

Assumption 9. *(i) $\mathscr{A}_{os}$ and $\mathscr{A}_{osn}$ are convex; (ii) $\mathbb{E}\left[||m(\mathbf{X};\alpha)||_E^2\right] \asymp ||\alpha - \alpha_0||^2$ for all $\alpha \in \mathscr{A}_{osn}$; (iii) there exists $\delta_0 > 0$ such that for all $t = 1,2,...,T$ and for any $\alpha \in \mathscr{A}_{os}$, $f_\varepsilon(X_t^{'}\theta + h(\mathbf{X})) \geq \delta_0$ almost surely.*

Assumption 9 is mild, in particular when we focus on a shrinking neighborhood of α_0.

Let $\overline{\mathbb{V}}$ be the closure of the linear span of $\mathscr{A}_{os} - \{\alpha_0\}$ under $||\cdot||$. For any $v_1, v_2 \in \overline{\mathbb{V}}$, define an inner product as

$$< v_1, v_2 >\equiv \mathbb{E}\left[\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v_1]\right)^{'}(\Sigma(\mathbf{X}))^{-1}\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v_2]\right)\right].$$

Then, $\left(\overline{\mathbb{V}}, <\cdot,\cdot>\right)$ is a Hilbert space. Note that

$$\overline{\mathbb{V}} = \mathbb{R}^{d_\beta} \times \overline{\mathscr{W}},$$

where $\overline{\mathscr{W}} \equiv \left\{ w : \mathbb{E}\left[\left|\left|\Sigma(\mathbf{X})^{-\frac{1}{2}}\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[w]\right|\right|_E^2\right] < \infty \right\}$. For a given component θ_j of θ, $j = 1,2,...,d_x$, let

$$D_{w_j}(\mathbf{X}) \equiv \frac{dm(\mathbf{X};\alpha_0)}{d\theta_j} - \frac{dm(\mathbf{X};\alpha_0)}{dh}[w_j]$$

and $w_j^* \in \overline{\mathscr{W}}$ denote the solution to

$$\inf_{w_j \in \overline{\mathscr{W}}} \mathbb{E}\left[D_{w_j}(\mathbf{X})^{'}\Sigma(\mathbf{X})^{-1}D_{w_j}(\mathbf{X})\right].$$

Let $\mathbf{w}^* \equiv \left(w_1^*, w_2^*, ..., w_{d_x}^*\right)$ and $D_{\mathbf{w}^*}(\mathbf{X}) = \frac{dm(\mathbf{X};\alpha_0)}{d\theta} - \frac{dm(\mathbf{X};\alpha_0)}{dh}[\mathbf{w}^*]$. Then, $D_{\mathbf{w}^*}(\mathbf{X})$ is the vector of the efficient score functions (Bickel et al., 1993; Ai and Chen, 2003; Chen and Pouzo, 2009).

Assumption 10. $\mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})^{'}\Sigma(\mathbf{X})^{-1}D_{\mathbf{w}^*}(\mathbf{X})\right]$ *is finite and positive definite.*

Theorem 3. *Suppose that Assumptions 1, and 2–10 hold. Then,*

$$||\hat{h}_n - h_0||_2 = O_p\left(\max\left\{\delta_{m,n}, k_n^{-\frac{p_h}{Td_x}}\right\}\right),$$

where $\delta_{m,n}^2 = \max\left\{\frac{J_n}{n}, J_n^{-\frac{2p_m}{Td_x}}\right\}$.

The L_2-convergence rate of $\hat{h}_n$ in Theorem 3 is a standard nonparametric convergence rate. Since the infinite dimensional parameter h_0 does not contain any endogenous regressors (i.e., $\varepsilon_t \perp \mathbf{X}$ for all $t = 1, 2, ..., T$), the PSMD estimator $\hat{\alpha}_n$ does not suffer from an ill-posed inverse problem (Carrasco *et al.*, 2007; Horowitz, 2014).

4.3. ASYMPTOTIC NORMALITY FOR THE FINITE DIMENSIONAL PARAMETER

We establish the asymptotic normality of $\hat{\theta}_n$. The convergence rate result in Theorem 3 allows us to focus on shrinking neighborhoods of α_0. Let

$$\mathscr{N}_0 \equiv \{\alpha \in \mathscr{A}_{os} : ||\alpha - \alpha_0||_2 \leq \delta_{2,n}, ||\alpha - \alpha_0||_\infty \leq M\}$$

for some $M > 0$ and $\mathscr{N}_n \equiv \mathscr{N}_0 \cap \mathscr{A}_n$.

Let $\lambda \in \mathbb{R}^{d_x} - \{0\}$ and define $f(\alpha_0) = \lambda^{'}\theta_0$. Under Assumption 10, there exists $v^* \in \overline{\mathbb{V}}$ such that $\lambda^{'}(\hat{\theta}_n - \theta_0) =< v^*, \hat{\alpha}_n - \alpha_0 >$ by the Riesz representation theorem. Furthermore, one can show that $v^* \equiv (v^*_\theta, v^*_h)$, where $v^*_\theta = \mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})^{'}\Sigma(\mathbf{X})^{-1}D_{\mathbf{w}^*}(\mathbf{X})\right]$ and $v^*_h = -\mathbf{w}^* \times v^*_\theta$.

For random variables X, Y, and Z, let $Corr(X,Y|Z)$ denote the conditional correlation coefficient between X and Y given Z.

Assumption 11. *For $t, s \in \{1, 2, ..., T\}$ such that $t \neq s$, $|Corr(Y_t, Y_s|\mathbf{X})| \neq 1$ almost surely.*

Assumption 12. *f_ε is continuously differentiable and its derivative is uniformly bounded.*

Assumption 13. *(i) v^*_n is the projection of v^* onto $\mathscr{A}_n - \{\alpha_0\}$ under $||\cdot||$ and satisfies $||v^*_n - v^*|| = o\left(n^{-1/4}\right)$; (ii) $\delta_{2,n} = o\left(n^{-1/4}\right)$.*

Assumption 14. *$\frac{dm(\cdot;\alpha_0)}{d\alpha}[v^*] \in \Lambda^{p_d}_{c_d}(Supp(\mathbf{X}))$ with $J_n^{-\frac{2p_d}{Td_x}} = o\left(n^{-1/2}\right)$.*

Assumption 11, together with condition (iv) of Assumption 2, implies that $Var\left(\rho(\mathbf{W};\alpha_0)|\mathbf{X}\right)$ is positive definite almost surely.

Assumption 12 strengthens the smoothness of F_ε in the sense that F_ε is twice continuously differentiable.

Assumption 13 restricts the rates of k_n and J_n. Condition (i) of Assumption 13 is required to eliminate the approximation error of the Riesz representer v^*. It is usually satisfied if v^* belongs to a class of smooth functions (e.g., Hölder spaces, Sobolev space) and the sieve space for $\mathscr{A}$, $\mathscr{A}_n$, does well approximate an element of the class of functions. Condition (ii) of Assumption 13 is satisfied with a proper choice on k_n and J_n under Assumptions 3 and 4.

Assumption 14 requires that the pathwise derivative of $m(\cdot;\alpha)$ with respect to α evaluated at α_0 in the direction v^* be in a class of smooth functions. The latter condition is required to eliminate the approximation error of $\frac{dm(\cdot;\alpha_0)}{d\alpha}[v^*]$ when using a series estimator of $\frac{dm(\cdot;\alpha_0)}{d\alpha}[v^*]$.

We now provide the asymptotic normality of the PSMD estimator of θ_0:

Theorem 4. *Suppose that Assumptions 1, and 2–14 hold. Then,*

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}N(0,V),$$

where

$$V\equiv\left(\mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})'D_{\mathbf{w}^*}(\mathbf{X})\right]^{-1}\mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})'\Sigma_0(\mathbf{X})D_{\mathbf{w}^*}(\mathbf{X})\right]\mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})'D_{\mathbf{w}^*}(\mathbf{X})\right]^{-1}\right).$$

4.4. WEIGHTED BOOTSTRAP

The asymptotic variance provided in Theorem 4 can be consistently estimated in a similar way to Ai and Chen (2003). However, it may be costly to estimate the asymptotic variance, in particular when the dimension of θ is large. We propose a weighted bootstrap procedure that can consistently estimate the asymptotic variance V.

Assumption 15. *Let $\{B_i\}_{i=1}^n$ be an i.i.d. sample of positive random variable B such that $E[B]=1$ and $Var(B)=b_0<\infty$ that is independent of $(\mathbf{W}_i)_{i=1}^n$.*

One can use a multinomial random variable or exponential random variable to draw bootstrap weights. Define

$$\hat{m}_{n,B}(\mathbf{X};\alpha)\equiv p^{J_n}(\mathbf{X})'\left(P'P\right)^{-}\sum_i^n p^{J_n}(X_i)\rho(\mathbf{W}_i;\alpha)\cdot B_i$$

and

$$\hat{\alpha}_n^*\equiv\arg\inf_{\alpha\in\mathscr{A}_n}\left\{\frac{1}{n}\sum_i^n\hat{m}_{n,B}(\mathbf{X}_i;\alpha)'\left[\hat{\Sigma}_n(\mathbf{X}_i)\right]^{-1}\hat{m}_{n,B}(\mathbf{X}_i;\alpha)+\lambda_n\hat{P}_n(h)\right\}.$$

The following theorem establishes the validity of the weighted bootstrap procedure:

Theorem 5. *Suppose that Assumptions 1, and 2–15 hold. Then, conditional on the data* $\{\mathbf{W}_i : i = 1,2,...,n\}$, $\sqrt{\frac{n}{b_0}}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ *has the same limiting distribution to that of* $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$.

4.5. SEMIPARAMETRIC EFFICIENCY

We now consider the semiparametric efficiency of the PSMD estimator of θ_0. Let $\Sigma_0(\mathbf{X}) \equiv Var(\rho(\mathbf{W};\alpha_0)|\mathbf{X})$ and $\hat{\Sigma}_{0,n}(\mathbf{X})$ be a consistent estimator of $\Sigma_0(\mathbf{X})$. Let $q_0(\mathbf{y},\mathbf{x},\alpha_0)$ be the true joint density function of $(\mathbf{Y},\mathbf{X})$ and consider $p(\mathbf{y},\mathbf{x},\theta,\xi) \equiv q_0(\mathbf{y},\mathbf{x},\theta,h_0+\xi(h-h_0))$ for a fixed $h \in \mathscr{H}$ and some small $\xi > 0$ such that $h_0+\xi(h-h_0) \in \mathscr{H}$. We impose the following conditions:

Assumption 16. *(i)* $\Pr\left(\hat{\Sigma}_n(\mathbf{X}) = \hat{\Sigma}_{0,n}(\mathbf{X})\right) = 1$ *for all n; (ii)* $\sup_{\mathbf{x}\in Supp(\mathbf{X})}\left|\hat{\Sigma}_{0,n}(\mathbf{x}) - \Sigma_0(\mathbf{x})\right| = O_p\left(\delta_{\Sigma,n}\right)$ *with* $\delta_{\Sigma,n} = o\left(n^{-1/4}\right)$.

Assumption 17. *For any* $h \in \mathscr{H}$, $p(\mathbf{y},\mathbf{x},\theta,\xi)$ *is smooth in the sense of Newey (1990).*

Assumption 16 requires that there exist a consistent estimator of the optimal weighting matrix Σ_0. Recall that the diagonal elements of $\Sigma_0(\mathbf{X})$ are the conditional variances of $\rho_t(\mathbf{W};\alpha_0)$'s given $\mathbf{X}$. The off-diagonal elements of $\Sigma_0(\mathbf{X})$ are the conditional covariances between $\rho_t(\mathbf{W};\alpha_0)$ and $\rho_s(\mathbf{W};\alpha_0)$, with $t \neq s$, given $\mathbf{X}$. Recall that

$$Var\left(\rho_t(\mathbf{W};\alpha_0)|\mathbf{X}\right) = F_\varepsilon\left(X_t^{'}\theta_0 + h_0(\mathbf{X})\right)\cdot\left(1 - F_\varepsilon\left(X_t^{'}\theta_0 + h_0(\mathbf{X})\right)\right) \tag{6}$$

and

$$\begin{aligned} Cov\left(\rho_t(\mathbf{W};\alpha_0),\rho_s(\mathbf{W};\alpha_0)|\mathbf{X}\right) =& \mathbb{E}\left[Y_tY_S|\mathbf{X}\right] \\ & - F_\varepsilon\left(X_t^{'}\theta_0 + h_0(\mathbf{X})\right)F_\varepsilon\left(X_s^{'}\theta_0 + h_0(\mathbf{X})\right) \end{aligned} \tag{7}$$

for $t \neq s$. One can easily obtain a consistent estimator of $Var\left(\rho_t(\mathbf{W};\alpha_0)|\mathbf{X}\right)$ by replacing the unknown parameters in (6) with their consistent PSMD estimators. To consistently estimator $Cov\left(\rho_t(\mathbf{W};\alpha_0),\rho_s(\mathbf{W};\alpha_0)|\mathbf{X}\right)$, it is needed to consistently estimate $\mathbb{E}\left[Y_tY_S|\mathbf{X}\right]$. There are many nonparametric approaches to estimating the conditional expectation function, including kernel regression and series estimation. Condition (ii) of Assumption 16 can hold with those nonparametric estimation approaches with an appropriate choice of tuning parameters.

Assumption 17 is a sufficient condition for the PSMD estimator of θ_0 to achieve the semiparametric efficiency bound.

Theorem 6. *Suppose that Assumptions 1, and 2–7, 9–14, and 16 hold. Then,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, V_0),$$

where $V_0 \equiv \mathbb{E}\left[D_{\mathbf{w}^*}(\mathbf{X})' \Sigma_0(\mathbf{X})^{-1} D_{\mathbf{w}^*}(\mathbf{X})\right]^{-1}$.

If Assumption 17 additionally holds, then the PSMD estimator of θ_0 *is semiparametrically efficient.*

5. MONTE CARLO SIMULATION

We conduct a small Monte Carlo simulation study to investigate the finite sample performance of the PSMD estimator $\hat{\alpha}_n$. Let $\Phi(\cdot)$ denote the standard normal distribution function and consider the following data generating process (DGP) with two time periods $(T = 2)$:

$$Y_t = \mathbf{1}(X_t\theta_0 + c \geq U_t),$$

for each $t = 1,2$, where $\theta_0 = 1$, $c = \Phi\left(\frac{\sum_{t=1}^{T} X_t}{T}\right) + V$ with $V \sim N(0, 0.2)$, $U_t \sim N(0, 0.8)$, $Cov(U_t, U_s) = 0.1$ for all $t,s \in \{1,2,...,T\}$ such that $t \neq s$. We generate $(X_1^*, X_2^*)'$ from $BVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}\right)$, and they are independent of U_1, U_2, and V. Then, we construct $X_1 \equiv 2\Phi(X_1^*) - 1$ and $X_2 \equiv 2\Phi(X_2^*) - 1$ so that $X_1$and X_2 have a bounded support. As a result, the model can be rewritten as follows: for each $t = 1,2$,

$$Y_t = \mathbf{1}\left(X_t\theta_0 + \Phi\left(\frac{X_1 + X_2}{2}\right) \geq \varepsilon_t\right),$$

where $\varepsilon_t | \mathbf{X} \sim N(0,1)$. Note that $h_0(\mathbf{x}) = \Phi\left(\frac{X_1+X_2}{2}\right)$. The sample size is set to be 500, and all Monte Carlo simulation results are obtained from 500 iterations.

We focus on the performance of $\hat{\theta}_n$, the PSMD estimator of θ_0. As the performance measure, we consider the bias, standard deviation (S.D), and the root mean squared error (RMSE). The sieve space for $\mathscr{H}$ is P-spline sieve spaces. We consider several values for the order (r) and number of interior knots (l).[4]

Table 1 presents the Monte Carlo simulation results. We find that the PSMD estimator of θ_0 performs well as the bias is negligible and the magnitude of the standard deviation is reasonable. The performance of the PSMD estimator is not sensitive to the choice of the order or P-spline functions and the number of interior knots.

[4] While it is challenging to choose r and l in a data-dependent way, one can use some information criteria (e.g., AIC, BIC) to choose r and/or l, as in Chen and Liao (2014).

Table 1: Monte Carlo Simulation for θ_0 ($n = 500$)

(r,l)	Bias	S.D	RMSE
$(3,3)$	0.0533	0.1343	0.1445
$(3,4)$	0.0408	0.1397	0.1456
$(4,3)$	0.0459	0.1337	0.1413
$(4,4)$	0.0465	0.1288	0.1370

Note: The number of simulations is set to be 500.

6. CONCLUSION

In this paper, we consider binary choice models with panel data. The model is a semiparametric extension of the model of Chamberlain (1980), who considered a CRE approach for binary choice models with panel data. Our model is different from that of Chamberlain (1980) in the sense that we do not specify the conditional expectation function of the time-invariant unobserved heterogeneity on time-varying regressors. In doing so, we alleviate the issue about model misspecification. The model parameters are identified under mind conditions. Then, we propose the PSMD approach to estimate the model. We establish the asymptotic theory for the PSMD estimator and show the bootstrap validity. Our PSMD estimator of the finite dimensional parameter is semiparametrically efficient once we use the optimal weighting matrix. The optimal weighting matrix is unknown, but it can be easily estimated in practice. The Monte Carlo simulation results confirm that our PSMD estimator performs well in finite samples, as it has small bias and standard deviation with a relatively small sample size.

While the semiparametric model proposed in this paper may reduce the possibility of model misspecification, we point out that our model relies on a distributional assumption on the latent error term. The distributional assumption is imposed for tractability of the model, and is not driven by some economic theory. There are several studies in the literature that relax such distributional assumptions in various contexts (e.g., Chen et al., 2006; Bierens, 2014; Han and Vytlacil, 2017; Han and Lee, 2019). It would be interesting to consider identification and estimation of a model where the distribution of the latent error term is unknown. We leave this important and interesting topic for future research.

A. MATHEMATICAL PROOF

For any positive real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ means that there exist a finite constant $C > 0$ and $N \in \mathbb{N}$ such that $a_n \leq Cb_n$ for all $n \geq N$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$, it is denoted by $a_n \asymp b_n$.

A.1. PROOF OF THEOREM 1

Proof. Note that

$$\begin{aligned}\Pr(Y_t = 1|\mathbf{X}) &= F_{\varepsilon_t|\mathbf{X}}(X_t^{'}\theta + h(\mathbf{X})) \\ &= F_{\varepsilon}\left(X_t^{'}\theta + h(\mathbf{X})\right).\end{aligned}$$

Since F_ε is strictly increasing over $\mathbb{R}$, there exists the inverse map $F_\varepsilon^{-1}(\cdot)$. Therefore,

$$F_\varepsilon^{-1}\left(\Pr(Y_t = 1|\mathbf{X})\right) = X_t^{'}\theta + h(\mathbf{X}), \tag{8}$$

and we have

$$F_\varepsilon^{-1}\left(\Pr(Y_t = 1|\mathbf{X})\right) - F_\varepsilon^{-1}\left(\Pr(Y_{t-1} = 1|\mathbf{X})\right) = \Delta X_t^{'}\theta. \tag{9}$$

Multiplying ΔX_t to the both sides of equation (9) and taking expectation, we obtain that

$$\mathbb{E}\left[\Delta X_t\left(F_\varepsilon^{-1}\left(\Pr(Y_t = 1|\mathbf{X})\right) - F_\varepsilon^{-1}\left(\Pr(Y_{t-1} = 1|\mathbf{X})\right)\right)\right] = \mathbb{E}\left[\Delta X_t \Delta X_t^{'}\right]\theta.$$

Under Assumption 1, $\mathbb{E}\left[\Delta X_t \Delta X_t^{'}\right]$ is invertible. Therefore,

$$\theta = \left(\mathbb{E}\left[\Delta X_t \Delta X_t^{'}\right]\right)^{-1}\mathbb{E}\left[\Delta X_t\left(F_\varepsilon^{-1}\left(\Pr(Y_t = 1|\mathbf{X})\right) - F_\varepsilon^{-1}\left(\Pr(Y_{t-1} = 1|\mathbf{X})\right)\right)\right],$$

implying that θ is identified. The identification of h is from equation (8) and identification of θ. □

A.2. PROOF OF THEOREM 2

Lemma 1. *Suppose that Assumptions 2–5 hold. Then, Assumption 2.6 in Chen and Pouzo (2009) is satisfied.*

Proof. We verify Assumptions 2.7 and 2.8 in Chen and Pouzo (2009). Assumption 2.7 in Chen and Pouzo (2009) is directly imposed by Assumptions 2, 3, and 4. Note that

$$\sup_{\alpha\in\mathscr{A}_n}\sup_{\mathbf{x}\in Supp(\mathbf{X})} Var\left(\rho\left(\mathbf{W};\alpha\right)|\mathbf{X}=\mathbf{x}\right)\leq 2$$

since

$$\begin{aligned}
& Var\left(\rho_t\left(\mathbf{W};\alpha\right)|\mathbf{X}=\mathbf{x}\right)\\
=&\mathbb{E}\left[\rho_t(\mathbf{W};\alpha)^2|\mathbf{X}=\mathbf{x}\right]\\
=&\mathbb{E}\left[Y_t^2|\mathbf{X}=\mathbf{x}\right]+2\mathbb{E}[Y_t|\mathbf{X}=\mathbf{x}]\cdot F_\varepsilon\left(x_t'\theta+h(\mathbf{x})\right)+\left(F_\varepsilon\left(x_t'\theta+h(\mathbf{x})\right)\right)^2\\
\leq&\left(F_\varepsilon\left(x_t'\theta_0+h_0(\mathbf{x})\right)+F_\varepsilon\left(x_t'\theta+h(\mathbf{x})\right)\right)^2\\
\leq&2
\end{aligned}$$

by the fact that $F_\varepsilon\left(X_t'\theta+h(\mathbf{X})\right)<1$ almost surely for all $t=1,2,...,T$ under Assumption 2.

Let $\xi_{0n}\equiv\sup_{\mathbf{x}\in Supp(\mathbf{X})}||p^{J_n}(\mathbf{x})||_E$. Under Assumption 4, $\xi_{0n}=O\left(J_n^{1/2}\right)$ by Newey (1997, p.151). By Assumption 5 and Newey (1997), there exists $(\pi_n^*)_n$ such that $||m(\cdot;\alpha)-p^{J_n}(\cdot)'\pi_n^*||_2^2=O\left(J_n^{-2p_m/Td_x}\right)$.

By Remark 2.1 in Chen and Pouzo (2009), Assumption 2.6 in Chen and Pouzo (2009) is satisfied with $\delta_{m,n}^2=\max\left\{\frac{J_n}{n},J_n^{-\frac{2p_m}{Td_x}}\right\}$. □

Proof of the theorem

Proof. We verify the conditions of Lemma 2.1 in Chen and Pouzo (2009). Under Assumption 1, the moment conditions in (3) holds if and only if $\theta=\theta_0$ and $h=h_0$. Therefore, Assumption 2.1 in Chen and Pouzo (2009) is satisfied under Assumptions 1, 2, and 3. Since $h_0\in\Lambda_{c_h}^{p_h}(Supp(\mathbf{X}))$ and the sieve space for $\mathscr{H}$ is the space of polynomial splines by Assumptions 3 and 4, there exists $(\beta_n^*)_n$ such that $||h_0-p^{k_n}(\cdot)'\beta_n^*||_\infty=O\left(k_n^{-p_h/T\cdot d_x}\right)=o(1)$. In addition, for each $t=1,2,...,T$, $m_t(\mathbf{X};\alpha)=F_\varepsilon\left(X_t'\theta_0+h_0(\mathbf{X})\right)-F_\varepsilon\left(X_t'\theta+h(\mathbf{X})\right)$ is continuous in α; and therefore, under Assumption 8, $\mathbb{E}\left[m(\mathbf{X};\alpha)'\cdot\Sigma(\mathbf{X})^{-1}m(\mathbf{X};\alpha)\right]$ is continuous at α_0 under $||\cdot||_\infty$. As a result, Assumptions 2.2 and 2.3 in Chen and Pouzo (2009) are satisfied. Since $\hat{P}_n(h)=P(h)=0$ for all $h\in\mathscr{H}$, Assumption 2.4 in Chen and Pouzo (2009) is met. Assumption 2.5 in Chen and Pouzo (2009) is satisfied by Assumption 8. By Lemma 1, Assumption 2.6 in Chen and Pouzo (2009) holds. In all, it follows from Lemma 2.1 in Chen and Pouzo (2009) that $||\hat{\alpha}_n - \alpha_0||_\infty = o_p(1)$. □

A.3. PROOF OF THEOREM 3

Proof. We use Lemma 2.3 in Chen and Pouzo (2009) to prove Theorem 3. Condition (i) of Assumption 2.9 in Chen and Pouzo (2009) is directly imposed by Assumption 9. Condition (ii) of Assumption 2.9 in Chen and Pouzo (2009) is implied by Lemma 1.

Note that under Assumption 8,

$$
\begin{aligned}
&||\alpha - \alpha_0||^2 \\
=&\mathbb{E}\left[\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[\alpha - \alpha_0]\right)^{'} \Sigma(\mathbf{X})^{-1}\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[\alpha - \alpha_0]\right)\right] \\
=&\mathbb{E}\left[\sum_{t\le T} f_\varepsilon(X_t^{'}\theta_0 + h_0(\mathbf{X}))^2(X_t^{'}(\theta - \theta_0) + (h - h_0))^2\right] \\
\lesssim&\mathbb{E}\left[(\theta - \theta_0)^{'}\left(\sum_{t\le T} f_\varepsilon(X_t^{'}\theta_0 + h_0(\mathbf{X}))^2 X_t X_t^{'}\right)(\theta - \theta_0) + (h - h_0)^2\sum_{t\le T} f_\varepsilon(X_t^{'}\theta_0 + h_0(\mathbf{X}))^2\right] \\
\lesssim&||\alpha - \alpha_0||_2^2 \\
\le&||\alpha - \alpha_0||_\infty^2,
\end{aligned}
$$

where the inequality in the third line holds by the c_r-inequality, and the inequality in the fourth line holds by Assumption 2(vii) and Assumption 2(viii). This leads to that condition (iii) of Assumption 2.9 in Chen and Pouzo (2009) is satisfied by Assumption 2. Therefore, Assumption 2.9 in Chen and Pouzo (2009) holds. It also implies that condition (i) of Assumption 2.10 in Chen and Pouzo (2009) is satisfied. Condition (ii) of Assumption 2.10 in Chen and Pouzo (2009) is directly imposed by Assumption 10. In all, Lemma 2.3 in Chen and Pouzo (2009) yields that

$$
||\hat{h}_n - h_0|| = O_p\left(\max\left\{\delta_{m,n}, k_n^{-\frac{p_h}{Td_x}}\right\}\right).
$$

Since $||\cdot|| \asymp ||\cdot||_2$ on $\mathscr{A}_{os}$ by Assumptions 2 and 9, this completes the proof. □

A.4. PROOF OF THEOREM 4

Proof. We verify the sufficient conditions of Theorem 3.1 in Chen and Pouzo (2009). Observe that for any $\alpha, \tilde{\alpha} \in \mathscr{A}_n$ such that $||\alpha - \tilde{\alpha}||_2 \leq \delta$

$$\begin{aligned}
\mathbb{E}\left[||\rho(W;\alpha) - \rho(W;\tilde{\alpha})||_E^2|\mathbf{X}\right] &\leq \sum_t \left|F_\varepsilon\left(X_t^{'}\theta + h(\mathbf{X})\right) - F_\varepsilon\left(X_t^{'}\tilde{\theta} + \tilde{h}(\mathbf{X})\right)\right|^2 \\
&\lesssim \sum_t \left|X_t^{'}(\theta - \tilde{\theta}) + h(\mathbf{X}) - \tilde{h}(\mathbf{X})\right|^2 \\
&\lesssim \sum_t \left(||X_t||_E^2 \cdot ||\theta - \tilde{\theta}||_E^2 + |h(\mathbf{X}) - \tilde{h}(\mathbf{X})|^2\right) \\
&\leq T\left(\max_{t\leq T} ||X_t||_E^2 + 1\right) \cdot ||\alpha - \tilde{\alpha}||_2^2.
\end{aligned}$$

By Assumption 2, condition (i) of Assumption 3.1 in Chen and Pouzo (2009) is satisfied with $r = 2$, $\kappa = 1$, and $b(\mathbf{x}) \equiv T\left(\max_{t\leq T} ||x_t||_E^2 + 1\right)$. Since $|\rho_t(\mathbf{W})| \leq 2$ for all $t = 1, 2, ..., T$, condition (ii) of Assumption 3.1 in Chen and Pouzo (2009) is satisfied. Since $\delta_{2,n} = o\left(n^{-1/4}\right)$ by condition (ii) of Assumption 13 and $\delta_{2,n} \asymp \delta_n$, we have $\delta_n^2 \cdot \delta_{2,n}^2 \asymp \delta_{2,n}^4 = o\left(n^{-1}\right)$, which implies condition (iii) of Assumption 3.1 in Chen and Pouzo (2009).

Condition (i) of Assumption 3.2 in Chen and Pouzo (2009) is directly imposed by Assumption 3. Note that since $Var(\rho_t(\mathbf{W};\alpha_0)|\mathbf{X})$ is bounded away from zero, $Var(\rho(\mathbf{W};\alpha_0)|\mathbf{X})$ is positive definite by Assumptions 2 and 11. In addition, Assumption 13 implies condition (iii) of Assumption 3.2 in Chen and Pouzo (2009).

Assumption 3.3 in Chen and Pouzo (2009) is implied by Assumptions 8 and 13. Under Assumptions 4 and 14, it follows that

$$\mathbb{E}\left[\left|\left|\frac{d\tilde{m}(\mathbf{X};\alpha_0)}{d\alpha}[v^*] - \frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v^*]\right|\right|_E^2\right] = O\left(J_n^{-\frac{2p_d}{Td_x}}\right)$$

by Newey (1997). Therefore, condition (i) of Assumption 3.4 in Chen and Pouzo (2009) is satisfied by Assumptions 4 and 14. Since $\Sigma(\mathbf{X}) = I_T$ by Assumption 8, condition (ii) of Assumption 3.4 in Chen and Pouzo (2009) is also met. Condition (b) of Assumption 3.5 in Chen and Pouzo (2009) is implied by Assumption 3.

By Assumption 12, $m(\mathbf{X};\alpha)$ is twice pathwise differentiable in $\alpha \in \mathscr{N}_{0n}$.

Note that for each $t = 1, 2, ..., T$,

$$\begin{aligned}
\frac{dm_t(\mathbf{X};\alpha_0)}{d\alpha}[v] &= -f_\varepsilon\left(X_t^{'}\theta_0 + h_0(\mathbf{X})\right)\left(X_t^{'}v_\theta + v_h\right), \\
\frac{d^2m_t(\mathbf{X};\alpha_0)}{d\alpha d\alpha}[v,v] &= -f_\varepsilon^{'}\left(X_t^{'}\theta_0 + h_0(\mathbf{X})\right)\left(X_t^{'}v_\theta + v_h\right)^2,
\end{aligned}$$

where $f_\varepsilon^{'}(x) \equiv \frac{df_\varepsilon(x)}{dx}$. Under Assumptions 2, 3, 4, and 12,

$$\mathbb{E}\left[\sup_{\alpha\in\mathscr{N}_{0n}}\left|\frac{d^2m(\mathbf{X};\alpha)}{d\alpha d\alpha}[v_n^*,v_n^*]\right|\right] \lesssim \mathbb{E}\left[\max_{t\leq T}\left(X_t^{'}v_{\theta,n}^* + v_{h,n}^*\right)^2\right] < \infty,$$

and thus, condition (i) of Assumption 3.6 in Chen and Pouzo (2009) is satisfied. For any $\alpha \in \mathscr{N}_{0n}$,

$$\left|\left|\frac{dm(\mathbf{X};\alpha)}{d\alpha}[v_n^*] - \frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v_n^*]\right|\right|_E^2 \lesssim \left(\sum_t^T\left(X_t^{'}v_{\theta,n}^* + v_{h,n}^*\right)^2\right)\cdot(\alpha-\alpha_0)^2$$

since $f_\varepsilon^{'}(\cdot)$ is bounded over $\mathbb{R}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\alpha\in\mathscr{N}_{0n}}\left|\left|\frac{dm(\mathbf{X};\alpha)}{d\alpha}[v_n^*] - \frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v_n^*]\right|\right|_E^2\right] \\
&\lesssim \mathbb{E}\left[\left(\sum_t^T\left(X_t^{'}v_{\theta,n}^* + v_{h,n}^*\right)^2\right)\right]\cdot\delta_{2,n}^2 \\
&= o\left(n^{-1/2}\right)
\end{aligned}$$

by Assumption 13, which implies condition (ii) of Assumption 3.6 in Chen and Pouzo (2009). Finally, since $||\alpha-\alpha_0||_2 \leq \delta_{2,n}$ and $||\bar{\alpha}-\alpha_0||_2 \leq \delta_{2,n}$ for any $\alpha \in \mathscr{N}_{0n}$ and $\bar{\alpha} \in \mathscr{N}_0$ by the definitions of $\mathscr{N}_{0n}$ and $\mathscr{N}_0$, it can be shown that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v^*]\right)^{'}\Sigma(\mathbf{X})^{-1}\cdot\left(\frac{dm(\mathbf{X};\bar{\alpha})}{d\alpha}[\bar{\alpha}-\alpha_0] - \frac{dm(\mathbf{X};\alpha)}{d\alpha}[\alpha-\alpha_0]\right)\right] \\
&\leq\sqrt{\mathbb{E}\left[\left|\left|\frac{dm(\mathbf{X};\alpha_0)}{d\alpha}[v^*]\right|\right|_E^2\right]}\cdot\sqrt{\mathbb{E}\left[\left|\left|\left(\frac{dm(\mathbf{X};\bar{\alpha})}{d\alpha}[\bar{\alpha}-\alpha_0] - \frac{dm(\mathbf{X};\alpha)}{d\alpha}[\alpha-\alpha_0]\right)\right|\right|_E^2\right]} \\
&\lesssim\sqrt{\mathbb{E}\left[\left|\left|\frac{d^2m(\mathbf{X};\tilde{\alpha})}{d\alpha d\alpha}[\bar{\alpha}-\alpha_0,\alpha-\alpha_0]\right|\right|_E^2 + \left|\left|\frac{d^2m(\mathbf{X};\dot{\alpha})}{d\alpha d\alpha}[\alpha-\alpha_0,\alpha-\alpha_0]\right|\right|_E^2\right]} \\
&\lesssim\delta_{2,n}^2,
\end{aligned}$$

where $\tilde{\alpha}$ lies between $\bar{\alpha}$ and α_0, and $\dot{\alpha}$ lies between α and α_0. Therefore, by condition (ii) of Assumption 13, condition (iii) of Assumption 3.6 in Chen and Pouzo (2009) is met.

In all, it follows from Theorem 3.1 in Chen and Pouzo (2009) that

$$\sqrt{n}\lambda'\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}N\left(0,\lambda'V\lambda\right).$$

By the Cramer-Wold device, we conclude that $\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}N\left(0,V\right)$. □

A.5. PROOF OF THEOREM 5

Proof. This is a direct consequence of Theorem 3.2 in Chen and Pouzo (2009). □

A.6. PROOF OF THEOREM 6

Proof. Assumption 16, together with Assumption 13, is sufficient for condition (ii) of Assumption 3.3 in Chen and Pouzo (2009) (i.e., $\delta_n\times\delta_{\Sigma,n}=o\left(n^{-1/2}\right)$). Applying Theorem 3.1 in Chen and Pouzo (2009) with the weighting matrix $\hat{\Sigma}_{0,n}$ establishes the asymptotic normality result in the theorem. The semiparametric efficiency follows from Theorem 6.1 in Ai and Chen (2003). □

REFERENCES

Abrevaya, J. and Dahl, C. M. (2008). "The effects of birth inputs on birthweight: evidence from quantile estimation on panel data," *Journal of Business & Economic Statistics* 26, 379-397.

Ai, C. and Chen, X. (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica* 71, 1795-1843.

Arellano, M. and Bonhomme, S. (2016). "Nonlinear panel data estimation via quantile regressions" *The Econometrics Journal* 19, C61-C94.

Bester, C. and Hansen, C. (2009). "Identification of marginal effects in a nonparametric correlated random effects model" *Journal of Business & Economic Statistics* 27, 235-250.

Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Springer

Bierens, H. J. (2014). "Consistency and asymptotic normality of sieve ML estimators under low-level conditions," *Econometric Theory* 30, 1021-1076.

Carrasco, M., J-P. Florens, and E. Renault (2007)."Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of Econometrics*, 6B, 5633-5751.

Chamberlain, G. (1980). "Analysis of covariance with qualitative data," *The Review of Economic Studies* 47, 225-238.

Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics* 6B, 5549-5632.

Chen, X., Fan, Y. and Tsyrennikov, V. (2006). "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association* 101, 1228-1240.

Chen, X. and Liao, Z. (2014). "Sieve M inference on irregular parameters," *Journal of Econometrics* 182, 70-86.

Chen, X. and D. Pouzo (2009). "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics* 152, 46-60.

Chen, X. and D. Pouzo (2012). “Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals,” *Econometrica* 80, 277-321.

Fernández-Val, I. and Weidner, M. (2018), “Fixed effects estimation of large-T panel data models,” *Annual Review of Economics* 10, 109-138.

Han, S. and Lee, S. (2019). “Estimation in a generalization of bivariate probit models with dummy endogenous regressors,” *Journal of Applied Econometrics* 34, 994-1015.

Han, S. and Vytlacil, E. (2017). “Identification in a generalization of bivariate probit models with dummy endogenous regressors,” *Journal of Econometrics* 199, 63-73.

Horowitz, J.L. (2014). “Ill-posed inverse problems in economics,” *Annual Review of Economics* 6, 21-51.

Lee, S. (2022). “Nonparametric Estimation of a Triangular System of Equations for Quantile Regression,” *Journal of Economic Theory and Econometrics* 33, 31-53.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*, Princeton University Press.

Mundlak, Y. (1978). “On the pooling of time series and cross section data,” *Econometrica* 46, 69-85.

Newey, W. K. (1990). “Semiparametric efficiency bounds,” *Journal of Applied Econometrics* 5, 99-135.

Newey, W. K. (1997). “Convergence rates and asymptotic normality for series estimators,” *Journal of Econometrics* 79, 147-168.

Neyman, J. and Scott, E. L. (1948). “Consistent estimates based on partially consistent observations,” *Econometrica* 16, 1-32.

Song, H. (2015). “A univariate sieve density estimation based on a simulated Kolmogorov-Smirnov test,” *Journal of Economic Theory and Econometrics* 26, 26-43.

Wooldridge, J. M. (1995). “Selection corrections for panel data models under conditional mean independence assumptions,” *Journal of Econometrics* 68, 115-132.

Wooldridge, J. M. (2005). "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics* 20, 39-54.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, MIT Press.