Nonlinear Panel Data Models with Distribution-Free Correlated Random Effects*

Yu-Chin Hsu† Ji-Liang Shiu‡

January 14, 2018

Abstract

This paper investigates identification and estimation of parametric nonlinear panel data models with distribution-free correlated random effects (CRE). Under the Mundlak-type CRE specification without distributional assumptions, we first show that the average (or integrated) likelihood is the convolution of the parametric panel data model and the conditional distribution of the unobserved heterogeneity which can be recovered by means of Fourier transformation. We then construct a semi-parametric family of average likelihood functions of observables by combining the parametric panel data model with the recovered distribution of the unobserved heterogeneity. We show that the parameter vector is identifiable and based on the identification result, we propose a sieve maximum likelihood estimator which is root-n consistent and asymptotically normal. Compared with the conventional parametric CRE approaches, the advantage of our method is that it is not subject to the misspecification on the distribution of the CRE. We investigate the finite sample properties of the proposed estimator through a Monte Carlo study and apply our method to estimate the persistence effects of union membership.

Keywords: Nonlinear panel data models, Semi-parametric identification, Correlated random effects, Sieve maximum likelihood estimator

*Helpful comments by Arthur Lewbel, and Matthew Shum are acknowledged. The authors are solely responsible for any remaining errors.

†Institute of Economics, Academia Sinica, Email: ychsu@econ.sinica.edu.tw.

‡Institute for Economic and Social Research, Jinan University. Email: jishiu.econ@gmail.com.

1. Introduction

One of the challenges in the panel data literature is how to model the unobserved heterogeneity across individuals when the time dimension is fixed. There is a fundamental difference between the linear and nonlinear models. For linear panel data models with an additive unobserved heterogeneity, one can use the fixed-effects formulation to eliminate the unobserved effects so there will be no requirement to postulate the distribution of the unobserved heterogeneity. Consistent estimators can be obtain by generalized method of moments methods after one applies within transformation. We refer to Baltagi (2008), Wooldridge (2010), and Hsiao (2015) for more complete literature reviews.

In nonlinear models it is not clear how to remove the unobserved heterogeneity[1] and there are two main treatments of the unobserved heterogeneity. One is to treat the unobserved heterogeneity as fixed parameters (fixed effects) and the other is to treat it as random variables (random effects). There are a surprising amount of differences in the identification and estimation between these two treatments. When the unobserved heterogeneity is modeled as fixed effects, the dimension of unknown parameters increases at the same rate as the sample size so estimators are often subject to an incidental parameter problem and a standard maximum likelihood estimator causes bias. Honoré and Kyriazidou (2000) generalized conditional maximum likelihood approaches of Andersen (1970) and Rasch (1993) to estimate the parameters of dynamic discrete choice logit models with strictly exogenous explanatory variables.[2] Chamberlain (2010) considered the identification of binary response models when the time dimension is fixed and the distribution of individual effects is unrestricted. He showed that identification is only possible in the logistic case.[3]

On the other hand, in the random effects approach, one would first specify a conditional distribution of the parametric nonlinear panel data model, i.e., a conditional distribution of

[1] Some progress has been made in this direction including Arellano and Carrasco (2003), Altonji and Matzkin (2005), Hoderlein and Mammen (2007), Bester and Hansen (2009), Chernozhukov, Fernandez-Val, Hoderlein, Holzmann, and Newey (2015), Hoderlein and White (2012), Graham and Powell (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013) and Browning and Carro (2014).

[2] Honoré and Lewbel (2002) provided a set of conditions for identification of the parameters of a binary choice model allowing for general predetermined explanatory variables and propose a root-n consistent GMM estimator to estimate the parameters.

[3] As discussed in Arellano and Bonhomme (2011), the identification problem is related to the situations where the information of outcomes is not enough to identify the unobserved heterogeneity. In the binary response model, the support of outcomes is less rich than the support of the unobserved heterogeneity.

the dependent variable, Y_t, conditional on a K-dimensional vector of time-varying explanatory variables, X_t, and an individual unobserved heterogeneity, C,:

$$f_{Y_t|X_t,C}(y_t|x_t,c;\theta), \text{ for all } t=1,...,T, \tag{1}$$

where y_t, x_t and c are points in the supports of Y_t, X_t and C, respectively, and θ is a vector of unknown parameters to be estimated. The second step in a conventional parametric random effects approach is to complete the model by specifying the statistical relationship between the unobserved heterogeneity and the observed covariates. To be specific, denote $x=(x_t,\ldots,x_T)$ a vector of explanatory variables in all periods and $f_{C|X}(c|x;\beta)$ as the parametric distribution of C conditional on the explanatory variables X. An average likelihood according to these two parametric densities can be constructed as follows:

$$f_{Y|X}(y|x;\theta,\beta)=\int\left(\prod_{t=1}^{T}f_{Y_t|X_t,C}(y_t|x_t,c;\theta)\right)f_{C|X}(c|x;\beta)dc. \tag{2}$$

The average likelihood in Eq. (2) is fully parametric in that it depends on a finite number of parameters, and the estimation and inference is possible under standard parametric framework. For example, Wooldridge (2005) handled the initial conditions problem of dynamic panel data problem by specifying the conditional distribution of the unobserved heterogeneity to be normal distributed with a mean which is a linear combination of the initial value, and exogenous explanatory variables. Alvarez and Arellano (2003) used a similar specification for models with large time and cross-sectional dimensions. Arellano and Bonhomme (2009) focused on estimators that maximize an average likelihood that assigns weights to different values of the unobserved heterogeneity. They provided a characterization of the class of weights that produce first-order unbiased estimators.

The disadvantage of parametric random effects approach is that the misspecification of $f_{C|X}(c|x;\beta)$ generally results in inconsistent estimates. To avoid this, we can alternatively characterize the identified region containing the true parameter so that we would lose point-identification of the parameters. For example, Honoré and Tamer (2006) relaxed the distributional assumption of the initial condition and calculated bounds on parameters in panel dynamic discrete choice models. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) showed

that bounds for marginal effects in nonlinear panel models can be tightened rapidly as the number of time series observations grows. As a result, a general way to handle the distribution misspecification of the average likelihood in Eq. (2) for fixed T while retaining point-identification of the parameters remains intangible.[4]

To fill this gap in the literature, we provide a data-driven specification of conditional distributions of the unobserved heterogeneity which is internally consistent with the parametric nonlinear panel data models in Eq. (1) and at the same time, to retain point-identification of the parameters of interest. To be specific, we consider a correlated random effects (CRE) approach without fully specifying the distribution of the unobserved heterogeneity conditional on the explanatory variables. Let $\overline{W}$ be a vector of time-invariant observed variables. Under the Mundlak-type specification, $C = \lambda\overline{W} + V$, Eq. (2) can be written as

$$f(y|x,\overline{w};\theta,\lambda) = \int \left(\prod_{t=1}^{T} f_{Y_t|X_t,C}(y_t|x_t,c;\theta) \right) f_V(c - \overline{w}\lambda) dc, \tag{3}$$

so the average likelihood is the convolution of the parametric nonlinear panel data models and the conditional distribution of the unobserved heterogeneity. In turn, the characteristic function of the conditional distribution of the unobserved heterogeneity can be obtained by the quotient of two characteristic functions related to a density of observables and the parametric panel data model at the true parameter. Next, we extend the relation of the characteristic function to parameters other than the true parameter and apply Fourier inversion formula to the extended characteristic function to devise a parametric distribution of the unobserved heterogeneity conditional on exogenous variables. Combining the parametric nonlinear panel data model and the recovered parametric distribution of the unobserved heterogeneity, we can construct a semi-parametric average likelihood of observables. The semi-parametric average likelihood is correctly specified in the sense that the semi-parametric average likelihood evaluated at the true parameter is equal to the density of observables. We can regard the specification as a data-driven one because the recovered semi-parametric distribution of the unobserved heterogeneity is based on the parametric nonlinear panel data model and the density of observables.

We then show that its parameter vector is identifiable by a standard maximum likelihood

[4] Section 3.3 of Arellano and Bonhomme (2011) provided detailed discussions.

condition, the negative definiteness of the information matrix. We propose a sieve maximum likelihood (henceforth sieve ML) estimator based on the identification result and show that the proposed estimator is consistent and asymptotically normal. More importantly, we propose a Hausman-type test to test the distributional assumption in the conventional parametric random effects approach. We also extend our method to dynamic nonlinear models and show that the average partial effects can be identified.

The key insight of our approach is to utilize the information of the observed time-invariant variable as a source of identification for a time-invariant structure of heterogeneity. Similar strategies are also considered in Honoré and Lewbel (2002), Hu and Shum (2012), and Shiu and Hu (2013). Our identification strategy is related to the literature on nonparametric deconvolution including the measurement error models (see Schennach (2004); Schennach (2007); Hu and Ridder (2010); Hu and Ridder (2012)), and panel data models (see Evdokimov (2011); Arellano and Bonhomme (2012)), etc. Evdokimov (2011) established nonparametric identification of a panel data model with nonadditive unobserved heterogeneity and developed a nonparametric estimation procedure. Arellano and Bonhomme (2012) considered random coefficients panel data models where the coefficients can be arbitrarily correlated with the covariates and obtained identification of the density of individual effects.

The rest of the paper is organized as follows. In Section 2, we present the identification results of an internally consistent likelihood function. In Section 3, we propose a sieve ML estimator. Section 4 provides a specification test for a parametric specification of the CRE model, an extension of the proposed method to dynamic nonlinear panel data models and identification results of partial effects. In Section 5, the finite-sample properties of the sieve ML estimator are investigated via Monte Carlo simulations. In Section 6, we apply our method in an empirical study to estimate the persistence effects of union membership using panel data. Section 7 concludes. The Appendix contains technical proofs of results.

2. Identification

For $t = 1, \ldots, T$, let Y_t denote the dependent variable of interest and X_t denote a K-dimensional vector of possibly time-varying explanatory variables with supports $\mathscr{Y}_t$ and $\mathscr{X}_t$, respectively. Let $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_T$, and $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_T$. Also, let C denote an individual unobserved

heterogeneity C with support $\mathscr{C}$. Consider the following parametric panel data model:

$$f_{Y_t|X_t,C}(y_t|x_t,c;\theta), \text{ for all } t=1,...,T, \tag{4}$$

where $y_t \in \mathscr{Y}_t$, $x_t \in \mathscr{X}_t$ and $c \in \mathscr{C}$. Also, $\theta \in \Theta$ is a vector of parameters which specifies the structure of the model and Θ is the parameter space.

The panel data model in Eq. (4) may be derived from more primitive econometric model. For example, for the following single-index binary-choice model:

$$Y_t = 1(X_t\theta + C + \varepsilon_t \geq 0), \text{ for all } t=1,...,T, \tag{5}$$

where $1(\cdot)$ is the indicator function, and ε_t is independent of X with a known time-specific distribution function F_{ε_t}, its corresponding conditional distribution is:

$$f_{Y_t|X_t,C}(y_t|x_t,c;\theta) = \left(1 - F_{\varepsilon_t}[-(x_t\theta + c)]\right)^{y_t} F_{\varepsilon_t}[-(x_t\theta + c)]^{1-y_t}. \tag{6}$$

However, the identification result of the panel data model in Eq. (4) developed below can be applied to other parametric nonlinear panel data models.

2.1. Assumptions and Results

We make assumptions for identification in this section. We first assume that the model in Eq. (4) is correctly specified.

Assumption 2.1. *Assume that (i) for $t=1,\ldots,T$, the density $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ is known up to a vector of parameters and is uniformly bounded above for all $y_t \in \mathscr{Y}_t$, $x_t \in \mathscr{X}_t$, $c \in \mathscr{C}$ and $\theta \in \Theta$;*
(ii) there exists a unique vector of parameters $\theta_0 \in \Theta$ such that $f_{Y_t|X_t,C}(y_t|x_t,c;\theta_0)$ is equal to the population density function, $f_{Y_t|X_t,C}(y_t|x_t,c)$;
(iii) the parameter space, Θ, is a compact subset of $\mathbb{R}^{d_\theta}$.

In order to control the possible correlation between X_t and C, we use a correlated random effects (CRE) condition to model the conditional mean of the unobserved effect as a linear function of the time average of some explanatory variables in X_t. Let $\overline{W} = \frac{1}{T}\sum_{t=1}^{T}(X_{t1},...,X_{tK_1}) = (\overline{X}_1,...,\overline{X}_{K_1})$ as the time average of the first K_1 explanatory variables in X_t with support $\overline{\mathscr{W}}$.

Assumption 2.2. *(Correlated Random Effects (CRE))*

Assume that there exists K_1*-dimensional vector of nonzero coefficients* $\lambda_0 = (\lambda_{01}, ..., \lambda_{0K_1})'$ *in* Λ *that is a compact subset of* R^{K_1} *such that*

$$C = \overline{W}\lambda_0 + V, \tag{7}$$

where the remainder term V *is independent of* $\overline{W}$.

Note that in contrast to conventional fully parametric approaches, Assumption 2.2 does not impose any distributional restriction on the remainder term V, i.e., we consider weaker restrictions on the unobserved individual-specific effect. Denote f_V as the PDF of the remainder term V. The independence between $\overline{W}$ and V, and the additive structure in Eq. (7) together imply that $f_{C|\overline{W}}(c|\overline{w}) = f_V(c - \overline{w}\lambda_0)$. We note that an alternative specification of the CRE condition is that $\overline{W}$ includes some time-invariant observed variables not in X_t.

Assumption 2.3. *(Movement of the Correlated Unobserved Effects)*

Assume that (i) $f_{Y|X,\overline{W},C}(y|x,\overline{w},c) = \prod_{t=1}^{T} f_{Y_t|X_t,\overline{W},C}(y_t|x_t,\overline{w},c)$ *for all* $(y_t, x_t, \overline{w}, c) \in \mathcal{Y} \times \mathcal{X} \times \overline{\mathcal{W}} \times \mathcal{C}$*;*

(ii) $f_{Y_t|X_t,\overline{W},C}(y_t|x_t,\overline{w},c) = f_{Y_t|X_t,C}(y_t|x_t,c)$ *for all* $(y_t, x_t, \overline{w}, c) \in \mathcal{Y}_t \times \mathcal{X}_t \times \overline{\mathcal{W}} \times \mathcal{C}$*;*

(iii) the conditional distribution of unobserved heterogeneity satisfies $f_{C|X,\overline{W}}(c|x,\overline{w}) = f_{C|\overline{W}}(c|\overline{w})$ *for all* $(c, x, \overline{w}) \in \mathcal{C} \times \mathcal{X} \times \overline{\mathcal{W}}$.

Although the variable $\overline{W}$ is observed and related to the dependent variable Y_t, Assumption 2.3(ii) requires that it does not provide any more information on Y_t when the regressors X_t, and C are controlled. Assumption 2.3(iii) requires that the time invariant unobservable C conditional on X and $\overline{W}$ does not depend upon X, and is connected with X only through the time average term $\overline{W}$. Under Assumption 2.2, a sufficient condition for Assumption 2.3(iii) is that the remainder error V is independent of X. Denote the parametric conditional joint density as $f_{Y|X,C}(y|x,c;\theta) = \prod_{t=1}^{T} f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$.

Assumption 2.4. *(Well Defined Characteristic Function)*

Assume that (i) there exists a constant c_1 *such that* $\sum_{j=1} \int_{\overline{\mathcal{W}}_j} f_{Y|X,\overline{W}}(y|x,\overline{w}) d\overline{w}_j < c_1 < \infty$ *for all* $(y, x, \overline{w}) \in \mathcal{Y} \times \mathcal{X} \times \overline{\mathcal{W}}$*;*

(ii) there exists a constant c_2 *such that* $\int_{\mathcal{C}} f_{Y|X,C}(y|x,c;\theta) dc < c_2 < \infty$ *for all* $(y, x) \in \mathcal{Y} \times \mathcal{X}$ *and all* $\theta \in \Theta$*;*

(iii) there exists a weighting function $\Omega(y,x)$ over $\mathscr{Y}\times\mathscr{X}$ such that for all $\xi\in\mathbb{R}$ and for all $\lambda\in\Theta$,

$$\left|\int_{\mathscr{C}} e^{-i\xi c}\left(\int_{\mathscr{Y}\times\mathscr{X}} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydx\right)dc\right|>0. \tag{8}$$

Assumptions 2.4(i) & (ii) ensure the characteristic functions are well defined, implying that the characteristic functions $\sum_{j=1}^{K_1}\left(\lambda_j\int_{\overline{\mathscr{W}}_j} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_j\right)$ for $(y,x,\overline{w})\in\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}$ and $\lambda\in\Lambda$, and $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc$ for all $(y,x)\in\mathscr{Y}\times\mathscr{X}$ and $\theta\in\Theta$ are both finite. Because $f_{Y|X,C}(y|x,c;\theta)$ is uniformly bounded for all $\theta\in\Theta$ by Assumption 2.1, if the support of the unobserved heterogeneity $\mathscr{C}$ is compact then Assumption 2.4(ii) holds. Thus, the density $f_{Y|X,C}(y|x,c;\theta)$ for the binary-choice model in Eq. (6) with compact unobserved effects satisfies Assumption 2.4(ii) and $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)dc$ is well-defined. When $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ can be written in the form $f(y_t-m(x_t,c;\theta))$, where f is a known density and $|\frac{\partial}{\partial c}m(x_t,c;\theta)|>a>0$, Assumption 2.4(ii) holds.[5] However, requiring the support of the unobserved heterogeneity to be equal to the real line may fail Assumption 2.4(ii) for binary-choice models and censored models. The characteristic functions in Eq. (8) appear as denominators in our identifying formula and Assumption 2.4(iii) rules out zero denominator. Note that all conditions in Assumption 2.4 are testable since they involve the density of observables and the parametric nonlinear panel data model.

Let $\alpha=(\theta,\lambda)$, $\alpha_0=(\theta_0,\lambda_0)$ and $\mathscr{A}=\Theta\times\Lambda$. Under Assumption 2.4, we can construct a semi-parametric family of functions related to the characteristic function of the remainder term V in Eq. (A.9) in Appendix.

Assumption 2.5. *(Continuous Parameter Structure)*

Assume that (i) The parametric panel data density function $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ is continuous at θ for all $\theta\in\Theta$ and $t=1,...,T$;

(ii) the semi-parametric family of functions $\{\phi_{V;\alpha}(\xi):\alpha\in\mathscr{A}\}$ defined in Eq. (A.9) belongs to $L^1(\mathbb{R})$.

Under Assumption 2.5, we can apply Fourier Inversion Formula in Proposition A.1 to the semi-parametric family of functions $\{\phi_{V;\alpha}(\xi):\alpha\in\Theta\times\Lambda\}$ in Eq. (A.9) to construct a semi-parametric family of density functions $\{f_{C|\overline{W}}(c|\overline{w};\alpha):\alpha\in\mathscr{A}\}$ in Eq. (A.15).[6] Because the semi-parametric

[5] Consider $|\int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)dc| = |\int_{\mathscr{C}} f(y-m(x,c;\theta))dc| = |\int f(u)\frac{-du}{\frac{\partial m}{\partial c}}| < \frac{1}{a}\int f(u)du<\infty$, where $u=y-m(x,c;\theta)$ and $|\frac{1}{\frac{\partial m}{\partial c}}|<\frac{1}{a}$.

[6] Lemma A.1 shows that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is a density function over $\mathscr{C}$.

characteristic function $\phi_{V;\alpha}(\xi)$ is derived from the data and the parametric nonlinear panel data model, $f_{C|\overline{W}}(c|\overline{w};\alpha)$ can be regarded as internally consistent semi-parametric distribution of the unobserved heterogeneity.[7]

The parameter structure is then described by a $(d_\theta + K_1)$-dimensional vector associated with the panel data density function $f_{Y|X,C}(y|x,c;\theta)$ and the conditional distribution of the unobserved heterogeneity $f_{C|\overline{W}}(c|\overline{w};\alpha)$. For the identification in the parameter structure, we have to distinguish the true parameter α_0 from other parameters in the neighborhood of α_0. This implies that there is a unique vector of parameters associated with each population structure in the parameter space $\mathscr{A}$.

Definition 2.1. *(i) Two vectors of parameters, $\alpha_0 = (\theta_0, \lambda_0)$ and $\widetilde{\alpha} = (\widetilde{\theta}, \widetilde{\lambda})$ in $\mathscr{A} \subset \mathbb{R}^{d_\theta + K_1}$ are observationally equivalent if $f_{Y|X,C}(y|x,c;\theta_0) = f_{Y|X,C}(y|x,c;\widetilde{\theta})$ and $f_{C|\overline{W}}(c|\overline{w};\alpha_0) = f_{C|\overline{W}}(c|\overline{w};\widetilde{\alpha})$ for all $(y,x,\overline{w},c) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}} \times \mathscr{C}$ with probability one at the probability distribution of the random variable $(Y,X,\overline{W},C)$.*
(ii) A vector of parameters α_0 is said to be identifiable if there exists an open neighborhood of α_0 in $\mathscr{A}$ containing no other vectors of parameters observationally equivalent to α_0.

Next, we provide sufficient conditions for the identification of α_0. First, combine the density $f_{C|\overline{W}}(c|\overline{w};\alpha)$ with the parametric panel data model $f_{Y|X,C}(y|x,c;\theta)$ to construct the following internally consistent semi-parametric density function of observable variables:

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) \underbrace{f_{C|\overline{W}}(c|\overline{w};\alpha)}_{\substack{\text{constructed from} \\ f(y|x,\overline{w}) \text{ and} \\ f_{Y|X,C}(y|x,c;\theta)}} dc. \tag{9}$$

The equation (9) is called a semi-parametric density function, because it also depends on the density of observables $f(y|x,\overline{w})$, which is not parametrically specified. As we will see in Appendix B, we will apply Fourier transformations to combine the parameter structure of $f_{Y|X,C}(y|x,c;\theta)$ with $f(y|x,\overline{w})$ and construct $f_{C|\overline{W}}(c|\overline{w};\alpha)$ under the CRE specification. The semi-parametric density function is correctly specified because $f(y|x,\overline{w};\alpha_0) = f(y|x,\overline{w})$.

We need an identification condition on the basis of sample information to pin down α_0 and the information conditions to distinguish between the parametric structures. Specifically,

[7] See details in Eq. (A.9).

the identification of the parametric system is approached via the concavity of the conditional Kullback-Leibler information criterion evaluated at α_0. Define

$$K(\alpha; x, \overline{w}) = \mathrm{E}\Big[\log\left(\frac{f(Y|X, \overline{W}; \alpha)}{f(Y|X, \overline{W}; \alpha_0)} \right) \Big| X = x, \overline{W} = \overline{w} \Big]$$

where the expectation is taken with respect to $f(y|x, \overline{w}; \alpha_0)$. It follows that a sufficient condition for the existence of a unique maximum is that the first derivative $K(\alpha; x, \overline{w})$ evaluated at α_0 is equal to zero and the second derivative of $K(\alpha; x, \overline{w})$ evaluated at α_0 is negative definite. Differentiating $K(\alpha; x, \overline{w})$ with respect to α_j for $j = 1, ..., d_\theta + K_1$, we have the gradient of $K(\alpha; x, \overline{w})$ being a $(d_\theta + K_1)$-dimensional vector,

$$\frac{\partial}{\partial \alpha} K(\alpha; x, \overline{w}) = \left(\frac{\partial K(\alpha; x, \overline{w})}{\partial \alpha_1}, ..., \frac{\partial K(\alpha; x, \overline{w})}{\partial \alpha_{d_\theta + K_1}} \right)'. \tag{10}$$

The matrix of the second derivative of $K(\alpha; x, \overline{w})$ can be written as minus outer product of the gradient of the log likelihood:

$$K''(\alpha_0; x, \overline{w}) = -\mathrm{E}\Big[\frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)\big|_{\alpha=\alpha_0} \cdot \frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)'\big|_{\alpha=\alpha_0} \Big| X = x, \overline{W} = \overline{w} \Big]. \tag{11}$$

Assumption 2.6. *(Concave Parameter Structure)*
Assume that the information matrix $K''(\alpha_0; x, \overline{w})$ in Eq. (11) is negative definite for $(x, \overline{w}) \in \mathscr{X} \times \overline{\mathscr{W}}$ with probability one, and the elements of the matrix exist and are continuous in $\mathscr{A}$.

Theorem 2.1. *Under Assumptions 2.1-2.6, the population parameters of the parametric panel data density in Eq. (4) and the correlated random effects in Assumption 2.2, θ_0 and λ_0, are identifiable from the joint distribution of a panel data sample $\{Y_t, X_t\}$ for $t = 1, 2, ..., T$.*

3. Sieve Maximum Likelihood Estimation

The identification results in Section 2 are constructive in that we can propose a sieve Maximum Likelihood (ML) estimator for the parameter $\alpha = (\theta, \lambda)$ based on the parametric specification for the density function of observable variables in Eq. (A.1). The proposed estimator will be semi-parametric in the sense that the distribution of the CRE remainder error V in Assumption 2.2

is not specified. As shown in Eq. (A.1), we consider

$$f_{Y|X,\overline{W}}(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_V(c - \overline{w}\lambda_0) dc. \tag{12}$$

The fact that the population parameters α_0 and the distribution of the remainder error f_V are identified implies that α_0 and f_V solve

$$\max_{\alpha, f_V} \mathrm{E}\Big[\log f(Y|X,\overline{W};\alpha,f_V)\Big]. \tag{13}$$

Therefore, one could maximize the likelihood with respect to θ, λ, and f_V, where the unknown density f_V is estimated using a series estimator. We consider Hermite polynomial series as our sieve basis functions for the nonparametric nuisance components, $f_V(\cdot)^{1/2}$. Denote $\phi(\cdot)$ and $H_i(\cdot)$ as the i-th order Hermite polynomial and the PDF of standard normal. It follows that $h_i(\cdot) = H_i(\cdot)\phi(\cdot)$ form an orthogonal series of the square-integral function space. A Hermite polynomial series estimator of $f_V(\cdot)^{1/2}$ can be constructed by

$$\widetilde{f}_V^{1/2}(v) = \sum_{i=0}^{J} \beta_i h_i(v). \tag{14}$$

Because $f_V(\cdot)$ is a density function, the density restriction $\int f_V(v)dv = 1$ imposes a restriction on these sieve coefficients, $\sqrt{2\pi}\sum_{i=0}^{J} i!\beta_i^2 = 1$. By substituting the parametric specification of $f_{Y|X,C}(y|x,c)$ and the Hermite polynomial series of f_V into Eq. (12) we obtain:

$$f_{Y|X,\overline{W}}(y|x,\overline{w};\theta,\lambda,\beta) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)\Big(\sum_{i=0}^{J} \beta_i h_i(c - \overline{w}\lambda)\Big)^2 dc. \tag{15}$$

Let $\{Y_i, X_i, \overline{W}_i\}_{1\le i\le N}$ be a sample of observed variables. The empirical analogue of the expression in Eq. (13) is given by

$$\widehat{Q}_N(\alpha,\beta) = \frac{1}{N}\sum_{i=1}^{N} \log f_{Y|X,\overline{W}}(y_i|x_i,\overline{w}_i;\alpha,\beta). \tag{16}$$

The sieve ML estimators of α_0 and f_V from this maximizing problem are

$$(\widehat{\alpha},\widehat{\beta}) \equiv \arg\max_{\alpha,\beta} \widehat{Q}_N(\alpha,\beta). \tag{17}$$

The estimation is a standard sieve ML procedure and there is a huge literature on the estimation such as Shen (1997), Chen and Shen (1998), and Ai and Chen (2003). Ai and Chen (2003) have shown the consistency and asymptotic normality of the parametric component α. We refer the literature for a review of the sieve ML estimation.

4. Discussions

We propose a Hausman-type test for the distribution assumption on V, extend the identification results to dynamic nonlinear panel data models and provide identification result on partial effects which are often the objects of interest in empirical studies

4.1. Specification Test for Normality Assumption

A popular approach in the literature is to impose that $V \sim N(0,\sigma_0^2)$ for some $\sigma_0^2 > 0$ and we provide a specification test for such assumption.[8] Recall that in our case

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha)dc,$$

where $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is constructed from data. However, if Assumption 2.2 holds with $V \sim N(0,\sigma^2)$, then we have $C|\overline{W} \sim N(\overline{W}\lambda,\sigma^2)$ and $f_{C|\overline{W}}(c|\overline{w};\lambda,\sigma) = \sigma^{-1}\phi((c-\overline{w}\lambda)/\sigma)$ where $\phi(\cdot)$ denotes the density function of standard normal. Therefore, the full parametric MLE for (α_0,σ_0^2) is defined as for some $M < \infty$,

$$(\hat{\alpha}_{pa},\hat{\sigma}^2_{pa}) \equiv \arg\max_{\alpha\in\mathscr{A},\sigma^2\le M} \frac{1}{N}\sum_{i=1}^{N} f_{pa}(Y_i|X_i,\overline{W}_i;\alpha,\sigma^2), \tag{18}$$

$$f_{pa}(y|x,\overline{w};\alpha,\sigma^2) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(c-\overline{w}\lambda)^2/2\sigma^2}dc.$$

Under suitable conditions, we can show that $\sqrt{N}(\hat{\alpha}_{pa}-\alpha_0) \xrightarrow{d} N(0,\mathscr{V}_{pa})$ where $\mathscr{V}_{pa}$ the asymptotic variance and covariance matrix of the fully parametric estimator.

We can construct a Hausman-type test for the null hypothesis that $V \sim N(0,\sigma_0^2)$ by testing if $\hat{\alpha}$ and $\hat{\alpha}_{pa}$ are close to each other. To be specific, both our estimator and the fully parametric

[8] Chamberlain (1980) used the specification in the static probit model. Wooldridge (2005) also used the specification in dynamic panel data models where the conditional mean of C is a linear combination of time-invariant variables and the initial condition.

estimator for α_0 are consistent for α_0 when $V \sim N(0, \sigma_0^2)$. However, if $V \nsim N(0, \sigma^2)$ for any $\sigma^2 \leq M$, then our estimator is still consistent, but in general, the fully parametric estimator will converge to a point other than α_0. Therefore, under the null hypothesis, we can show that $\sqrt{N}(\hat{\alpha} - \hat{\alpha}_{pa}) \xrightarrow{d} N(0, \mathcal{V}_{di})$ where $\mathcal{V}_{di}$ stands for the asymptotic covariance of $\sqrt{N}(\hat{\alpha} - \hat{\alpha}_{pa})$ under null. Let $\widehat{\mathcal{V}}_{di}^{b}$ denote the bootstrapped estimator for $\mathcal{V}_{di}$. Define the test statistics as

$$\widehat{S}_N = N(\hat{\alpha} - \hat{\alpha}_{pa})(\widehat{\mathcal{V}}_{di}^{b})^{-1}(\hat{\alpha} - \hat{\alpha}_{pa})' \tag{19}$$

and the null distribution of $\widehat{S}_N$ would be a Chi-squared distribution with degrees of freedom equal to the dimension of α_0.

4.2. Extension to Dynamic Nonlinear Panel Data Models

Consider a parametric dynamic panel data density function:

$$f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c; \theta), \text{ for all } t = 2, ..., T. \tag{20}$$

Time series dependence arises naturally in the context of dynamic panel data models that can be used to investigate the effects of lagged outcomes on current outcomes. Thus, the identification and estimation of models in Eq. (20) are of great practical value. We impose the following assumptions for identification in this setting.

Assumption 4.1. *(Movement of the Unobserved Effects)*

Assume that (i) $f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t, y_{t-1}, \overline{w}, c) = f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c)$ *for all* $(y_t, x_t, y_{t-1}, \overline{w}, c) \in \mathcal{Y}_t \times \mathcal{X}_t \times \mathcal{Y}_{t-1} \times \overline{\mathcal{W}} \times \mathcal{C}$;

(ii) $f_{C|X,Y_1,\overline{W}}(c|x, y_1, \overline{w}) = f_{C|\overline{W}}(c|\overline{w})$ *for all* $(c, x, y_1, \overline{w}) \in \mathcal{C} \times \mathcal{X} \times \mathcal{Y}_1 \times \overline{\mathcal{W}}$.

Similar to Eq. (A.1), we can use Assumptions 2.2 and 4.1(i) &(ii) to obtain

$$
\begin{aligned}
&f_{Y_2,Y_3,\ldots,Y_T|X,Y_1,\overline{W}}(y_2,y_3,\ldots,y_T|x,y_1,\overline{w}) \\
&= \int_{\mathscr{C}_T} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t,y_{t-1},\overline{w},c) f_{C|X_t,Y_{t-1},\overline{W}}(c|x,y_1,\overline{w})dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c) f_{C|\overline{W}}(c|\overline{w})dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c;\theta_0) f_v(c-\overline{w}\lambda_0)dc. \qquad (21)
\end{aligned}
$$

That is, the observable density function $f_{Y_2,Y_3,\ldots,Y_T|X,Y_1,\overline{W}}$ can be written as the convolution of the dynamic panel data density function $\prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}$ and the distribution of the CRE remainder error f_V. Therefore, sufficient conditions for identification are similar to those in Theorem 2.1 and the identification results follow.

Furthermore, Assumption 4.1(ii) can be generalized to include the initial condition of the outcome Y_1 as a time-invariant covariate in the CRE specification. For example, consider

$$C = \gamma_0 Y_1 + \overline{W}\lambda_0 + V, \qquad (22)$$

where the error term V is independent of Y_1, and $\overline{W}$. Then, $f_{C|X,Y_1,\overline{W}}(c|x,y_1,\overline{w}) = f_V(C-\gamma_0 Y_1 - \overline{W}\lambda_0)$. Note that our approach still applies in this case. Wooldridge (2005) considered the same specification, but assumed V is normally distributed, so the Hausman test proposed in Section 4.1 can be applied here too.

4.3. Identification of Partial Effects

In most empirical applications, researchers are also interested in partial effects that are defined as the marginal effects of an explanatory variable on the conditional expectation of the dependent variable holding other explanatory variables fixed. For a given value of the explanatory variables (X_t, C), the partial effect of continuous X_{tk} on Y_t is the partial derivative of $\mathrm{E}[Y_t|X_t,C]$ with respect to X_{tk}:

$$\frac{\partial \mathrm{E}[Y_t|X_t,C]}{\partial X_{tk}}. \qquad (23)$$

If X_{tk} is a discrete variable, partial effects are computed by comparing $\mathrm{E}[Y_t|X_t, C]$ at different values of X_{tk}, holding other variables fixed. However, the partial effects of interest depend on the unobserved heterogeneity C and it is not clear which value of C one should consider, so the common practice is to average the partial effect across the population distribution of C which leads to the average partial effect (APE) in the literature. Note that the marginal distribution of the unobserved heterogeneity C can also be identified by the results in Theorem 2.1:

$$f_C(c) = \int_{\overline{\mathscr{W}}} f_{C|\overline{W}}(c|\overline{w};\alpha_0) f_{\overline{W}}(\overline{w}) d\overline{w} = \mathrm{E}_{\overline{W}}\left[f_{C|\overline{W}}(c|\overline{w};\alpha_0) \right].$$

Then the APE is identified by:

(24)
$$\mathrm{APE}(x_{tk}) = \int_{\mathscr{C}} \left(\frac{\partial \mathrm{E}[Y_t|X_t = x_t, C = c]}{\partial X_{tk}} \right) f_C(c) dc = \int_{\mathscr{C}} \left[\int_{\mathscr{Y}_t} y_t \frac{\partial f_{Y_t|X_t,C}(y_t|x_t, c;\theta_0)}{\partial x_{tk}} dy_t \right] f_C(c) dc.$$

Corollary 4.1. *Under Assumptions 2.1- 2.6, the APE defined in Eq. (24) is identified from the joint distribution of a panel data sample,* $\{Y_t, X_t\}$ *for* $t = 1, 2, ..., T$.

For the binary-choice model in Eq. (5) with a continuous explanatory variable, the APE is given by

(25)
$$\mathrm{APE}(x_{tk}) = \theta_k \int_{\mathscr{C}} \frac{\partial F_{\varepsilon_t}(-x_t\theta - c)}{\partial x_{tk}} f_C(c) dc,$$

where F_{ε_t} is the CDF of the error term in the latent variable model. If x_{td_θ} is a binary explanatory variable, then the partial effect from changing x_{td_θ} from zero to one, holding all other variables fixed, is

(26) $\mathrm{APE}(x_{td_\theta})$
$$= \int_{\mathscr{C}} (F_{\varepsilon_t}\left(-x_{t1}\theta_1 - ... - x_{td_\theta - 1}\theta_{d_\theta - 1} - \theta_{1d_\theta} - c\right) - F_{\varepsilon_t}\left(-x_{t1}\theta_1 - ... - x_{td_\theta - 1}\theta_{d_\theta - 1} - c\right)) f_C(c) dc.$$

For a parametric dynamic panel data model in Eq. (20), researches may be interested in whether there is state dependence–that is, the partial effect of Y_{t-1} on Y_t after controlling for the unobserved heterogeneity, C. The magnitude of state dependence can be defined as an average partial effect from $Y_{t-1} = 0$ to $Y_{t-1} = 1$ at fixed values of all other variables. We now

state our identification result for APE.

5. Monte Carlo Simulation

In this section, we present simulation results to illustrate the finite sample performance of the proposed sieve ML estimation procedure of a panel data probit model in Section 3. We assess the finite sample performance of the proposed estimator in a variety of models including static models and dynamic models.

5.1. Static Model

The data generating process (DGP) is defined as follows:

$$
\begin{aligned}
&Y_t = \mathbf{1}(\theta X_t + C + \varepsilon_t \geq 0), \text{ for } t = 1,2, \\
&C = \lambda \overline{W} + V, \ \overline{W} = \frac{1}{2}\sum_{t=1}^{2} X_t, \\
&X_2 = 0.8X_1 + \xi, X_1 \sim U(0,2), \xi \sim N(0,1), \\
&(\varepsilon_1, \varepsilon_2) \sim N(0, I_2),
\end{aligned}
$$

where I_2 is the 2×2 identify matrix and we set $(\theta, \lambda) = (0.5, 0.5)$. We consider various distributions of V. For a random variable Q, we denote the corresponding truncated random variable over interval $[a,b]$ as $Trun(Q,[a,b])$.[9] Let μ_ω be the mean of ω. Two specifications of V are considered:

$$
\begin{aligned}
&\text{DGP I: } V \sim Trun(N(0,1),[-2,2]), \\
&\text{DGP II: } V = \omega - \mu_\omega \text{ with } \omega \sim Trun\big(H,[0,2]\big) \text{ and } \ln H = N(0,5), \\
&\text{DGP III: } V = \omega - \mu_\omega \text{ with } \sqrt{\omega} \sim Trun(Rayleigh(1),[-2,15]).
\end{aligned}
$$

The unobserved heterogeneities in all the simulation designs are with bounded supports so Assumption 2.4(ii) is satisfied in all cases. We consider sample sizes 500, and 1,000 and for each case, we consider 150 simulation replications. For comparison, we also consider the other

[9] $Trun(Q,[a,b])$ is a random variable generated by $F_Q^{-1}(u \cdot (F_Q(b) - F_Q(a)) + F_Q(a))$ where F_Q is the CDF of Q random variable, F_Q^{-1} is the inverse of F_Q and u is a uniform random variable on $[0,1]$.

two estimators. The first one is an infeasible estimator that treats V as known. The second one is the conventional random effects estimator which specifies the unobserved heterogeneity to be normally distributed. The simulation results for parameters and APE are presented in Tables 1–2 and 3–4, respectively.

The estimation results of the parameters in DGP I show a little bias in all the three estimators. In this case, the normal specification in the conventional random effects estimator is close to the true distribution of the data so the estimation does not suffer from the misspecification of the estimator. The proposed sieve ML estimator exhibits little degrees of biases in DGP II but the conventional random effects estimator exhibits conspicuous downward bias in θ, λ, and σ for all sample sizes.

Overall, the simulation results show that the proposed sieve ML estimator works well in simulation designs. As expected, the infeasible estimator outperforms the proposed estimator in RMSE. The conventional estimator does a good job in estimating θ and λ in DGP I but causes bias in DGPs II. The estimation results for APEs in Tables 3–4 have a similar pattern. While the infeasible estimator and the proposed sieve ML estimator perform well in all simulations, the conventional estimator perform well only in DGP I.

We also consider the Hausman type test proposed in Section 4.1 for the normality assumption of V and the results are summarized in Table 5. For DGP I, the rejection rates are 0.047 and 0.053 which are close to the nominal size 5% given that the normality assumption hold.[10] For DGPs II and III, the rejection rates are much higher than the nominal size 5% and increase with sample size indicating that our test are consistent when the normality assumption is violated.

5.2. Dynamic Model

The simulation design for dynamic models is close to the static models in Section 5.1. The DGP for dynamic models is defined as follows:

$$Y_t = \mathbf{1}\left(\gamma Y_{t-1} + \theta X_t + C + \varepsilon_t \geq 0\right), \text{ for } t = 1, ..., 7,,$$

[10]Because it takes more time to implement the bootstrap estimator for the covariance matrix, instead we use empirical covariance matrix which is the average of the 150 simulated estimators when we conduct the Hausman test in the simulations.

where $(\gamma, \theta, \lambda) = (0.8, 0.5, 0.5)$ and DGPs for X_t and C are the same as the ones in the static models. Tables 6–7 and 8–9 present the estimation results for parameters and the magnitudes of state dependence. We reach the same conclusion as the estimation results of the static models. While the proposed sieve ML estimator performs well in all DGPs, the conventional random effects estimator cannot deliver a consistent estimation for the parameters γ, λ and σ in DGP II. The simulation results for the Hausman type test are similar to the static case too.

6. Empirical Application

We apply the proposed sieve ML estimator to estimate the persistence effects of union membership using the panel data in Wooldridge (2005). There are 7 periods in which the first period corresponds to year 1981 and the last period corresponds to year 1987. The dynamic behavior of union membership may come from its formulation of behavior maximizing the future utility and mechanical dynamics for membership. We model the membership decisions as the following dynamic probit model:

$$\begin{aligned}(27)\qquad &\text{Prob}(\text{Union}_t = 1 | \text{Married}_t, \text{Union}_{t-1}, D_{1982}, ..., D_{1987}, C) \\ &= \Phi\left(\beta_0 + \beta_1 \text{Married}_t + \rho \text{Union}_{t-1} + \gamma_1 D_{1982} + ... + \gamma_6 D_{1987} + C\right), \text{ for all } t = 1, ..., 7.\end{aligned}$$

where $D_{1982}, ..., D_{1987}$ are year dummies for 1982-1987 respectively. The correlated random effect specification for the individual unobserved heterogeneity C follows Eq. (22), where $Y_1 = \text{Union}_0$, and $\overline{W}$ is the 1×2 vector of the variables, Education and Black.

As the identification of the parametric nonlinear panel data model hinges on assumptions in Section 2.1, it is necessary to discuss the plausibility of the identification conditions in this application. Assumption 2.1 is to specify the distribution of the disturbance in the latent variable formulation for the binary membership decisions to be normally distributed and this results in the probit model in Eq. (27). Because the disturbance is the sum of many unobserved factors after controlling the covariates, we can invoke the central limit theorem to conclude that the disturbance has an approximate normal distribution. If C represents individual union preference, Assumption 2.2 is to replace C with its linear projection onto the time-invariant observed variables, the initial union condition, and the marital status indicators, and the projection er-

ror V. Assumption 2.3 states that (a) the union decision is independent of the initial union condition, and the marital status indicators conditional on the current marital status and the previous union decision; (b) given the initial union condition, and the marital status indicators, individual union preference is independent of the time-varying covariates such as the current marital status and the previous union decision. Assumptions 2.4 and 2.5 require that the conditional densities $f_{Y_t|X_t,Y_{t-1},\overline{W}}(y_t|x_t,y_{t-1},\overline{w})$ and $f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c;\theta)$ satisfy regular conditions, which are reasonable for this application. Assumption 2.6 is an identification condition for the semiparametric probit model and it makes sure that we have sufficient information to distinguish between alternative population structures.

The estimated coefficients are shown in Table 10. To facilitate comparisons with the conventional approach, we also consider the fully parametric approach which assumes the normality of V. The estimated coefficients in the two methods exhibit the same sign but with different magnitudes. The coefficients for the marital status are 19.7% and 29.0% in the conventional and the sieve ML methods, respectively. The coefficient on the previous union decision is smaller than the coefficient on the initial union decision. According to Heckman (1978, 1981a,b), the estimation results show that there exists small "true" state dependence and large “spurious" state dependence.

To obtain the magnitude of the state dependence, we adopt the following definition of estimated probabilities of being union in 1987,

$$\int_{\mathscr{C}} \text{Prob}(\text{Union}_t = 1|\text{Married}_t = a, \text{Union}_{t-1} = b, D_{1982} = 0, ..., D_{1987} = 1, C) f_C(c) dc. \tag{28}$$

where $a, b \in \{0, 1\}$. A similar definition can be applied to Union_{t-1}. The estimation results for the APEs are in Table 11. While the estimated state dependence in $\text{Married}_t = 1$ are 0.286=0.485-0.199 and 0.232=0.580-0.348, the estimated APEs for $\text{Married}_t = 0$ are 0.269=0.444-0.175 and 0.218=0.494-0.276 for conventional and sieve ML estimators. This indicates that union membership has a strong state dependence. The conventional approach gives much higher estimate than the proposed sieve ML approach around 5% in these two conditions. Finally, we use the Hausman-type test in Section 4.1 to test the null hypothesis that $V \sim N(0, \sigma_0^2)$. We fail to reject the null hypothesis at 10% significance level because the test statistics is $\widehat{S}_N = 18.390$ and the critical value based on a $\chi^2(12)$ distribution is 18.549.

7. Conclusion

This paper addresses unsolved issues of the distribution misspecification of the random effect approach for nonlinear panel data models with a fixed time dimension. The main insight of our approach is to use the information of the time-invariant observed covariates as a source of identification for a time-invariant heterogeneity structure in a Mundlak-type specification. Then, the average likelihood takes the form of the convolution of the parametric panel data model and the conditional distribution of the unobserved heterogeneity. By Fourier transformations, we provide a data-driven specification of conditional distributions of the unobserved heterogeneity which is internally consistent with the parametric nonlinear panel data models. That is, the average likelihood is correctly specified. A sieve ML estimator is provided based on the identification result. Under appropriate regularity conditions, the estimator is root N consistent and asymptotically normal. The identification strategy can also be applied to dynamic nonlinear panel data models. In the application to investigate the persistence effects of union membership, we show that union membership has a strong state dependence and the magnitudes are up to 20%.

Appendix

A. Proof of Theorem 2.1

Consider

$$
\begin{aligned}
f_{Y|X,\overline{W}}(y|x,\overline{w}) &= \int_{\mathscr{C}} f_{Y|X,\overline{W},C}(y|x,\overline{w},c) f_{C|X,\overline{W}}(c|x,\overline{w})dc \\
&= \int_{\mathscr{C}} f_{Y|X,\overline{W},C}(y|x,\overline{w},c) f_{C|\overline{W}}(c|\overline{w})dc \\
&= \int_{\mathscr{C}} \left(\prod_{t=1}^{T} f_{Y_t|X_t,\overline{W},C}(y_t|x_t,\overline{w},c) \right) f_{C|\overline{W}}(c|\overline{w})dc \\
&= \int_{\mathscr{C}} \left(\prod_{t=1}^{T} f_{Y_t|X_t,C}(y_t|x_t,c) \right) f_{C|\overline{W}}(c|\overline{w})dc \\
&= \int_{\mathscr{C}} f_{Y|X,C}(y|x,c) f_V(c-\overline{w}\lambda_0)dc, \qquad \text{(A.1)}
\end{aligned}
$$

where we have used (a) the law of the total probability, (b) Assumptions 2.2 and 2.3(i)(ii)&(iii), and (c) $f_{Y|X,C}(y|x,c) = \prod_{t=1}^{T} f_{Y_t|X_t,C}(y_t|x_t,c)$.

Given each (y,x), constructing a characteristic function of $f_{Y|X,\overline{W}}(y|x,\overline{w})$ with respect to $\overline{w}_1$ and interchanging integrations yields the following equation: for all real-valued ξ,

$$
\begin{aligned}
&\int_{\overline{\mathscr{W}}_1} e^{i\xi\overline{w}_1} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1 \\
&= \int_{\overline{\mathscr{W}}_1} e^{i\xi\overline{w}_1} \left(\int_{\mathscr{C}} f_{Y|X,C}(y|x,c) f_V(c-\overline{w}\lambda_0)dc \right) d\overline{w}_1 \\
&= \int_{\mathscr{C}} \left(\int_{\overline{\mathscr{W}}_1} e^{i\xi\overline{w}_1} f_V(c-\overline{w}\lambda_0)d\overline{w}_1 \right) f_{Y|X,C}(y|x,c)dc \\
&= \int_{\mathscr{C}} \left(\int_{\mathscr{C}} e^{i\xi \frac{c-v-\lambda_{02}\overline{w}_2-\ldots-\lambda_{0K_1}\overline{w}_{K_1}}{\lambda_{01}}} f_V(v)\frac{dv}{-\lambda_{01}} \right) f_{Y|X,C}(y|x,c)dc \\
&= \frac{-1}{\lambda_{01}} e^{i\xi \frac{-\lambda_{02}\overline{w}_2-\ldots-\lambda_{0K_1}\overline{w}_{K_1}}{\lambda_{01}}} \left(\int_{\mathscr{C}} e^{i\xi\frac{-v}{\lambda_{01}}} f_V(v)dv \right) \int_{\mathscr{C}} e^{i\xi\frac{c}{\lambda_{01}}} f_{Y|X,C}(y|x,c)dc \\
&= \frac{-1}{\lambda_{01}} e^{-i\xi \frac{\sum_{k=2}^{K_1}\lambda_{0k}\overline{w}_k}{\lambda_{01}}} \phi_v\left(\frac{-\xi}{\lambda_{01}}\right) \int_{\mathscr{C}} e^{i\xi\frac{c}{\lambda_{01}}} f_{Y|X,C}(y|x,c)dc, \qquad \text{(A.2)}
\end{aligned}
$$

where $\phi_v(\xi) = \int_v e^{i\xi v} f_V(v)dv$. Rescale ξ by $-\lambda_{01}\xi$ in Eq. (A.2) and the equation becomes

$$
-\lambda_{01} \int_{\overline{\mathscr{W}}_1} e^{-i\xi\lambda_{01}\overline{w}_1} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1 = \phi_v(\xi) e^{i\xi\sum_{k=2}^{K_1}\lambda_{0k}\overline{w}_k} \int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc. \qquad \text{(A.3)}
$$

Multiplying each side of Eq. (A.3) by $e^{-i\xi\sum_{k=2}^{K_1}\lambda_{0k}\overline{w}_k}$ establishes that

$$-\lambda_{01}\int_{\overline{\mathscr{W}}_1} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1 = \phi_v(\xi)\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc. \tag{A.4}$$

Because $\lambda_{0j}\neq 0$, for $j=1,...,K_1$ by Assumption 2.2(i), following the derivation of Eq. (A.4) we can obtain: for $j=1,...,K_1$,

$$-\lambda_{0j}\int_{\overline{\mathscr{W}}_j} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_j = \phi_v(\xi)\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc. \tag{A.5}$$

The intuition of the above expression is that the Fourier transform of the convolution of two functions is the product of their individual Fourier transforms and there is a convolution type function in Eq. (A.1).[11] For example, if we consider the simplest case, $\lambda_0=(\lambda_{01},...,\lambda_{0K_1})^T=(-1,0,...,0)'$, Eq. (A.4) becomes

$$\underbrace{\int_{\overline{\mathscr{W}}_1} e^{i\xi\overline{w}_1} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1}_{\text{Fourier transform of } f_{Y|X,\overline{W}}} = \underbrace{\phi_v(\xi)}_{\substack{\text{Fourier}\\ \text{transform}\\ \text{of } f_V}} \times \underbrace{\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c)dc}_{\text{Fourier transform of } f_{Y|X,C}}.$$

Averaging Eq. (A.5) across $j=1,...,K_1$ yields

$$\frac{-1}{K_1}\sum_{j=1}^{K_1}\left(\lambda_{0j}\int_{\overline{\mathscr{W}}_j} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_j\right) = \phi_v(\xi)\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc. \tag{A.6}$$

The result in Eq. (A.6) holds for every $(y,x,\overline{w})\in\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}$. As such, we can utilize a positive weighting function $\Omega(y,x,\overline{w})$. Multiplying the equation by $\Omega(y,x,\overline{w})$, integrating out the variables $(y,x,\overline{w})$ over the domain, and then interchanging the integrations, we obtain

$$\begin{aligned}&\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}}\frac{-1}{K_1}\sum_{j=1}^{N}\left(\lambda_{0j}\int_{\overline{\mathscr{W}}_j} e^{-i\xi\sum_{k=1}^{K}\lambda_{0k}\overline{w}_k}\underbrace{f_{Y|X,\overline{W}}(y|x,\overline{w})}_{\substack{\text{observable}\\ \text{from data}}}d\overline{w}_j\right)\Omega(y,x,\overline{w})dydxd\overline{w}\\ &=\phi_v(\xi)\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c}\underbrace{f_{Y|X,C}(y|x,c;\theta_0)}_{\substack{\text{population}\\ \text{density}}}\Omega(y,x)dydxdc,\end{aligned} \tag{A.7}$$

where $\int_{\overline{\mathscr{W}}}\Omega(y,x,\overline{w})d\overline{w}=\Omega(y,x)$.

Assumption 2.4(i) &(ii) implies that the characteristic functions other than $\phi_v(\xi)$ in Eq. (A.7) are well defined. Denote $h(\xi,y,x,\overline{w};\lambda_0)=\frac{-1}{K_1}\sum_{j=1}^{K_1}\left(\lambda_{0j}\int_{\overline{\mathscr{W}}_j} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_j\right)$. By Assumption 2.4(iii), dividing both sides of Eq. (A.7) by $\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)\Omega(y,x)dydxdc$ yields the char-

[11] In probabilistic language, a Fourier transform is simply the characteristic function.

acteristic function of the remainder term V in the CRE assumption,

$$\phi_v(\xi) = \frac{\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} h(\xi,y,x,\overline{w};\lambda_0)\Omega(y,x,\overline{w})dydxd\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)\Omega(y,x)dydxdc}. \tag{A.8}$$

Let $\alpha = (\theta,\lambda)$. With the correctly specified parametric family $\{f_{Y_t|X_t,C}(y_t|x_t,c;\theta)|\theta\in\Theta\}$ and CRE condition in Assumption 2.2, we can use Assumption 2.4 to extend Eq. (A.8) to all $\alpha\in\Theta\times\Lambda$ to obtain a potential semi-parametric family of functions connected to the characteristic functions of the distribution of the remainder term v in the following form

$$\phi_{v;\alpha}(\xi) \equiv \frac{\int_{\mathscr{Y}_t\times\mathscr{X}_t\times\overline{\mathscr{W}}} h(\xi,y_t,x_t,\overline{w};\lambda)\Omega(y_t,x_t,\overline{w})dy_tdx_td\overline{w}}{\int_{\mathscr{Y}_t\times\mathscr{X}_t\times\mathscr{C}} e^{-i\xi c} f_{Y_t|X_t,C}(y_t|x_t,c;\theta)\Omega(y_t,x_t)dy_tdx_tdc}, \tag{A.9}$$

where $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$. Notice that the terms in the numerator and denominator of the fraction in Eq. (A.8) are known. While the term in the numerator can be estimated directly from data, the term in the denominator can be constructed by the parametric panel data density function $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$.

Applying the inverse Fourier transform to $\phi_{v;\alpha}(t)$ yields a semi-parametric family of the function of V as

$$f_{V;\alpha}(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi v}\phi_{v;\alpha}(\xi)d\xi. \tag{A.10}$$

Under Assumption 2.5(ii), the characteristic function $\phi_v(\cdot)$ belongs to $L^1(\mathbb{R})$, we can apply the Fourier inversion theorem to obtain $f_{v;\alpha_0}(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi v}\phi_{v;\alpha_0}(\xi)d\xi = f_V(v)$. This suggests that the PDF of the remainder term V, $f_V(v)$, needs to be expressed in terms of $\phi_v(\cdot)$ by means of Fourier inversion formula. In order to show which function space satisfies Fourier inversion formula, we introduce the Schwartz class $\mathscr{S}(\mathbb{R}^n)$ as follows. Given a $n\times 1$ vector of nonnegative integers, $a = (a_1,...,a_n)'$, denote $[a] = a_1+...+a_n$, and let D^a denote the differential operator defined by $D^a = \frac{\partial^{[a]}}{\partial x_1^{a_1}...\partial x_d^{a_d}}$. The space $\mathscr{S}(\mathbb{R}^n)$ is a collection of smooth functions $g(x)$ such that for all multi-indices a,b,

$$\sup_{x\in\mathbb{R}^n} |x^a D^b g(x)| = c_{a,b}(g) < \infty, \tag{A.11}$$

where $x = (x_1,...,x_n)'$ and $x^a = x_1^{a_1}...x_n^{a_n}$. $\mathscr{S}(\mathbb{R}^n)$ contains those smooth functions with compact support, and functions with infinite supports like $e^{-|x|^2}$. The following result comes from Proposition 1.1 of Chapter X in Torchinsky (2012):

Proposition A.1. *(Fourier Inversion Formula) Suppose $f\in\mathscr{S}(\mathbb{R}^n)$ and $\widehat{f}$ is its Fourier transform. Then*

$$f(x) = (2\pi)^{-n}\int_{\mathbb{R}^n} e^{ix\cdot\xi}\widehat{f}(\xi)d\xi. \tag{A.12}$$

Next, we will try to connect this semi-parametric family of unobservable V to a parametric family of density functions of observable variables and then use sample observations of the observable variables to pin down the population parameter (θ_0, λ_0). Integrating out $f_{V;\alpha}(c - \overline{w}\lambda)$ over the domain $\mathscr{C}$ yields

$$c_\alpha(\overline{w}) \equiv \int_{\mathscr{C}} f_{v;\alpha}(c - \overline{w}\lambda)dc = \int_{-\infty}^{\infty} \left(e^{i\xi\overline{w}\lambda}\phi_{v;\alpha}(\xi)\right)\left(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc\right)d\xi. \tag{A.13}$$

Note that when $\mathscr{C} = \mathbb{R}$, the last term of the integrand becomes $\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi c}dc = \delta(\xi)$ where $\delta(\xi)$ is the Dirac delta function and it is the Fourier transform of a constant function. The important property of the delta function is that $\int f(\xi)\delta(\xi)dt = f(0)$ for all continuous compactly supported functions $f(\cdot)$. With this property, if $\phi_{v;\alpha}(\xi)$ is continuous compactly supported then Eq. (A.13) can be further reduced as

$$c_\alpha(\overline{w}) = \int_{-\infty}^{\infty} \underbrace{\left(e^{i\xi\overline{w}\lambda}\phi_{v;\alpha}(\xi)\right)}_{\text{a function of } \xi} \underbrace{\left(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc\right)}_{\psi_c(\xi)} d\xi = \phi_{v;\alpha}(0). \tag{A.14}$$

Use $f_{v;\alpha}$ to construct the following semi-parametric family of conditional density functions of the unobserved heterogeneity

$$f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{c_\alpha(\overline{w})} f_{v;\alpha}(c - \overline{w}\lambda) \tag{A.15}$$

such that $f_{C|\overline{W}}(c|\overline{w};\alpha_0) = \frac{1}{c_{\alpha_0}(\overline{w})} f_v(c - \overline{w}\lambda_0)$.

Lemma A.1. *Under Assumptions 2.2(i)&(ii), 2.3(i)(ii)&(iii), 2.4, and 2.5(i)&(ii), there exists an open neighborhood of α_0 such that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is a conditional density function for α in the neighborhood.*

Proof: First, by Assumptions 2.2(i)&(ii), 2.3(i)(ii)&(iii), 2.4, and 2.5(ii), $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is well defined. Then, Assumption 2.5(i) & (ii) implies that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is continuous for all α and is nonnegative for α in some open neighborhood of α_0. With the definition of $c_\alpha(\overline{w})$ in Eq. (A.13), we can obtain the integration of $f_{C|\overline{W}}(c|\overline{w};\alpha)$ over the domain $\mathscr{C}$ is equal to one. *Q.E.D.*

Combining this semi-parametric PDF with the parametric known density functions $f_{Y|X,C}(y|x,c;\theta)$ lead to the following semi-parametric function of observable variables:

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha)dc. \tag{A.16}$$

Integrating out $f(y|x,\overline{w};\alpha)$ over the domain $\mathscr{Y}$ and interchanging the integrations yields

$$\begin{aligned}\int_{\mathscr{Y}} f(y|x,\overline{w};\alpha)dy_t &= \int_{\mathscr{C}}\left(\int_{\mathscr{Y}_t} f_{Y|X,C}(y|x,c;\theta)dy_t\right) f_{C|\overline{W}}(c|\overline{w};\alpha)dc \\ &= \int_{\mathscr{C}} f_{C|\overline{W}}(c|\overline{w};\alpha)dc = 1. \end{aligned} \tag{A.17}$$

Under general framework of conditional maximum likelihood estimation, we have the following result.

Lemma A.2. *If α_0 is identified and $E\left[\log f(Y|X,\overline{W};\alpha)\middle|X=x,\overline{W}=\overline{w}\right] < \infty$ for all α and for all $(x,\overline{w}) = \mathscr{X}\times\overline{\mathscr{W}}$, then $K(\alpha;x,\overline{w})$ has a unique maximum at α_0 for all $(x,\overline{w}) = \mathscr{X}\times\overline{\mathscr{W}}$.*

Proof: The proof uses the Jensen's inequality. For $\alpha \neq \alpha_0$,

$$\begin{aligned} K(\alpha;x,\overline{w}) &= \mathrm{E}\Big[\log\left(\frac{f(Y|X,\overline{W};\alpha)}{f(Y|X,\overline{W};\alpha_0)}\right)\Big|X=x,\overline{W}=\overline{w}\Big] \\ &< \log\mathrm{E}\Big[\frac{f(Y|X,\overline{W};\alpha)}{f(Y|X,\overline{W};\alpha_0)}\Big|X=x,\overline{W}=\overline{w}\Big] \\ &= \log\left(\int_{\mathscr{Y}} f(Y|X,\overline{W};\alpha)dy\right) \\ &= \log 1 = 0 \end{aligned}$$

where we have used the strict concavity of $\log(\cdot)$. *Q.E.D.*

Differentiating Eq. (A.17) with respect to α_j and evaluating at α_0 yields

$$0 = \int_{\mathscr{Y}} \frac{\partial}{\partial\alpha_j} f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0} dy_t = \mathrm{E}\Big[\frac{\frac{\partial}{\partial\alpha_j} f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0}}{f(y_i|x_i,\overline{w}_i;\alpha_0)}\Big|X=x,\overline{W}=\overline{w}\Big] = \frac{\partial K(\alpha;x,\overline{w})}{\partial\alpha_j}\big|_{\alpha=\alpha_0}$$

Applying the above result to Eq. (10), we have $\frac{\partial}{\partial\alpha}K(\alpha;x,\overline{w})\big|_{\alpha=\alpha_0} = 0$. Similarly, differentiating Eq. (A.17) twice and applying the result to the second derivative of $K(\alpha;x,\overline{w})$, the matrix of the second derivative can be written as minus outer product of the gradient of the log likelihood:

$$K''(\alpha_0;x,\overline{w}) = -\mathrm{E}\Big[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0}\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)'\big|_{\alpha=\alpha_0}\Big|X=x,\overline{W}=\overline{w}\Big]. \tag{A.18}$$

Proof of Theorem 2.1: First we have discussed that the semi-parametric density function of observable variables in Eq. (A.17) is well defined using Assumptions 2.2(i)&(ii), 2.3(i)(ii)&(iii), 2.4, and 2.5(i)&(ii). We next proceed to prove the result using concavity of conditional Kullback-Leibler information criterion, i.e., the second derivative of $K(\alpha;x,\overline{w})$ in Eq. (11) is negative definite. *Q.E.D.*

A.1. Summary of Identification Steps

In this subsection, we present heuristic sketch of how to utilizes CRE specification and the two properties of the Fourier transform: (i) the Fourier transform of the convolution of the two functions is the product of their individual Fourier transforms, and (ii) the Fourier inversion formula to construct an internal consistent likelihood function.

There are four steps toward the construction of the internally consistent average likelihood function and we start with the parametric density function, $f_{Y_t|X_t,C;\theta}$.

Step 1: A convolution type function.

Under Assumptions 2.2, and 2.3, we use the law of total probability to obtain Eq. (A.1). This equation takes the form of a convolution type function:

$$f * g(w) = \int f(w-c)g(c)dc.$$

Step 2: Apply the Fourier transform.

Under Assumption 2.4, we apply the Fourier transform to the convolution type function in the first step to have the product of the Fourier transforms in Eq. (A.7) and then extend to relationship to obtain the semi-parametric function in Eq. (A.9).

Step 3: Apply the inverse Fourier transform.

Under Assumption 2.5, the inverse Fourier transform is applicable and we can recover the semi-parametric distribution of the unobserved heterogeneity in Eq. (A.15) using the inverse transform and normalization.

Step 4: Construct an internally consistent average likelihood.

We can then combine the semi-parametric distribution of the unobserved heterogeneity in Step 3 with the parametric panel data models to obtain the internally consistent average likelihood Eq. (A.16).

References

AI, C., AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

ALTONJI, J., AND R. MATZKIN (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

ALVAREZ, J., AND M. ARELLANO (2003): "The Time Series and Cross-section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71(4), 1121–1159.

ANDERSEN, E. B. (1970): "Asymptotic Properties of Conditional Maximum-likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(3), 283–301.

ARELLANO, M., AND S. BONHOMME (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

——— (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395–424.

——— (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *Review of Economic Studies*, 79(3), 987–1020.

ARELLANO, M., AND R. CARRASCO (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics*, 115(1), 125–157.

BALTAGI, B. H. (2008): *Econometric Analysis of Panel Data*. John Wiley & Sons.

BESTER, C. A., AND C. HANSEN (2009): "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects," *Journal of Business & Economic Statistics*, 27(2), 131–148.

BROWNING, M., AND J. M. CARRO (2014): "Dynamic Binary Outcome Models with Maximal Heterogeneity," *Journal of Econometrics*, 178(2), 805–823.

CHAMBERLAIN, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47(1), 225–238.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(1), 159–168.

CHEN, X., AND X. SHEN (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66(2), 289–314.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

CHERNOZHUKOV, V., I. FERNANDEZ-VAL, S. HODERLEIN, S. HOLZMANN, AND W. NEWEY (2015): "Nonparametric Identification in Panels using Quantiles," *Journal of Econometrics*, 188(2), 378–392.

EVDOKIMOV, K. (2011): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," *Working Paper*.

GRAHAM, B., AND J. POWELL (2012): "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica*, 80(5), 2105–2152.

HECKMAN, J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence," in *Annales de l'INSEE*, pp. 227–269. Institut national de la statistique et des études économiques.

——— (1981a): "Statistical Models for Discrete Panel Data," *in Structural Analysis of Discrete Panel Data with Econometric Applications, ed. by C. Manski and D. McFadden. Cambridge: MIT Press*, pp. 179–195.

——— (1981b): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," *in Structural Analysis of Discrete Data with Econometric Applications, ed. by C. F. Manski and D. L. McFadden*, pp. 114–178.

HODERLEIN, S., AND E. MAMMEN (2007): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 75(5), 1513–1518.

HODERLEIN, S., AND H. WHITE (2012): "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects," *Journal of Econometrics*, 168(2), 300–314.

HONORÉ, B., AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

HONORÉ, B., AND E. TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

HONORÉ, B. E., AND A. LEWBEL (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70(5), 2053–2063.

HSIAO, C. (2015): *Analysis of Panel Data*. Cambridge University Press.

HU, Y., AND G. RIDDER (2010): "On Deconvolution as a First Stage Nonparametric Estimator," *Econometric Reviews*, 29(4), 365–396.

——— (2012): "Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Information," *Journal of Applied Econometrics*, 27(3), 347–385.

HU, Y., AND M. SHUM (2012): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," *Journal of Econometrics*, 171(1), 32–44.

RASCH, G. (1993): *Probabilistic Models for Some Intelligence and Attainment Tests.* ERIC.

SCHENNACH, S. (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-variables Models," *Econometrica*, 75(1), 201–239.

SCHENNACH, S. M. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72(1), 33–75.

SHEN, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics*, 25, 2555–2591.

SHIU, J., AND Y. HU (2013): "Identification and Estimation of Nonlinear Dynamic Panel Data Models with Unobserved Covariates," *Journal of Econometrics*, 175(2), 116–131.

TORCHINSKY, A. (2012): *Real-variable Methods in Harmonic Analysis*. Courier Corporation.

WOOLDRIDGE, J. (2005): "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20(1), 39–54.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*. The MIT press.

Table 1: Simulations of Static Models

	Infeasible		Conventional			Sieve ML	
Sample Size=500	θ	λ	θ	λ	σ	θ	λ
True	0.5	0.5	0.5	0.5	1	0.5	0.5
DGP I:							
Mean	0.509	0.495	0.505	0.493	0.881	0.485	0.505
Std.dev.	0.510	0.493	0.508	0.495	0.884	0.486	0.501
RMSE	0.048	0.062	0.061	0.079	0.172	0.123	0.128
DGP II:							
Mean	0.507	0.495	0.391	0.219	1.200	0.468	0.472
Std.dev.	0.510	0.497	0.389	0.223	1.193	0.467	0.481
RMSE	0.052	0.061	0.130	0.290	0.256	0.102	0.103
DGP III:							
Mean	0.505	0.493	0.404	0.239	1.278	0.469	0.469
Std.dev.	0.506	0.500	0.399	0.239	1.265	0.462	0.470
RMSE	0.051	0.057	0.120	0.271	0.325	0.102	0.108

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 2: Simulations of Static Models

	Infeasible		Conventional			Sieve ML	
Sample Size=1000	θ	λ	θ	λ	σ	θ	λ
True	0.5	0.5	0.5	0.5	1	0.5	0.5
DGP I:							
Mean	0.504	0.496	0.506	0.495	0.897	0.485	0.504
Std. dev.	0.509	0.496	0.508	0.496	0.891	0.487	0.503
RMSE	0.040	0.039	0.048	0.053	0.130	0.097	0.119
DGP II:							
Mean	0.506	0.494	0.394	0.217	1.208	0.469	0.474
Std. dev.	0.508	0.493	0.397	0.217	1.204	0.469	0.471
RMSE	0.037	0.041	0.115	0.287	0.238	0.095	0.101
DGP III:							
Mean	0.506	0.496	0.408	0.240	1.292	0.478	0.477
Std. dev.	0.509	0.497	0.410	0.240	1.279	0.490	0.473
RMSE	0.038	0.041	0.103	0.265	0.313	0.095	0.093

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 3: Simulation of the APE($\bar{x}$) in Static Models

Sample Size=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	0.117	0.118	0.117
Std. dev.	0.003	0.014	0.027
RMSE	–	0.014	0.026
DGP II:			
Mean	0.121	0.093	0.115
Std. dev.	0.004	0.016	0.022
RMSE	–	0.032	0.022
DGP III:			
Mean	0.116	0.093	0.115
Std. dev.	0.003	0.015	0.023
RMSE	–	0.028	0.023

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 4: Simulation of the APE($\bar{x}$) in Static Models

Sample Size=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	0.117	0.118	0.118
Std. dev.	0.002	0.011	0.021
RMSE	–	0.011	0.021
DGP II:			
Mean	0.121	0.094	0.116
Std. dev.	0.003	0.011	0.020
RMSE	–	0.029	0.020
DGP III:			
Mean	0.116	0.093	0.117
Std. dev.	0.003	0.011	0.020
RMSE	–	0.025	0.020

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 5: Hausman-type Test for Normality: Empirical Size

	Static Models		Dynamic Models	
	N=500	N=1000	N=500	N=1000
DGP I:	0.047	0.053	0.040	0.073
DGP II:	0.667	0.733	0.620	0.700
DGP III:	0.500	0.707	0.500	0.653

Note: The p-values of 0.05 of Chi-distributions for static models and dynamic models are 5.991 and 7.815 respectively. Empirical size refers to the fraction of rejections when using these values as the critical values.

Table 6: Simulations of Dynamic Models

	Infeasible			Conventional				Sieve ML		
Sample Size=500	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	0.8	0.5	0.5	0.8	0.5	0.5	1	0.8	0.5	0.5
DGP I:										
Mean	0.805	0.502	0.496	0.800	0.504	0.502	0.902	0.805	0.441	0.448
Std. dev.	0.806	0.504	0.493	0.803	0.505	0.499	0.898	0.781	0.452	0.441
RMSE	0.051	0.027	0.048	0.064	0.037	0.092	0.119	0.177	0.130	0.123
DGP II:										
Mean	0.792	0.503	0.507	0.648	0.501	0.230	1.355	0.764	0.460	0.473
Std. dev.	0.794	0.502	0.507	0.643	0.503	0.236	1.353	0.764	0.464	0.467
RMSE	0.055	0.024	0.045	0.167	0.038	0.284	0.377	0.107	0.093	0.094
DGP III:										
Mean	0.793	0.503	0.506	0.664	0.502	0.253	1.412	0.781	0.467	0.473
Std. dev.	0.798	0.503	0.510	0.668	0.499	0.254	1.417	0.787	0.464	0.467
RMSE	0.054	0.024	0.045	0.154	0.039	0.266	0.429	0.102	0.094	0.097

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 7: Simulations of Dynamic Models

	Infeasible			Conventional				Sieve ML		
Sample Size=1000	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	0.8	0.5	0.5	0.8	0.5	0.5	1	0.8	0.5	0.5
DGP I:										
Mean	0.805	0.502	0.496	0.800	0.504	0.502	0.902	0.837	0.420	0.463
Std. dev.	0.806	0.504	0.493	0.803	0.505	0.499	0.898	0.824	0.429	0.462
RMSE	0.051	0.027	0.048	0.064	0.037	0.092	0.119	0.155	0.132	0.117
DGP II:										
Mean	0.792	0.503	0.507	0.648	0.501	0.230	1.355	0.784	0.460	0.466
Std. dev.	0.794	0.502	0.507	0.643	0.503	0.236	1.353	0.783	0.453	0.468
RMSE	0.055	0.024	0.045	0.167	0.038	0.284	0.377	0.094	0.100	0.099
DGP III:										
Mean	0.793	0.503	0.506	0.664	0.502	0.253	1.412	0.801	0.458	0.474
Std. dev.	0.798	0.503	0.510	0.668	0.499	0.254	1.417	0.807	0.453	0.478
RMSE	0.054	0.024	0.045	0.154	0.039	0.266	0.429	0.106	0.104	0.100

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 8: Simulation of the State Dependence in Dynamic Models

Sample Size=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	0.181	0.180	0.191
Std. dev.	0.006	0.017	0.037
RMSE	–	0.017	0.038
DGP II:			
Mean	0.192	0.139	0.182
Std. dev.	0.006	0.019	0.021
RMSE	–	0.056	0.023
DGP III:			
Mean	0.184	0.139	0.185
Std. dev.	0.005	0.019	0.022
RMSE	–	0.049	0.021

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 9: Simulation of the State Dependence in Dynamic Models

Sample Size=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	0.181	0.180	0.198
Std. dev.	0.006	0.017	0.032
RMSE	–	0.017	0.036
DGP II:			
Mean	0.192	0.139	0.186
Std. dev.	0.006	0.019	0.019
RMSE	–	0.056	0.020
DGP III:			
Mean	0.184	0.139	0.189
Std. dev.	0.005	0.019	0.022
RMSE	–	0.049	0.023

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 10: Estimates of the Persistent Effects of Union Membership

Explanatory Variables	Conventional Estimator	Sieve ML Estimator
Married_t	0.197	0.270
	(0.386)	(0.004)
Union_{t-1}	1.095	0.752
	(1.634)	(0.011)
D_{82}	0.028	0.015
	(0.112)	(0.000)
D_{83}	-0.090	-0.028
	(0.359)	(0.000)
D_{84}	-0.053	-0.066
	(0.161)	(0.001)
D_{85}	-0.257	-0.298
	(0.841)	(0.003)
D_{86}	-0.293	-0.319
	(1.173)	(0.003)
D_{87}	0.068	0.067
	(0.207)	(0.001)
Union_0	1.125	1.467
	(1.559)	(0.027)
Education	-0.011	-0.010
	(0.035)	(0.000)
Black	0.456	0.167
	(1.831)	(0.002)
Constant	-1.608	-1.814
	(2.517)	(0.090)
Variance, σ	0.814	–
	(0.144)	–

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations.

Table 11: Estimated Probability of Being in a Union, 1987

Explanatory Variables	Conventional Estimator	Sieve ML Estimator
$\text{Married}_t = 1$,	0.485	0.580
$\text{Union}_{t-1} = 1$	(0.242)	(0.022)
$\text{Married}_t = 1$,	0.199	0.348
$\text{Union}_{t-1} = 0$	(0.084)	(0.014)
$\text{Married}_t = 0$,	0.444	0.494
$\text{Union}_{t-1} = 1$	(0.221)	(0.020)
$\text{Married}_t = 0$,	0.175	0.276
$\text{Union}_{t-1} = 0$	(0.058)	(0.010)

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 150 simulations. The definition of estimated probabilities is in Eq. (28).