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

Yu-Chin Hsu† Ji-Liang Shiu‡

August 27, 2019

Abstract

Under a Mundlak-type correlated random effect (CRE) specification, we first show that the average likelihood of a parametric nonlinear panel data model is the convolution of the conditional distribution of the model and the distribution of the unobserved heterogeneity. Hence, the distribution of the unobserved heterogeneity can be recovered by means of Fourier transformation without imposing a distributional assumption on the CRE specification. We then construct a semi-parametric family of average likelihood functions of observables by combining the conditional distribution of the model and the recovered distribution of the unobserved heterogeneity, and show that the parameters in the nonlinear panel data model and in the CER specification are identified. Based on the identification result, we propose a sieve maximum likelihood estimator. Compared with the conventional parametric CRE approaches, the advantage of our method is that it is not subject to misspecification on the distribution of CRE. Furthermore, we show that average partial effects are identifiable and extend our results to dynamic nonlinear panel data models. We investigate the finite sample properties of the proposed estimator through a Monte Carlo study.

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

*Helpful comments by Arthur Lewbel, and Matthew Shum are acknowledged. We also thank the editors, and two anonymous referees for useful comments. The authors are solely responsible for any remaining errors. Yu-Chin Hsu gratefully acknowledges research support from the Ministry of Science and Technology of Taiwan (MOST107-2410-H-001-034-MY3) and Career Development Award of Academia Sinica, Taiwan.

†Institute of Economics, Academia Sinica; Department of Finance, National Central University; Department of Economics, National Chengchi University. Email: ychsu@econ.sinica.edu.tw.

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

1. Introduction

How to model unobserved heterogeneity across individuals when the time dimension is fixed is one of the challenges in the panel data literature. In particular, there is a fundamental difference between linear and nonlinear models. For linear panel data models with additive unobserved heterogeneity, under the fixed-effects formulation, one can apply within transformation to eliminate the unobserved effects and consistent estimators can be obtained by generalized method of moments methods without specifying the distribution of the unobserved heterogeneity. 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 Bonhomme (2012) provides a systematic functional differencing approach to construct moment restrictions on common parameters that are free from the individual fixed effects.[2] Other than this, there are two main approaches in the literature. One is to treat the unobserved heterogeneity of each individual as a fixed parameter (fixed effect) and the other is to treat it as a random variable (random effect). However, the resulting identification and estimation strategies of these two approaches are very different.

For the fixed effect approach, the number of parameters increases at the same rate as the sample size so an incidental parameter problem often arises and a standard maximum likelihood estimator is in general biased. For example, Honoré and Kyriazidou (2000) generalize the 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.[3] Chamberlain (2010) considers the identification of binary response models when the time dimension is fixed and the distribution of individual effects is unrestricted. He shows that identification is only possible in the logistic case.[4]

On the other hand, for the random effects approach, one would first specify the conditional

[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), Hoderlein and White (2012), Graham and Powell (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013), Browning and Carro (2014) and Chernozhukov, Fernandez-Val, Hoderlein, Holzmann, and Newey (2015).

[2] The method of Bonhomme (2012) mainly applies to likelihood models with continuous dependent variables.

[3] Honoré and Lewbel (2002) provide 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.

[4] As discussed in Arellano and Bonhomme (2011), the identification problem is related to 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.

distribution of a parametric nonlinear panel data model, i.e., a parametric conditional distribution of the dependent variable, Y_t, conditional on a 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. In the second step of a conventional parametric random effects approach, one can complete the model by specifying the statistical relationship between the unobserved heterogeneity and the observed covariates. To be specific, denote $X=(X_1,\ldots,X_T)$ as 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 of Y given X can then 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 are possible under a standard maximum likelihood estimation (MLE) framework. For example, Wooldridge (2005) handles the initial conditions problem of a dynamic panel data problem by specifying the conditional distribution of the unobserved heterogeneity to be normally distributed with a mean which is a linear combination of the initial value and exogenous explanatory variables. Alvarez and Arellano (2003) use a similar specification for models with large time and cross-sectional dimensions. Arellano and Bonhomme (2009) focus on estimators that maximize an average likelihood that assigns weights to different values of the unobserved heterogeneity. They provide a characterization of the class of weights that produce first-order unbiased estimators.

The disadvantage of the parametric random effects approach is that the misspecification of $f_{C|X}(c|x;\beta)$ would generally result in inconsistent estimates.[5] To avoid this, one can alternatively characterize the identified set of the true parameter, but one would lose point-identification of the parameters. For example, Honoré and Tamer (2006) relax the distributional assump-

[5] See detailed discussion in Hsiao (2015).

tion of the initial condition and calculate bounds on parameters in panel dynamic discrete choice models. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) show 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 method not subject to the distribution misspecification of the average likelihood in Eq. (2) for fixed T that also retains point-identification of the parameters remains intangible.[6]

To fill this gap in the literature, we provide a correlated random effects (CRE) approach in which it is not necessary to fully specify the conditional distributions of the unobserved heterogeneity. Our approach is internally consistent with the parametric nonlinear panel data models in Eq. (1) and we can retain point-identification of the parameters. To be specific, let $\overline{W}$ be a vector of time-invariant observed variables. We consider a Mundlak-type specification such that $C = \lambda\overline{W} + V$, so Eq. (2) can be written as

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

Therefore, the average likelihood is the convolution of the conditional distribution of the parametric nonlinear panel data models and that of the unobserved heterogeneity.[7] The Fourier transform of the conditional distribution of the unobserved heterogeneity can be obtained by the quotient of two Fourier transforms of the density of observables and that of the parametric panel data model evaluated at the true parameter. Next, we extend the relation of the Fourier transform to parameters other than the true parameter and apply the Fourier inversion formula to the extended Fourier transform to devise a parametric distribution of the unobserved heterogeneity conditional on exogenous variables. Combining the parametric nonlinear panel data model, Mundlak-type specification and the recovered distribution of the unobserved heterogeneity, we can construct a correctly specified semi-parametric average likelihood of observables in that it is equal to the density of observables when evaluated at the true parameter, (θ_0, λ_0). We regard the specification as a data-driven one because the distribution of the unobserved heterogeneity is recovered from the parametric nonlinear panel data model and the density of observables.

[6] Section 3.3 of Arellano and Bonhomme (2011) provides detailed discussion.

[7] Our (3) is similar to Wooldridge (2005) except that we require that $C = \lambda\overline{W} + V$.

We then show that its parameter vector is identifiable by the negative definiteness of the information matrix, a standard maximum likelihood condition. Based on the identification result, a sieve maximum likelihood (henceforth sieve ML) estimator of the parameters, (θ_0, λ_0), is proposed, and it is consistent and asymptotically normal. Furthermore, we propose a Hausman-type test to test the distributional assumption in the conventional parametric random effects approach. We also show that the average partial effects can be identified and then extend our method to dynamic nonlinear models.

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 connections are also considered in Honoré and Lewbel (2002), Hu and Shum (2012), and Shiu and Hu (2013). Among them, Shiu and Hu (2013) use the spectral decomposition of linear operators related to observable and unobservable conditional densities and provide nonparametric identification of nonlinear dynamic panel data models. This method relies on the assumption that the dynamic process of the future covariates are independent of the current and lagged dependent variables conditional on the current observed and unobserved covariates. Our identification strategy is also 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. Among them, Evdokimov (2011) establishes nonparametric identification of a panel data model with nonadditive unobserved heterogeneity and develops a nonparametric estimation procedure. Arellano and Bonhomme (2012) consider random coefficients panel data models where the coefficients can be arbitrarily correlated with the covariates and obtained identification of the density of individual effects.

Similar extensions to the Mundlak assumption are considered in regression function models in which the dependent variable is specified as a function of the covariates including Gayle and Viauroux (2007), Gayle (2013), Gayle and Namoro (2013), and Chen, Si, Zhang, and Zhou (2017). Specifically, Gayle (2013) investigates identification and root-n-consistent estimation of a class of linear-single-index panel data models in which the regression function is unknown, the individual effect C depends on $\overline{W}$ in an unknown way, and the distribution of V is also unknown. Compared with Gayle (2013) and Gayle and Namoro (2013), we do not require the

linear-single-index specification and we also provide an identification result of average partial 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. Section 6 concludes. Technical proofs of results are in the Appendix.

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 a more primitive econometric model. Suppose $m(x,c;\theta)$ is a known parametric function. Consider the following binary-choice model:

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

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:

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

This parametric specification contains nonlinear terms of (x_t, c) and includes the linear single-

index binary-choice model considered in the literature.

Another example is the panel data Poisson model, which can also be written as the functional form $f_{Y_t|X_t,C}$. Consider

$$f_{Y_t|X_t,C}(y_t|x_t,c;\theta) = \frac{\exp(-m(x_t,c,;\theta))m(x_t,c,;\theta)^{y_t}}{y_t!} \text{ with } \quad y_t = 0,1,..., \tag{7}$$

where $m(x_t,c;\theta) = \mathrm{E}(Y_t|x_t,c)$ is a parametric mean function. Therefore, the identification result of the panel data model in Eq. (4) developed below is general and can be applied to many types of 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 $\overline{W}$ be a $K_1 \times 1$ vector of time-invariant observed variables with support $\overline{W} \equiv \overline{W}_1 \times \cdots \times \overline{W}_{K_1}$ that is not included in X_t. Alternatively, we can have $\overline{W} = \frac{1}{T+1}\sum_{t=0}^{T}(X_{t1},...,X_{tK_1}) = (\overline{X}_1,...,\overline{X}_{K_1})$ as the time average of the first K_1 explanatory variables of X_t for $t = 0,1,\ldots,T$.[8]

Assumption 2.2. *(Correlated Random Effects)*
Assume that there exists a K_1-dimensional vector of coefficients $\lambda_0 = (\lambda_{01},...,\lambda_{0K_1})' \in \Lambda$ with at least one of these coefficients being non-zero, where Λ is a compact subset of R^{K_1} such that

$$C = \overline{W}\lambda_0 + V, \tag{8}$$

[8]It is necessary to include X_0 so $\overline{W}$ still has variation once we condition on $X_1,\ldots,X_T$. We thank a referee for pointing this out.

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

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. (8) together imply that $f_{C|\overline{W}}(c|\overline{w}) = f_V(c - \overline{w}\lambda_0)$. We note that our Assumption 2.2 is weaker than Wooldridge (2005) because we do not impose a parametric assumption on V; however, ours is less general than Wooldridge (2005) in that we do require that $C = \overline{W}\lambda_0 + V$ and λ_0 is not a zero vector, and Wooldridge (2005) allows C to be arbitrarily correlated with X, while we only allow C to be correlated with X through $\overline{W}$.

The non-zero restriction of λ_0 rules out nonlinear panel-data models with pure random effects and this cannot be tested because our identification results below rely on this condition, but this is not restrictive. For example, this restriction generally holds in empirical applications when the individual unobserved heterogeneity, C, is used to model the time-invariant features of an individual such as cognitive ability or early family upbringing. Because such individual unobserved heterogeneity is generally correlated with the time-invariant explanatory variables such as education or parents' education, a Mundlak-type CRE specification can be justified.

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* $(y_t,x_t,\overline{w},c) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}} \times \mathscr{C}$*;*
(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* $(y_t,x_t,\overline{w},c) \in \mathscr{Y}_t \times \mathscr{X}_t \times \overline{\mathscr{W}} \times \mathscr{C}$*;*
(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* $(c,x,\overline{w}) \in \mathscr{C} \times \mathscr{X} \times \overline{\mathscr{W}}$.

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. Assumption 2.3 allows time-invariant variables to be included in X_t but only those that are not related to individual unobserved heterogeneity.

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 Fourier Transforms)*

Under Assumption 2.2, without loss of generality, set $\lambda_{01} \neq 0$. *Assume that (i) there exists a constant* c_1 *such that* $\int_{\overline{\mathscr{W}}_1} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1 < c_1 < \infty$ *for all* $(y,x,\overline{w}_{-1}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}_{-1}$;

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

$$\left| \int_{\mathscr{C}} e^{-i\xi c} \left(\int_{\mathscr{Y}\times\mathscr{X}} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydx \right) dc \right| > 0. \tag{9}$$

Assumptions 2.4(i) & (ii) ensure that the Fourier transforms of the density of observables $f_{Y|X,\overline{W}}$ and the proposed panel data model $f_{Y|X,C}$ are well defined. The assumption implies that the Fourier transforms $\lambda_1 \int_{\overline{\mathscr{W}}_1} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1$ for $(y,x,\overline{w}_{-1}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}_{-1}$ 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 all finite. $f_{Y|X,C}(y|x,c;\theta)$ is uniformly bounded for all $\theta \in \Theta$ by Assumption 2.1, so 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(i-i) 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(i-i) holds.[9] 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 Fourier transforms in Eq. (9) appear as denominators in our identifying formula and Assumption 2.4(iii) rules out a zero denominator. 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 $\mathscr{A} = \Theta \times \Lambda$. Consider the following parametric function

$$\phi_{v;\alpha}(\xi) \equiv \frac{\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}_{-1}} h(\xi,y,x,\overline{w};\lambda)\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}_{-1}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydxdc}, \tag{10}$$

where $h(\xi,y,x,\overline{w};\lambda_0) = -\lambda_{01}\int_{\overline{\mathscr{W}}_1} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_1$, $\Omega(y,x,\overline{w}_{-1})$ is a positive weighting function, and $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$.[10] While Assumption 2.4(i) ensures the term on the top of the

[9] 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 $|\frac{1}{\frac{\partial m}{\partial c}}| < \frac{1}{a}$.

[10] The function $\phi_{v;\alpha}$ is also defined in Eq. (A.7) and its detailed derivation can be found there.

quotient in Eq. (10) is well defined, Assumption 2.4(ii) and (iii) guarantees the term on the bottom of the quotient in Eq. (10) is well defined and non-zero. We can use $\phi_{v;\alpha}$ to construct a semi-parametric family of functions related to the Fourier transform of the remainder term V.

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) there exists a nonnegative integrable function g such that for all α

$$|\phi_{v;\alpha}(\xi)| \leq g(\xi). \tag{11}$$

A sufficient condition for Assumption 2.5(ii) is that the density function $f_{Y|X,\overline{W}}$ and the domain $\overline{\mathscr{W}}_1$ are bounded, and there exists $p > 1$ such that for all $\theta \in \Theta$,

$$\left| \int_{\mathscr{C}} e^{-i\xi c} \left(\int_{\mathscr{Y}\times\mathscr{X}} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydx \right) dc \right| \geq c(1+|\xi|)^p. \tag{12}$$

Under the sufficient condition, we obtain

$$|\phi_{v;\alpha}(\xi)| \leq c|\overline{\mathscr{W}}_1|(1+|\xi|)^{-p} = g(\xi), \tag{13}$$

where c is some constant and $|\overline{\mathscr{W}}_1|$ is the length of $\overline{\mathscr{W}}_1$. Under Assumption 2.5, we can apply Fourier Inversion Formula in Proposition A.2 to the semi-parametric family of functions $\{\phi_{V;\alpha}(\xi) : \alpha \in \Theta \times \Lambda\}$ in Eq. (A.7) to construct a semi-parametric family of density functions $\{f_{C|\overline{W}}(c|\overline{w};\alpha) : \alpha \in \mathscr{A}\}$ in Eq. (A.13). Because the semi-parametric Fourier transform $\phi_{V;\alpha}(\xi)$ 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.[11] We summarize that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is a parametric density function over $\mathscr{C}$ in the following result.

Lemma 2.1. *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 $f_{C|\overline{W}}(c|\overline{w};\alpha)$ in Eq. (A.13) is a conditional density function for α in the neighborhood.*

[11] See details in Eq. (A.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:

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)\ \underbrace{f_{C|\overline{W}}(c|\overline{w};\alpha)}_{\substack{\text{constructed from} \\ f(y|x,\overline{w}) \text{ and} \\ f_{Y|X,C}(y|x,c;\theta)}}\ dc. \tag{14}$$

Equation (14) 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 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, the identification of the parametric system is approached via the concavity of the conditional Kullback-Leibler information criterion evaluated at α_0. Define

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

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,

$$\frac{\partial}{\partial\alpha}K(\alpha;x,\overline{w})=\left(\frac{\partial K(\alpha;x,\overline{w})}{\partial\alpha_1},...,\frac{\partial K(\alpha;x,\overline{w})}{\partial\alpha_{d_\theta+K_1}}\right)'. \tag{15}$$

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:

$$K''(\alpha_0;x,\overline{w})=-\mathrm{E}\Big[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0}\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)'\big|_{\alpha=\alpha_0}\Big|X=x,\overline{W}=\overline{w}\Big]. \tag{16}$$

Assumption 2.6. *(Concave Parameter Structure)*
Assume that the information matrix $K''(\alpha_0;x,\overline{w})$ in Eq. (16) 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 $t=1,2,...,T$. In particular, the density function of the remainder term V in CRE is also identified.*

We prove Theorem 2.1 in four steps and we have summarized them in Appendix A.1 to make the identification strategy more transparent.

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 population parameter $\alpha_0=(\theta_0,\lambda_0)$ and f_V based on the parametric specification for the density function of observable variables in Eq. (A.1):

$$f_{Y|X,\overline{W}}(y|x,\overline{w};\alpha,f_1)=\int_{\mathscr{C}}f_{Y|X,C}(y|x,c;\theta)f_1(c-\overline{w}\lambda)dc, \tag{17}$$

where $\alpha = (\theta, \lambda)$ is a finite-dimensional parameter and f_1 is an infinite-dimensional parameter. The proposed estimator is semi-parametric in that the distribution of the CRE remainder error V in Assumption 2.2 is not specified. The fact that the population parameters α_0 and the distribution of the remainder error f_V are identified implies that (α_0, f_V) is the unique solution to the following problem:

$$(\alpha_0, f_V) \equiv \arg \max_{\alpha \in \mathscr{A}, f_1 \in \mathscr{F}_1} \mathrm{E}\Big[\log f(Y|X, \overline{W}; \alpha, f_1) \Big], \tag{18}$$

where $\mathscr{A} \subset R^{d_\theta + K_1}$ is compact with a non-empty interior and $\mathscr{F}_1$ is a space of density functions with support equal to that of V. However, when the parameter function space containing f_1 is large, the direct ML estimation method based on the sample analog of (18) could yield an inconsistent estimator or a consistent estimator which converges very slowly. Thus, we replace $\mathscr{F}_1$ with a finite dimensional sieve space $\mathscr{F}_{1N}$ that becomes dense in $\mathscr{F}_1$ where the sample size N increases. To be specific, let $\{(Y_i, X_i, \overline{W}_i)\}_{i=1}^N$ be a sample of observed variables. The empirical analogue of the expression in Eq. (18) is given by

$$\widehat{Q}_N(\alpha, f_1) = \frac{1}{N} \sum_{i=1}^{N} \log f_{Y|X,\overline{W}}(Y_i|X_i, \overline{W}_i; \alpha, f_1). \tag{19}$$

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

$$(\widehat{\alpha}, \widehat{f_1}) \equiv \arg \max_{\alpha \in \mathscr{A}, f_1 \in \mathscr{F}_{1N}} \widehat{Q}_N(\alpha, f_1). \tag{20}$$

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), Ai and Chen (2003), Chen, Liao, and Sun (2014) and Hahn, Liao, and Ridder (2018). We give the expression of the limiting distribution of the $\widehat{\alpha}$ and sufficient conditions in the Appendix. In the Appendix, we also give a consistent estimator for the asymptotic covariance matrix for inference. In particular, by Hahn, Liao, and Ridder (2018), if the sieve space is generated by a finite number of basis functions such as the Hermite polynomial series we use in the simulations, then the expression of the sample analog consistent estimator for the asymptotic covariance matrix can be obtained through the conventional MLE method. Please see Appendix C for details.

We consider Hermite polynomial series as our sieve basis functions for the nonparamet-

ric nuisance components, $\sqrt{f_1(\cdot)}$. Denote $\phi(\cdot)$ and $H_j(\cdot)$ as the PDF of standard normal and the j-th order Hermite polynomial. It follows that $h_j(\cdot) = H_j(\cdot)\phi(\cdot)$ form an orthogonal series of the square-integral function space. A Hermite polynomial series estimator $\sqrt{f_1(\cdot)}$ can be constructed by

$$\sqrt{f_1(v)} = \sum_{j=0}^{J} \beta_j h_j(v). \tag{21}$$

Because $f_V(\cdot)$ is a density function, the density restriction $\int f_V(v)dv = 1$ imposes a restriction on these sieve coefficients, $\sqrt{2\pi}\sum_{j=0}^{J} j!\beta_j^2 = 1$. By substituting the parametric specification of $f_{Y|X,C}(y|x,c)$ and the Hermite polynomial series of f_1 into Eq. (17) we obtain:

$$f_{Y|X,\overline{W}}(y|x,\overline{w};\theta,\lambda,\beta) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)(\sum_{j=0}^{J} \beta_j h_j(c-\overline{w}\lambda))^2 dc. \tag{22}$$

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. (18) is given by

$$\widehat{Q}_N(\alpha,\beta) = \frac{1}{N}\sum_{i=1}^{N} \log f_{Y|X,\overline{W}}(y_i|x_i,\overline{w}_i;\alpha,\beta). \tag{23}$$

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

$$(\widehat{\alpha},\widehat{\beta}) \equiv \arg\max_{\alpha,\beta} \widehat{Q}_N(\alpha,\beta), \tag{24}$$

with $\widehat{f}_1(v) = \big(\sum_{j=0}^{J} \widehat{\beta}_j h_j(v)\big)^2$.

4. Discussion

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

4.1. Specification Test for Fully Parametric Model

A popular approach in the literature is to impose a distribution assumption on $V \sim f_V(v;\tau_0)$ so as to obtain a fully parametric likelihood model and apply the parametric ML method. For example, one may impose a normality assumption on V so that $V \sim f_V(v;\tau_0) = \tau_0^{-1/2}\phi(v \cdot \tau_0^{-1/2})$ for some $\tau_0 > 0$.[12] Here, we provide a specification test for such an assumption. 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 $V \sim f_V(v;\tau_0)$ for some τ_0, then we have $C|\overline{W} \sim f_V(c - \overline{W}\lambda;\tau_0)$. Therefore, the full parametric M-LE for (α_0,τ_0) is defined as,

$$(\widehat{\alpha}_{pa},\widehat{\tau}_{pa}) \equiv \arg\max_{\alpha\in\mathscr{A},\tau\in\mathscr{T}} \frac{1}{N}\sum_{i=1}^{N} f_{pa}(Y_i|X_i,\overline{W}_i;\alpha,\tau), \tag{25}$$

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

where $\mathscr{T}$ is a compact parameter space of τ. Under suitable conditions, we can show that $\sqrt{N}(\widehat{\alpha}_{pa} - \alpha_0) \xrightarrow{d} N(0,\Omega_{pa})$ where $\mathscr{V}_{pa}$ is the asymptotic variance and covariance matrix of the fully parametric estimator.

We can construct a Hausman-type test for the null hypothesis that $V \sim f_V(v;\tau_0)$ by testing if $\widehat{\alpha}$ and $\widehat{\alpha}_{pa}$ are close to each other. To be specific, both our estimator and the fully parametric estimator for α_0 are consistent for α_0 when $V \sim f(v;\tau)$. However, if $V \nsim f_V(v;\tau)$ for any $\tau \in \mathscr{T}$, 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 $\sqrt{N}(\widehat{\alpha} - \widehat{\alpha}_{pa}) \xrightarrow{d} N(0,\Omega_{di})$, where Ω_{di} stands for the asymptotic covariance of $\sqrt{N}(\widehat{\alpha} - \widehat{\alpha}_{pa})$ under null. Let $\widehat{\Omega}_{di}$ denote a consistent estimator for Ω_{di}. Define the test statistics as

$$\widehat{S}_N = N(\widehat{\alpha} - \widehat{\alpha}_{pa})(\widehat{\Omega}_{di})^{-1}(\widehat{\alpha} - \widehat{\alpha}_{pa})' \tag{26}$$

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

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. A formula of $\widehat{\Omega}_{di}$ is given in Appendix D.

4.2. Extension to Dynamic Nonlinear Panel Data Models

Consider a parametric dynamic panel data density function:

$$f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c;\theta), \text{ for all } t=2,...,T. \tag{27}$$

Dynamic panel data models can be used to investigate the effects of lagged outcomes on current outcomes, so the identification and estimation of models in Eq. (27) are of great practical value. We show that the identification strategy for the static panel data model in Theorem 2.1 can be adapted easily to dynamic panel data models. First, it is straightforward to maintain Assumptions 2.1, 2.2, 2.4, 2.5, and 2.6 in this dynamic setting. However, we also need the following assumptions for identification in the setting.

Assumption 4.1. *(Correlated Random Effects)*
Assume that there exist a $\gamma_0 \in \Gamma$ *and a* K_1*-dimensional vector of coefficients* $\lambda_0 = (\lambda_{01},...,\lambda_{0K_1})' \in \Lambda$ *with at least one of these coefficients being non-zero, where* Γ *is a compact interval and* Λ *is a compact subset of* R^{K_1+1} *such that*

$$C = \gamma_0 Y_1 + \overline{W}\lambda_0 + V, \tag{28}$$

where the remainder term V *is independent of* Y_1 *and* $\overline{W}$.

Assumption 4.1 is an extension of Assumption 2.2 so we can include the initial condition of the outcome Y_1. Assumption 4.1 implies that $f_{C|X,Y_1,\overline{W}}(c|x,y_1,\overline{w}) = f_{C|Y_1,\overline{W}}(c|y_1,\overline{w})$ for all $(c,x,y_1,\overline{w}) \in \mathscr{C} \times \mathscr{X} \times \mathscr{Y}_1 \times \overline{\mathscr{W}}$.

Assumption 4.2. *(Movement of the Unobserved Effects)*
Assume that $f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t,y_{t-1},\overline{w},c) = f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c)$ *for all* $(y_t,x_t,y_{t-1},\overline{w},c) \in \mathscr{Y}_t \times \mathscr{X}_t \times \mathscr{Y}_{t-1} \times \overline{\mathscr{W}} \times \mathscr{C}$.

Similar to Eq. (A.1), we can use Assumptions 4.1 and 4.2(i) & (ii) to obtain

$$
\begin{aligned}
&f_{Y_2,Y_3,...,Y_T|X,Y_1,\overline{W}}(y_2,y_3,...,y_T|x,y_1,\overline{w}) \\
&= \int_{\mathscr{C}_T} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t,y_{t-1},\overline{w},c) f_{C|X_t,Y_{t-1},\overline{W}}(c|x,y_1,\overline{w})dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c) f_{C|Y_1,\overline{W}}(c|y_1,\overline{w})dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t,y_{t-1},c;\theta_0) f_V(c-y_1\gamma_0-\overline{w}\lambda_0)dc. \qquad (29)
\end{aligned}
$$

That is, the observable density function $f_{Y_2,Y_3,...,Y_T|X,Y_1,\overline{W}}$ 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 V. Therefore, sufficient conditions for identification are similar to those in Theorem 2.1 and the identification results follow. Wooldridge (2005) considers the same specification, but assumes that 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}:

$$
\frac{\partial \mathrm{E}[Y_t|X_t,C]}{\partial X_{tk}}. \qquad (30)
$$

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:

$$\begin{aligned} \mathrm{APE}(x_{tk}) &= \int_{\mathscr{C}} \left(\frac{\partial \mathrm{E}[Y_t | X_t = x_t, C = c]}{\partial X_{tk}} \right) f_C(c) dc \\ &= \int_{\mathscr{C}} \left[\int_{\mathscr{Y}_t} y_t \frac{\partial f_{Y_t|X_t,C}(y_t|x_t,c;\theta_0)}{\partial x_{tk}} dy_t \right] f_C(c) dc. \end{aligned} \tag{31}$$

We now state our identification result for APE.

Corollary 4.1. *Under Assumptions 2.1- 2.6, the APE defined in Eq. (31) 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

$$\mathrm{APE}(x_{tk}) = \int_{\mathscr{C}} \frac{\partial F_{\varepsilon_t}(-m(x,c;\theta))}{\partial x_{tk}} \frac{\partial m(x,c;\theta)}{\partial x_{tk}} f_C(c) dc, \tag{32}$$

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

$$\begin{aligned} \mathrm{APE}(x_{td_\theta}) = \int_{\mathscr{C}} \Big(& F_{\varepsilon_t}\left(-m(x_{t1}, ..., x_{td_\theta - 1}, 1, x_{td_\theta + 1}, ..., x_{tK}, c; \theta\right) \\ & - F_{\varepsilon_t}\left(-m(x_{t1}, ..., x_{td_\theta - 1}, 0, x_{td_\theta + 1}, ..., x_{tK}, c; \theta\right) \Big) f_C(c) dc. \end{aligned} \tag{33}$$

For a parametric dynamic panel data model in Eq. (27), researchers 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.

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 model in Section 3. We consider both panel data probit model models and panel data Poisson models.

5.1. Panel Data Probit Models

We demonstrate our simulation results through static and dynamic setting. The *static* data generating process (DGP) is defined as follows:

$$
\begin{aligned}
&Y_t = \mathbf{1}(\theta X_t + C + \varepsilon_t \geq 0), \text{ for } t = 1,2, \\
&C = \lambda \overline{W} + V, \ \ \overline{W} = \frac{1}{2}\sum_{t=1}^{2} X_t, \\
&X_2 = 0.5X_1 + \xi, \ \ X_1 \sim U(0,2), \ \ \xi \sim N(0,1), \\
&(\varepsilon_1, \varepsilon_2) \sim N(0, I_2),
\end{aligned}
$$

where I_2 is the 2×2 identify matrix and $(\theta, \lambda) = (-0.5, 0.5)$. For a random variable Q, we denote the corresponding truncated random variable over interval $[a,b]$ as $Trun(Q,[a,b])$.[13] Let μ_ω be the mean of ω. Three specifications of V are considered:

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

The unobserved heterogeneities in all the simulation designs have 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 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

[13] $Trun(Q,[a,b])$ 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 $[0,1]$.

Tables 1–2 and 3–4, respectively.

The estimation results of the parameters in DGP I show a little bias in all 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 small degrees of bias in DGPs II & III but the conventional random effects estimator exhibits conspicuous 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 & III. 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 performs 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 9. For DGP I, the rejection rates are 0.080 and 0.047, which are close to the nominal size 5% given that the normality assumption holds.[14] 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.

The simulation design for dynamic panel data probit models is close to the static panel data probit models. The DGP for *dynamic* models is defined as follows:

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

where $(\gamma, \theta, \lambda) = (-0.5, 0.5, 0.5)$ 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

[14] We use an empirical covariance matrix which is the average of the 150 simulated estimators to conduct the Hausman test in the simulations.

in DGPs II & III. The simulation results for the Hausman-type test are similar to the static case too.

5.2. Dynamic Panel Data Poisson Models

We consider a data generating process for dynamic panel data Poisson models as follows:

$$m(y_{t-1}, x_t, c; \theta) = \gamma Y_{t-1} + \theta X_t + C \text{ with } \quad y_t = 0, 1, ..., \tag{34}$$

where $m(y_{t-1}, x_t, c; \theta) = \mathrm{E}(Y_t | y_{t-1}, x_t, c)$ and $(\gamma, \theta, \lambda) = (-0.5, 0.5)$. While the DGP for X_t is $X_1 \sim U(0,2)$, $X_t \sim N(0,1)$ for $t > 1$, the DGP for C is $C = 0.5\overline{W} + V$, $\overline{W} = \frac{1}{T}\sum_{t=1}^{T} X_t + z$ with $z \sim N(0,1)$. In this case, we consider unbounded support for V and three specifications are considered:

DGP IV: $V \sim N(0,1)$,

DGP V: $V = \omega - \mu_\omega$ with $\omega \sim$ Student's t with 100 degrees of freedom,

DGP VI: $V = \omega - \mu_\omega$ with $\omega = \omega_1 + \omega_2, \omega_1 \sim N(0,1), \omega_2 \sim Trun(Rayleigh(1), [-2, 12])$.

The finite-sample performance is evaluated over two different time dimensions, $T = 2$ and $T = 4$.

While Tables 10–13 report the estimated results for panel data of two periods, Tables 14–17 report the estimated results for panel data of four periods. From the estimation results we find that in the designs of panel data of two periods, the proposed estimator outperforms the conventional estimator in the parameter estimation. However, for a longer periods such as four periods, the conventional estimator performs well and is close to the proposed estimator. Table 18 reports the Hausman-type test proposed for the normality assumption of V. In the simulated data of two periods, the rejection rates of the proposed Hausman-type test are much higher than the nominal size 5% and increase with sample size in all designs. This implies that the conventional estimator performs poorly even for the case in which V is normally distributed. However, in the simulated data of four periods with a sample size of 1000, the estimation results of the conventional estimator are close to that of the proposed estimator and the rejection rates are 0.060, 0.067, and 0.087. This suggests that under an unbounded assumption of V, the conventional estimator may perform better for panel data with a longer period and with a

larger sample size.

6. 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. We show that 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. The identification strategy can also be applied to dynamic nonlinear panel data models.

Appendix

A. Proof of Theorem 2.1

Consider

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

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

Under Assumption 2.2, without loss of generality, assume $\lambda_{01} \neq 0$. Given each (y,x), constructing a Fourier transform 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,

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

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

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

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

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

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). For example, if we consider the simplest case, $\lambda_0 = (\lambda_{01},...,\lambda_{0K_1})^T = (-1,0,...,0)'$, Eq. (A.4) becomes

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

Denote $w_{-1} = (w_2, w_3, \ldots, w_T)$ and $w_{-1} \in \overline{W}_{-1} \equiv \overline{\mathscr{W}}_2 \times \cdots \times \overline{\mathscr{W}}_T$. The result in Eq. (A.4) holds for every $(y,x,\overline{w}_{-1}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}_{-1}$. As such, we can utilize a positive weighting function $\Omega(y,x,\overline{w}_{-1})$. Multiplying the equation by $\Omega(y,x,\overline{w}_{-1})$, integrating out the variables $(y,x,\overline{w}_{-1})$ over the domain, and then interchanging the integrations, we obtain

$$\begin{aligned}&\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}_{-1}} \left(-\lambda_{0j}\int_{\overline{\mathscr{W}}_1} e^{-i\xi\sum_{k=1}^{K}\lambda_{0k}\overline{w}_k} \underbrace{f_{Y|X,\overline{W}}(y|x,\overline{w})}_{\substack{\text{observable}\\ \text{from data}}} d\overline{w}_1\right)\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}_{-1} \\ &= \phi_v(\xi)\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} \underbrace{f_{Y|X,C}(y|x,c;\theta_0)}_{\substack{\text{population}\\ \text{density}}} \Omega(y,x)dydxdc,\end{aligned} \tag{A.5}$$

where $\int_{\overline{\mathscr{W}}_{-1}} \Omega(y,x,\overline{w}_{-1})d\overline{w}_{-1} = \Omega(y,x)$.

Assumptions 2.4(i) & (ii) imply that the Fourier transforms other than $\phi_v(\xi)$ in Eq. (A.5) are well defined. By Assumption 2.4(iii), dividing both sides of Eq. (A.5) 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 Fourier transform of the remainder term V in the CRE assumption,

$$\phi_v(\xi) = \frac{-\lambda_{01}\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)\Omega(y,x)dydxdc}. \tag{A.6}$$

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.6) to all $\alpha \in \Theta \times \Lambda$ to obtain a potential semi-parametric family of functions connected to the Fourier transforms of the distribution of the remainder term v in the following form

$$\phi_{v;\alpha}(\xi) \equiv \frac{-\lambda_1\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydxdc}, \tag{A.7}$$

where $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$. Notice that the terms in the numerator and denominator of the fraction in Eq. (A.6) 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 $\phi_{v;\alpha}(\xi)$ yields a semi-parametric family of the function of V as

$$f_{v;\alpha}(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi v}\phi_{v;\alpha}(\xi)d\xi. \tag{A.8}$$

Under Assumption 2.5(ii), the Fourier transform $\phi_v(\cdot)$ belongs to $L^1(\mathbb{R})$ and 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 the Fourier inversion formula. The following Fourier inversion theorem comes from Chapter 7.5 in Folland (2009):

Proposition A.1. *(Fourier Inversion Theorem) Suppose $\widehat{f}$ is the Fourier transform of f. If f and $\widehat{f}$ are both integrable and f is continuous, then*

$$f(x) = (2\pi)^{-n}\int_{\mathbb{R}^n} e^{ix\cdot\xi}\widehat{f}(\xi)d\xi. \tag{A.9}$$

In order to show which function space satisfies the Fourier inversion theorem, we introduce the Schwartz class $\mathscr{S}(\mathbb{R}^n)$ as follows. Given an $n\times 1$ 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,

$$\sup_{x\in\mathbb{R}^n}|x^a D^b g(x)| = c_{a,b}(g) < \infty, \tag{A.10}$$

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

$$f(x) = (2\pi)^{-n}\int_{\mathbb{R}^n} e^{ix\cdot\xi}\widehat{f}(\xi)d\xi. \tag{A.11}$$

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

$$c_\alpha(\overline{w}) \equiv \int_{\mathscr{C}} f_{v;\alpha}(c - \overline{w}\lambda) dc, \tag{A.12}$$

where $c_{\alpha_0}(\overline{w}) = \int_{\mathscr{C}} f_{v;\alpha_0}(c - \overline{w}\lambda_0) dc = \int_{\mathscr{C}} f_V(c - \overline{w}\lambda_0) dc = 1$. Use the scale factor $c_\alpha(\overline{w})$ to normalize $f_{v;\alpha}$ to construct the following semi-parametric family of conditional density functions of the unobserved heterogeneity

$$f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{c_\alpha(\overline{w})} f_{v;\alpha}(c - \overline{w}\lambda) \tag{A.13}$$

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

Note that when $\mathscr{C} = \mathbb{R}$, the scale factor $c_\alpha(\overline{w})$ has a simpler expression. The last term of the integrand in Eq. (A.12) 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 and compactly supported functions $f(\cdot)$. With this property, if $\phi_{v;\alpha}(\xi)$ is continuous compactly supported then Eq. (A.12) can be further reduced as

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

Proof of Lemma 2.1: 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.12), we can obtain that 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 $f_{Y|X,C}(y|x,c;\theta)$ leads to the following semi-parametric function of observable variables:

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha) dc. \tag{A.15}$$

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

$$\begin{aligned} \int_{\mathscr{Y}} f(y|x,\overline{w};\alpha) dy &= \int_{\mathscr{C}} \left(\int_{\mathscr{Y}} f_{Y|X,C}(y|x,c;\theta) dy \right) f_{C|\overline{W}}(c|\overline{w};\alpha) dc \\ &= \int_{\mathscr{C}} f_{C|\overline{W}}(c|\overline{w};\alpha) dc = 1. \end{aligned} \tag{A.16}$$

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

Lemma A.1. *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 $(x,\overline{w})\in\mathscr{X}\times\overline{\mathscr{W}}$.*

Proof: The proof uses 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.16) 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. (B.4), we have $\frac{\partial}{\partial\alpha}K(\alpha;x,\overline{w})\big|_{\alpha=\alpha_0}=0$. Similarly, differentiating Eq. (A.16) twice and applying the result to the second derivative of $K(\alpha;x,\overline{w})$, the matrix of the second derivatives can be written as minus outer product of the gradient of the log likelihood:

$$K''(\alpha_0;x,\overline{w})=-\mathrm{E}\Big[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0}\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)'\big|_{\alpha=\alpha_0}\Big|X=x,\overline{W}=\overline{w}\Big]. \tag{A.17}$$

Proof of Theorem 2.1: First we have discussed that the semi-parametric density function of observable variables in Eq. (A.16) 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 identification result of α_0 using concavity of conditional Kullback-Leibler information criterion in Assumption 2.6, i.e., the second derivative of $K(\alpha;x,\overline{w})$ in Eq. (16) is negative definite. Applying the identification result of α_0 to Eq. (A.8) leads to the identification of the density $f_V(\cdot)$.

Q.E.D.

A.1. Summary of Identification Steps

In this subsection, we present a heuristic sketch of how to utilize 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.5) and then extend the relationship to parameters other than the true parameter to obtain the semi-parametric function in Eq. (A.7).

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.13) 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. (B.1).

B. Primitive Conditions for Assumption 2.6

The internally consistent semi-parametric density of observable variables has the following form:

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha)dc, \tag{B.1}$$

where $f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi$ with

$$\phi_{v;\alpha}(\xi) = \frac{-\lambda_1 \int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydxdc}. \tag{B.2}$$

For illustrate, we consider $\alpha = (\theta_1, \lambda_1, \lambda_2)$, i.e. θ is a scalar and λ is a 2×1 vector. Similar lines of argument show that the same conclusion holds for general cases. The scale factor can be written as

follows:

$$c_\alpha(\overline{w}) \equiv \int_{\mathscr{C}} f_{v;\alpha}(c-\overline{w}\lambda)dc = \frac{1}{2\pi}\int_{\mathscr{C}}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi dc \\ = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc\right)e^{i\xi\overline{w}\lambda}\phi_{v;\alpha}(\xi)d\xi. \tag{B.3}$$

The gradient of the log likelihood at the true value α_0 in this simple case is

$$\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0) \\ = \frac{1}{f(Y|X,\overline{W};\alpha)}\Big(\frac{\partial}{\partial\theta_1}f(Y|X,\overline{W};\alpha_0), \frac{\partial}{\partial\lambda_1}f(Y|X,\overline{W};\alpha_0), \frac{\partial}{\partial\lambda_2}f(Y|X,\overline{W};\alpha_0)\Big)', \tag{B.4}$$

with

$$\frac{\partial}{\partial\theta_1}f(Y|X,\overline{W};\alpha_0) = \int_{\mathscr{C}}\frac{\partial}{\partial\theta_1}f_{Y|X,C}(y|x,c;\theta_0)f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc \\ + \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0)\underbrace{\frac{\partial}{\partial\theta_1}f_{C|\overline{W}}(c|\overline{w};\alpha_0)}_{\substack{\text{related to } f(y|x,\overline{w}) \\ \text{and } f_{Y|X,C}(y|x,c;\theta)}}dc, \tag{B.5}$$

$$\frac{\partial}{\partial\lambda_1}f(Y|X,\overline{W};\alpha_0) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0)\frac{\partial}{\partial\lambda_1}f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc, \tag{B.6}$$

$$\frac{\partial}{\partial\lambda_2}f(Y|X,\overline{W};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0)\frac{\partial}{\partial\lambda_2}f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc. \tag{B.7}$$

Compared with the conventional CRE approach that $V \sim N(0,1)$, and $f_{C|\overline{W}}(c|\overline{w};\lambda) = \phi(c-\overline{w}\lambda)$ where $\phi(\cdot)$ denotes the density function of standard normal, the derivative term $\frac{\partial}{\partial\theta_1}f_{C|\overline{W}}(c|\overline{w};\alpha)$ in Eq. (B.5) in our model is generally non-zero and related to the sample density and the proposed model. This implies that the term $\int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0)\frac{\partial}{\partial\theta_1}f_{C|\overline{W}}(c|\overline{w};\alpha_0)$ in our model is used to accommodate the modelling of the panel data structure $f_{Y|X,C}(y|x,c;\theta)$ and the distribution of the unobserved heterogeneity.

To provide a more primitive condition on the densities, we need to write out the derivatives of the composite distribution of the unobserved heterogeneity $\frac{\partial}{\partial\theta_1}f_{C|\overline{W};\alpha}$, $\frac{\partial}{\partial\lambda_1}f_{C|\overline{W};\alpha}$, and $\frac{\partial}{\partial\lambda_2}f_{C|\overline{W};\alpha}$ at the true

value α_0, where $f_{C|\overline{W};\alpha} \equiv f_{C|\overline{W}}(c|\overline{w};\alpha)$. We have

$$\begin{aligned}
\frac{\partial \phi_{v;\alpha}(\xi)}{\partial \theta_1} &= \phi_{v;\alpha}(\xi) \times \frac{-\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} \frac{\partial f_{Y|X,C}(y|x,c;\theta)}{\partial \theta_1} \Omega(y,x) dy dx dc}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x) dy dx dc} \equiv \phi_{v;\alpha}(\xi)\gamma_{\theta_1}(\xi;\theta), \\
\frac{\partial \phi_{v;\alpha}(\xi)}{\partial \lambda_1} &= \frac{1}{\lambda_1}\phi_{v;\alpha}(\xi) + \frac{i\lambda_1\xi \int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k}\overline{w}_1 f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1}) dy dx d\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x) dy dx dc} \\
&\equiv \frac{1}{\lambda_1}\phi_{v;\alpha}(\xi) + i\lambda_1\xi\gamma_{\lambda_1}(\xi;\alpha), \\
\frac{\partial \phi_{v;\alpha}(\xi)}{\partial \lambda_2} &= \frac{i\lambda_1\xi \int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k}\overline{w}_2 f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1}) dy dx d\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x) dy dx dc} \\
&\equiv i\lambda_1\xi\gamma_{\lambda_2}(\xi;\alpha).
\end{aligned}$$

Notice that these derivatives are mainly stated in terms of the proposed parametric nonlinear panel data model $f_{Y|X,C}(y|x,c;\theta)$ and the density of observables $f_{Y|X,\overline{W}}(y|x,\overline{w})$. These derivatives evaluated at the true parameter α_0 are

$$\begin{aligned}
\frac{\partial \phi_{v;\alpha_0}(\xi)}{\partial \theta_1} &= \phi_v(\xi)\gamma_{\theta_1}(\xi;\theta_0), \\
\frac{\partial \phi_{v;\alpha_0}(\xi)}{\partial \lambda_1} &= \frac{1}{\lambda_1}\phi_v(\xi) + i\lambda_{01}\xi\gamma_{\lambda_1}(\xi;\alpha_0), \\
\frac{\partial \phi_{v;\alpha_0}(\xi)}{\partial \lambda_2} &= i\lambda_{01}\xi\gamma_{\lambda_2}(\xi;\alpha).
\end{aligned}$$

It follows that

(B.8)
$$\begin{aligned}
&\frac{\partial}{\partial\theta_1} f_{C|\overline{W};\alpha} \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\frac{\partial c_\alpha(\overline{w})}{\partial\theta_1}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\theta_1}d\xi \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\theta_1}d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi \\
&\quad + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\theta_1}d\xi \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda}\phi_{v;\alpha}(\xi)\gamma_{\theta_1}(\xi;\theta)d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi \\
&\quad + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)\gamma_{\theta_1}(\xi;\theta)d\xi.
\end{aligned}$$

(B.9)
$$\begin{aligned}
&\frac{\partial}{\partial\lambda_1} f_{C|\overline{W};\alpha} \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\frac{\partial c_\alpha(\overline{w})}{\partial\lambda_1}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} i\xi\overline{w}_1 e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi \\
&\quad + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\lambda_1}d\xi \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda}(i\xi\overline{w}_1\phi_{v;\alpha}(\xi) + \frac{\partial\phi_{v;\alpha}(\xi)}{\partial\lambda_1})d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi. \\
&\quad + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} i\xi\overline{w}_1 e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\lambda_1}d\xi.
\end{aligned}$$

(B.10)
$$\begin{aligned}
&\frac{\partial}{\partial\lambda_2} f_{C|\overline{W};\alpha} \\
&= \frac{-1}{2\pi c_\alpha(\overline{w})^2}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda}(i\xi\overline{w}_2\phi_{v;\alpha}(\xi) + \frac{\partial\phi_{v;\alpha}(\xi)}{\partial\lambda_2})d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi \\
&\quad + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} i\xi\overline{w}_2 e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi + \frac{1}{2\pi c_\alpha(\overline{w})}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\frac{\partial\phi_{v;\alpha}(\xi)}{\partial\lambda_2}d\xi.
\end{aligned}$$

The derivatives at the true value α_0 can be stated as follows

$$\begin{aligned}
\text{(B.11)} \qquad \frac{\partial}{\partial\theta_1} f_{C|\overline{W};\alpha_0} &= \frac{-1}{2\pi}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda_0}\phi_v(\xi)\gamma_{\theta_{01}}(\xi;\theta_0)d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi \\
&\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)\gamma_{\theta_1}(\xi;\theta_0)d\xi, \\
\text{(B.12)} \qquad &\frac{\partial}{\partial\lambda_1} f_{C|\overline{W};\alpha_0} \\
&= \frac{-1}{2\pi}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda_0}(i\xi\overline{w}_{01}\phi_v(\xi) + \frac{\partial\phi_{v;\alpha_0}(\xi)}{\partial\lambda_1})d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi \\
&\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} i\xi\overline{w}_{01}e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\frac{\partial\phi_{v;\alpha_0}(\xi)}{\partial\lambda_1}d\xi, \\
\text{(B.13)} \qquad &\frac{\partial}{\partial\lambda_2} f_{C|\overline{W};\alpha_0} \\
&= \frac{-1}{2\pi}\left(\int_{-\infty}^{\infty}(\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc)e^{i\xi\overline{w}\lambda_0}(i\xi\overline{w}_{02}\phi_v(\xi) + \frac{\partial\phi_{v;\alpha_0}(\xi)}{\partial\lambda_2})d\xi\right)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi \\
&\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} i\xi\overline{w}_{02}e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\frac{\partial\phi_{v;\alpha_0}(\xi)}{\partial\lambda_2}d\xi.
\end{aligned}$$

where $c_{\alpha_0}(\overline{w}) = 1$, and $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$.

When $\mathscr{C} = \mathbb{R}$, or the domain of C is a real line. Then, using the relationship in Eq. (A.14) we obtain

$$\text{(B.14)} \qquad f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{2\pi\phi_{v;\alpha}(0)}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda)}\phi_{v;\alpha}(\xi)d\xi.$$

Applying the relations $\frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c}dc = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi c}dc = \delta(\xi)$ and $\int f(\xi)\delta(\xi)dt = f(0)$ into Eqs. (B.11)-(B.13) yields

$$\begin{aligned}
\text{(B.15)} \qquad \frac{\partial}{\partial\theta_1} f_{C|\overline{W};\alpha_0} &= \frac{-1}{2\pi}\phi_v(0)\gamma_{\theta_{01}}(0;\theta_0)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)\gamma_{\theta_1}(\xi;\theta_0)d\xi, \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)\left(\gamma_{\theta_1}(\xi;\theta_0) - \gamma_{\theta_{01}}(0;\theta_0)\right)d\xi, \\
\frac{\partial}{\partial\lambda_1} f_{C|\overline{W};\alpha_0} &= \frac{-1}{2\pi}\frac{1}{\lambda_1}\phi_v(0)\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi + \frac{1}{2\pi}\int_{-\infty}^{\infty} i\xi\overline{w}_{01}e^{-i\xi(c-\overline{w}\lambda_0)}\phi_v(\xi)d\xi \\
&\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\left(\frac{1}{\lambda_1}\phi_v(\xi) + i\lambda_{01}\xi\gamma_{\lambda_1}(\xi;\alpha_0)\right)d\xi \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\left(i\xi\overline{w}_{01} + i\lambda_{01}\xi\gamma_{\lambda_1}(\xi;\alpha_0)\right)d\xi, \\
\frac{\partial}{\partial\lambda_2} f_{C|\overline{W};\alpha_0} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi(c-\overline{w}\lambda_0)}\left(i\xi\overline{w}_{02}\phi_v(\xi) + i\lambda_{01}\xi\gamma_{\lambda_2}(\xi;\alpha_0)\right)d\xi.
\end{aligned}$$

Therefore, the gradient of the log likelihood in Eq. (B.4) is expressed in terms of the population panel data model $f_{Y|X,C}(y|x,c;\theta_0)$ and the density of observables $f_{Y|X,\overline{W}}(y|x,\overline{w})$. The negative definiteness of the information matrix in Assumption 2.6 is equivalent to the positive definiteness of the outer product of the gradient of the log likelihood. The property of the positive definiteness can be imposed by choosing the outer product $\mathrm{E}\left[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0)\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0)'\middle|X=x,\overline{W}=\overline{w}\right]$ as a strictly diagonally

dominant matrix. To provide a more transparent condition, define the vector of the derivatives of the average likelihood as follows:

$$Df(y|x,\overline{w};\alpha_0) = \begin{pmatrix} \int_{\mathscr{C}} \frac{\partial}{\partial\theta_1} f_{Y|X,C}(y|x,c;\theta_0) f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc \\ \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0) \frac{\partial}{\partial\theta_1} f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc \\ \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0) \frac{\partial}{\partial\lambda_1} f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc \\ \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta_0) \frac{\partial}{\partial\lambda_2} f_{C|\overline{W}}(c|\overline{w};\alpha_0)dc \end{pmatrix}. \tag{B.16}$$

The detailed formulas of $\frac{\partial}{\partial\theta_1} f_{C|\overline{W}}(c|\overline{w};\alpha_0)$, $\frac{\partial}{\partial\lambda_1} f_{C|\overline{W}}(c|\overline{w};\alpha_0)$, and $\frac{\partial}{\partial\lambda_2} f_{C|\overline{W}}(c|\overline{w};\alpha_0)$ are in Eqs. (B.11)-(B.13) respectively.

Definition B.1. *A square matrix $A = [a_{ij}]$ is strictly diagonally dominant if*

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \quad \text{for all } i.$$

Assumption B.1. *(Strictly Diagonally Dominant)*
Assume that every element of $Df(y|x,\overline{w};\alpha_0)$ is non-zero and the outer product of the derivatives of the average likelihood

$$E\left[Df(y|x,\overline{w};\alpha_0)\cdot Df(y|x,\overline{w};\alpha_0)'\middle|X=x,\overline{W}=\overline{w}\right] \tag{B.17}$$

is strictly diagonally dominant.

Assumption B.1 implies that the expectation of the squared term of each derivative of the average likelihood is larger than the sum of the expectation of the magnitudes of the products of the derivative and other derivatives of the average likelihood. It is straightforward to verify that Assumption B.1 implies that every element of the gradient of the log likelihood $\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0)$ is non-zero and its outer product strictly diagonally dominant. Because every diagonal element of the matrix $\mathrm{E}\left[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0)\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha_0)'\middle|X=x,\overline{W}=\overline{w}\right]$ is positive, symmetric, and strictly diagonally dominant, the matrix is positive definite.[15] Thus, Assumption B.1 is intuitive and provides a sufficient condition for Assumption 2.6.

C. Sieve Maximum Likelihood Estimators

In this section, we give the expression of the limiting distribution of $\widehat{\theta}$ and give sufficient conditions. The results for λ are similar and we omit it.

[15] The result can be found as Corollary 6.1.10. in Horn and Johnson (1985).

Let $g = (\theta, \lambda, f_1)$ and $g_0 = (\theta_0, \lambda_0, f_V)$. Let $\mathscr{G}$ denote the space of g and $(\mathscr{G}, \|\cdot\|_{\mathscr{G}})$ be an infinite dimensional separable complete metric space. Let $Z = (Y, X, \overline{W})$ and $\psi(Z, g) = \log f_{Y|X,\overline{W}}(y|x, \overline{w}; \theta, \lambda, f_1)$. We have for all g, $\psi(Z, g) - \psi(Z, g_0)$ are approximated by $\Delta_\psi(Z, g_0)[g - g_0]$ such that $\Delta_\psi(Z, g_0)[g - g_0]$ is linear in $g - g_0$. Specifically, in our case, we have

$$\begin{aligned}\Delta_\psi(Z, g_0)[g - g_0] = \frac{1}{\int_{\mathscr{C}} f_{Y|X,C}(y|x, c; \theta_0) f_V(c - \overline{w}\lambda_0) dc} \Big\{ & \int_{\mathscr{C}} \nabla_\theta f_{Y|X,C}(y|x, c; \theta_0)[\theta - \theta_0] f_V(c - \overline{w}\lambda_0) dc \\ & - \int_{\mathscr{C}} f_{Y|X,C}(y|x, c; \theta_0) f_V'(c - \overline{w}\lambda_0) \overline{w}[\lambda - \lambda_0] dc \\ & + \int_{\mathscr{C}} f_{Y|X,C}(y|x, c; \theta_0) \big(f_1(c - \overline{w}\lambda_0) - f_V(c - \overline{w}\lambda_0)\big) dc \Big\}.\end{aligned}$$

For any $v_1, v_2 \in \mathscr{G} - g_0$, we define their inner product as

$$\langle v_1, v_2 \rangle_\psi = \frac{-dE[\Delta_\psi(Z, g_0 + \tau v_1)[v_2]]}{d\tau}\Big|_{\tau=0}.$$

The expression is tedious so we omit it. Note that we can use this inner product to define a norm on $\mathscr{G}$ which is $\|g_1 - g_2\|_\psi^2 = \langle (g_1 - g_2), (g_1 - g_2) \rangle_\psi$.

For any $g \in \mathscr{G}$ and for any $\gamma \in R^{d_\alpha}$, let $\rho_\gamma(g) = \gamma'\alpha$ so that $d\rho_\gamma(g_0)[v] = d\rho(g_0 + \tau v)/d\tau\big|_{\tau=0} = \gamma' v$, which is a linear functional of g. By the Riesz representation theorem, there is a sieve Riesz representer $v^* = (v_1^*, \ldots, v_{d_\alpha}^*) \in \mathscr{V}^{d_\alpha}$ where $\mathscr{V}$ is the closed linear span of $\mathscr{G} - g_o$ such that for all $v \in \mathscr{V}$

$$d\rho_\gamma(g_0)[v] = \langle \gamma' v^*, v \rangle_\psi,$$

and v_j^* satisfies that

$$d\rho_j(g_0)[v] = \langle v_j^*, v \rangle_\psi,$$

and $\rho_j(g) = \alpha_j$ for $j = 1, \ldots, d_\alpha$.

Define

$$\sigma_\gamma^2 = Var\Big(\gamma'\big(\Delta_\psi(Z, g_0)[v_1^*], \ldots, \Delta_\psi(Z, g_0)[v_{d_\alpha}^*]\big)\Big) = \gamma'\Omega\gamma,$$

where $\Omega = Var\big((\Delta_\psi(Z, g_0)[v_1^*], \ldots, \Delta_\psi(Z, g_0)[v_{d_\alpha}^*])\big)$. We have

$$\frac{\gamma\sqrt{N}(\widehat{\alpha} - \alpha_0)}{\sigma_\gamma^2} \xrightarrow{d} \mathscr{N}(0, 1). \tag{C.1}$$

Therefore, (C.1) implies that

$$\sqrt{N}(\widehat{\alpha}-\alpha_0) \xrightarrow{d} \mathcal{N}(0,\Omega), \tag{C.2}$$

where Ω is the asymptotic covariance matrix and $\Omega = Var\big((\Delta_\psi(Z,g_0)[v_1^*],\ldots,\Delta_\psi(Z,g_0)[v_{d_\alpha}^*])\big)$. (C.2) gives the result we need.

We derive the limiting distribution of $\sqrt{N}(\widehat{\alpha}-\alpha_0)$. We first give conditions so that the estimator is consistent. Let $\widehat{g} \equiv (\widehat{\alpha},\widehat{f}_1)$.

Assumption C.1. *For any $\epsilon > 0$, there exists a non-increasing positive sequence $c_N(\epsilon)$ such that for all $N \geq 1$, we have*

$$E[\psi(Z,g_o)] - \sup_{g\in\mathcal{G}_N:\|g-g_o\|_{\mathcal{G}}\geq\epsilon} E[\psi(Z,g)] \geq c_N(\epsilon) \tag{C.3}$$

and $\liminf c_N(\epsilon) > 0$.

Define $\mu_N(f) = N^{-1}\sum_{i=1}^N [f_(Z_i) - E[f(Z_i)]]$ which denotes the empirical process indexed by f.

Assumption C.2. *Assume that (i) for all N, there exists some $g_N \in \mathcal{G}_N$ such that*

$$|E[\psi(Z,g_N)] - E[\psi(Z,g_o)]| = o(1); \tag{C.4}$$

(ii) $\sup_{g\in\mathcal{G}_N} |\mu_N(\psi(Z,g))| = o_P(1)$.

Theorem C.1. *Suppose Assumptions C.1 and C.2 hold. Then $\|\widehat{g}-g_o\| = o_p(1)$.*

Let C denote a positive finite number. Define $\mathcal{N}_{g,C} = \{g \in \mathcal{G} : \|g-g_o\| \leq C\}$ and $\mathcal{N}_{g,N,C} = \{g \in \mathcal{G}_N : \|g-g_o\| \leq C\}$. By Theorem C.1, we have $\widehat{g} \in \mathcal{N}_{g,N,C}$ with probability approaching 1. Next, we give conditions to derive the convergence rate of $\widehat{g}$.

Assumption C.3. *There exist some finite, positive and non-increasing sequences δ_{1N}, δ_{2N}, and δ_{3N} that are all $o(1)$ such that (i)*

$$\sup_{g\in\mathcal{N}_{g,N,C}} \Big|\mu_N(\psi(Z,g)-\psi(Z,g_o))\Big| = O_p(\delta_{1N}^2); \tag{C.5}$$

(ii) for all N large enough and for any $\delta > 0$ small enough,

$$E\Big[\sup_{g\in\mathcal{N}_{g,N,C}:\|g-g_o\|_{\mathcal{G}}\leq\delta} \Big|\mu_N(\psi(Z,g)-\psi(Z,g_o))\Big|\Big] \leq \frac{c_1\phi_N(\delta)}{\sqrt{N}} \tag{C.6}$$

where c_1 is some positive number and $\phi_N(\cdot)$ is some function such that $\delta^{\nu}\psi_N(\delta)$ is a decreasing function for some $\nu \in (0,2)$;

(iii) $\delta_{2N}^{-2}\phi(\delta_{2N}) \leq c_2\sqrt{N}$ for some finite positive $c_2 > 0$; (iv) $\|g_N - g_o\|_{\mathscr{G}} = O(\delta_{3N})$.

Theorem C.2. *Assumptions C.1, C.2 and C.3 hold. Then we have $\|\widehat{g} - g_o\| = O_p(\epsilon_N^*)$ where $\epsilon_N^* = \max\{\delta_{1N}, \delta_{2N}, \delta_{3N}\} = o(1)$.*

Assumption C.4. *Assume that there exists $c_\psi < \infty$ such that $\|v\|_\psi \leq c_\psi \|v\|_{\mathscr{G}}$.*

Let $\epsilon_N = \epsilon_N^* \cdot \delta_N$ with $\delta_N \to \infty$ such that ϵ_N remains $o_p(1)$. Let $\mathscr{G}_N \equiv \mathscr{A} \times \mathscr{F}_{1N}$. Let $\mathscr{N}_g = \{(g \in \mathscr{G} : \|g - g_o\|_{\mathscr{G}} \leq \epsilon_N)\}$ and $\mathscr{N}_{g,N} = \mathscr{N}_g \cap \mathscr{G}_N$. It is true that $\widehat{g} \in \mathscr{N}_{g,N}$ with probability approaching one. Let Π_N denote the projection of g on $\mathscr{G}_N$ under the norm $\|\cdot\|_\psi$ and let $g_{o,N} = \Pi_N g_o$. Assumption C.4 implies that $g_{o,N} \in \mathscr{N}_{g,N}$.

Π_N is also the projection of v on $\mathscr{V}_N$ which is the closed linear span of $\mathscr{G}_N - g_{o,N}$. Let $v_N^* \equiv (\Pi_N \cdot v_1^*, \ldots, \Pi_N \cdot v_{d_\alpha}^*) \in \mathscr{V}_N^{d_\alpha}$. By construction, we have for all $v \in \mathscr{V}_N$

$$d\rho_\gamma(g_0)[v] = \langle \gamma' v_N^*, v \rangle_\psi.$$

Let κ_N denote a sequence of positive numbers and $\kappa_N = o(N^{-1/2})$. For any g, let $g_\lambda^* = g \pm \kappa_N \cdot \gamma' v_N^*$.

Assumption C.5. *Assume that (i)*

$$\sup_{g \in \mathscr{N}_{g,N}} \left| \mu_N\left(\psi(Z, g^*) - \psi(Z, g) - \Delta_\psi(Z, g)[\pm\kappa_N \cdot \gamma' v_N^*]\right)\right| = O_p(\kappa_N^2), \tag{C.7}$$

$$\sup_{g \in \mathscr{N}_{g,N}} \left| \mu_N\left(\Delta_\psi(Z, g)[\gamma' v_N^*] - \Delta_\psi(Z, g_o)[\gamma' v_N^*]\right)\right| = O_p(\kappa_N); \tag{C.8}$$

(ii) let $K_\psi(g) \equiv E[\psi(Z, g) - \psi(Z, g_o)]$, then

$$\sup_{g \in \mathscr{N}_{g,N}} \left| K_\psi(g) - K_\psi(g_o) - \frac{\|g^* - g_o\|_\psi^2 - \|g - g_o\|_\psi^2}{2}\right| = O_p(\kappa_N^2); \tag{C.9}$$

(iii) $\kappa_N/\epsilon_N^* = o(1)$ and (iv) $\|v_j^*\|_\psi < \infty$ for all $j = 1, \ldots, d_\alpha$.

Theorem C.3. *Suppose Assumptions C.1, C.2, C.3, C.4 and C.5 hold. Then*

$$\lambda\sqrt{N}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathscr{N}(0, \sigma_\gamma^2) = \mathscr{N}(0, \gamma'\Omega\gamma).$$

The proofs of Theorems, C.1, C.2 and C.3 follow the same arguments as Chen, Liao, and Sun (2014) and Hahn, Liao, and Ridder (2018), so we omit the details. By the Cramér-Wold theorem, Theorem C.3 implies that $\sqrt{N}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathscr{N}(0, \Omega)$.

To make inference, one would need a consistent estimator for Ω. The estimator is a sample analog. To be specific, for $j = 1, \ldots, d_\alpha$, define the empirical Riesz representer $\widehat{v}_j^*$ as

$$d\rho_j(\widehat{g})[v] = \langle \widehat{v}_j^*, v \rangle_{N,\psi}, \text{ where}$$
$$\langle v_1, v_2 \rangle_{N,\psi} = \frac{1}{N} \sum_{i=1}^{N} - \frac{d\Delta_\psi(Z_i, \widehat{g} + \tau v_1)[v_2]}{d\tau}\Big|_{\tau=0}.$$

A consistent estimator for $\widehat{\Omega}$ is given as

$$\widehat{\Omega} = \frac{1}{N} \sum_{i=1}^{N} \left(\Delta_\psi(Z_i, \widehat{g})[\widehat{v}_1^*], \ldots, \Delta_\psi(Z_i, \widehat{g})[\widehat{v}_{d_\alpha}^*]\right)' \left(\Delta_\psi(Z_i, \widehat{g})[\widehat{v}_1^*], \ldots, \Delta_\psi(Z_i, \widehat{g})[\widehat{v}_{d_\alpha}^*]\right). \tag{C.10}$$

The consistency of $\widehat{\Omega}$ can be proved by Theorem 4.1 of Hahn, Liao, and Ridder (2018). More importantly, even if (C.10) looks complicated, one can apply Theorem 6.1 and Remark 6.1 of Hahn, Liao, and Ridder (2018) to use the variance estimator of the parametric MLE model when the sieve space is generated by a finite number of basis functions. To be specific, suppose that f_1 is approximated by $f_1(v;\beta)$ where β is of $K(N)$ dimensions. Then we write

$$f_{Y|X,\overline{W}}(y|x,\overline{w};\alpha,\beta) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_1(c - \overline{w}\lambda;\beta)) dc.$$

and the sieve estimator is given by

$$(\widehat{\alpha}, \widehat{\beta}) \equiv \arg\max_{\alpha,\beta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_{Y|X,\overline{W}}(Y_i|X_i, \overline{W}_i;\theta,\lambda,\beta)\right), \text{ and}$$
$$\widehat{f}_1(v) = f_1(v;\widehat{\beta}).$$

Let $\widehat{s}_i = \nabla \log\left(f_{Y|X,\overline{W}}(Y_i|X_i, \overline{W}_i;\alpha,\beta)\right)$ and $\widehat{H}_i = \nabla^2 \log\left(f_{Y|X,\overline{W}}(Y_i|X_i, \overline{W}_i;\alpha,\beta)\right)$ be the gradient and the Hessian of $\log\left(f_{Y|X,\overline{W}}(Y_i|X_i, \overline{W}_i;\alpha,\beta)\right)$ evaluated at $(\widehat{\alpha}, \widehat{\beta})$. Let $\widehat{V}$ be the variance matrix estimator of MLE estimator that is given by

$$\widehat{V} = \left(\frac{1}{N} \sum_{i=1}^{N} -\widehat{H}_i\right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{s}_i \widehat{s}_i'\right) \left(\frac{1}{N} \sum_{i=1}^{N} -\widehat{H}_i\right)^{-1}.$$

Then the expression of $\widehat{\Omega}$ will be the upper-left $d_\alpha \times d_\alpha$ matrix of $\widehat{V}$. Equivalently, let $\widehat{s}_{\alpha,i}$ be the first d_α row of the vector of $\widehat{\mathscr{H}}^{-1} \cdot \widehat{s}_i$ with $\widehat{\mathscr{H}} = N^{-1} \sum_{i=1}^{N} -\widehat{H}_i$, so $\widehat{\Omega}$ is given by $N^{-1} \sum_{i=1}^{N} \widehat{s}_{\alpha,i} \widehat{s}_{\alpha,i}'$.

D. A Consistent Estimator for Ω_{di}

In this section, we give a consistent estimator for Ω_{di}. Recall that the parameter MLE estimator given in Section 4.1 is

$$(\widehat{\alpha}_{pa}, \widehat{\tau}_{pa}) \equiv \arg\max_{\alpha \in \mathscr{A}, \tau \in \mathscr{T}} \frac{1}{N} \sum_{i=1}^{N} f_{pa}(Y_i | X_i, \overline{W}_i; \alpha, \tau),$$
$$f_{pa}(y | x, \overline{w}; \alpha, \tau) = \int_{\mathscr{C}} f_{Y|X,C}(y | x, c; \theta) f_V(c - \overline{W}\lambda; \tau) dc.$$

Let $\widehat{s}_{pa,i} = \nabla \log\left(f_{pa}(Y_i | X_i, \overline{W}_i; \alpha, \beta)\right)$ and $\widehat{H}_{pa,i} = \nabla^2 \log\left(f_{pa}(Y_i | X_i, \overline{W}_i; \alpha, \beta)\right)$ be the gradient and the Hessian of $\log\left(f_{pa}(Y_i | X_i, \overline{W}_i; \alpha, \beta)\right)$ evaluated at $(\widehat{\alpha}_{pa}, \widehat{\tau}_{pa})$. Let $\widehat{V}_{pa}$ be the variance matrix estimator of MLE estimator that is given by

$$\widehat{V}_{pa} = \left(\frac{1}{N} \sum_{i=1}^{N} -\widehat{H}_{pa,i}\right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{s}_{pa,i} \widehat{s}'_{pa,i}\right) \left(\frac{1}{N} \sum_{i=1}^{N} -\widehat{H}_{pa,i}\right)^{-1}.$$

Then the expression of $\widehat{\Omega}_{pa}$ will be the upper-left $d_\alpha \times d_\alpha$ matrix of $\widehat{V}_{pa}$. Equivalently, let $\widehat{s}_{pa,\alpha,i}$ be the first d_α row of the vector of $\widehat{\mathscr{H}}_{pa}^{-1} \cdot \widehat{s}_{pa,i}$ with $\widehat{\mathscr{H}}_{pa} = N^{-1} \sum_{i=1}^{N} -\widehat{H}_{pa,i}$, so $\widehat{\Omega}_{pa}$ is given by $N^{-1} \sum_{i=1}^{N} \widehat{s}_{pa,\alpha,i} \widehat{s}'_{pa,\alpha,i}$.

Given these results, a consistent estimator for Ω_{di} is given by

$$\widehat{\Omega}_{di} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{s}_{pa,\alpha,i} - \widehat{s}_{\alpha,i}\right) \cdot \left(\widehat{s}_{pa,\alpha,i} - \widehat{s}_{\alpha,i}\right)'.$$

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.

BONHOMME, S. (2012): "Functional Differencing," *Econometrica*, 80(4), 1337–1385.

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, S., J. SI, H. ZHANG, AND Y. ZHOU (2017): "Root-N Consistent Estimation of a Panel Data Binary Response Model With Unknown Correlated Random Effects," *Journal of Business & Economic Statistics, forthcoming*.

CHEN, X., Z. LIAO, AND Y. SUN (2014): "Sieve Inference on Possibly Misspecified Semi-nonparametric Time Series Models," *Journal of Econometrics*, 178(1), 639–658.

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*.

FOLLAND, G. B. (2009): *Fourier Analysis and its Applications*, vol. 4. American Mathematical Soc.

GAYLE, G.-L., AND C. VIAUROUX (2007): "Root-N Consistent Semiparametric Estimators of a Dynamic Panel-sample-selection Model," *Journal of Econometrics*, 141(1), 179–212.

GAYLE, W.-R. (2013): "Identification and $\sqrt{N}$-consistent Estimation of a Nonlinear Panel Data Model with Correlated Unobserved Effects," *Journal of Econometrics*, 175(2), 71–83.

GAYLE, W.-R., AND S. D. NAMORO (2013): "Estimation of a Nonlinear Panel Data Model with Semiparametric Individual Effects," *Journal of Econometrics*, 175(1), 46–59.

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.

HAHN, J., Z. LIAO, AND G. RIDDER (2018): "Nonparametric Two-step Sieve M Estimation and Inference," *Econometric Theory, forthcoming*.

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.

HORN, R., AND C. JOHNSON (1985): *Matrix Analysis*. Cambridge University Press.

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 Binary Models

	Infeasible		Conventional			Sieve ML	
N=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	-0.498	0.496	-0.501	0.499	0.916	-0.509	0.418
Std. dev.	0.052	0.053	0.070	0.079	0.116	0.139	0.152
RMSE	0.052	0.053	0.070	0.079	0.143	0.139	0.172
DGP II:							
Mean	-0.502	0.493	-0.609	0.195	1.868	-0.506	0.488
Std. dev.	0.059	0.062	0.088	0.105	0.244	0.119	0.121
RMSE	0.059	0.062	0.140	0.322	0.901	0.119	0.121
DGP III:							
Mean	-0.498	0.494	-0.595	0.229	1.761	-0.497	0.484
Std. dev.	0.059	0.061	0.085	0.101	0.200	0.128	0.130
RMSE	0.059	0.061	0.127	0.289	0.787	0.127	0.131

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

Table 2: Simulations of Static Binary Models

	Infeasible		Conventional			Sieve ML	
N=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	-0.496	0.504	-0.496	0.507	0.919	-0.512	0.523
Std. dev.	0.036	0.043	0.046	0.060	0.083	0.094	0.107
RMSE	0.036	0.043	0.046	0.061	0.116	0.094	0.109
DGP II:							
Mean	-0.498	0.503	-0.601	0.207	1.843	-0.504	0.523
Std. dev.	0.040	0.042	0.064	0.069	0.160	0.095	0.097
RMSE	0.040	0.042	0.119	0.301	0.858	0.094	0.099
DGP III:							
Mean	-0.494	0.503	-0.582	0.235	1.736	-0.515	0.531
Std. dev.	0.041	0.042	0.058	0.068	0.141	0.095	0.112
RMSE	0.041	0.042	0.100	0.273	0.749	0.096	0.116

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 Binary Models

N=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	-0.168	-0.143	-0.158
Std. dev.	0.001	0.017	0.040
RMSE	–	0.017	0.041
DGP II:			
Mean	-0.181	-0.113	-0.159
Std. dev.	0.001	0.015	0.037
RMSE	–	0.026	0.043
DGP III:			
Mean	-0.187	-0.116	-0.155
Std. dev.	0.001	0.014	0.038
RMSE	–	0.024	0.049

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 Binary Models

N=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	-0.168	-0.141	-0.151
Std. dev.	0.001	0.011	0.027
RMSE	–	0.011	0.032
DGP II:			
Mean	-0.181	-0.113	-0.146
Std. dev.	0.001	0.011	0.027
RMSE	–	0.023	0.045
DGP III:			
Mean	-0.188	-0.114	-0.152
Std. dev.	0.001	0.011	0.028
RMSE	–	0.021	0.046

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

Table 5: Simulations of Dynamic Binary Models

	Infeasible			Conventional				Sieve ML		
N=500	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP I:										
Mean	-0.503	0.504	0.494	-0.493	0.493	0.478	0.826	-0.505	0.511	0.515
Std. dev.	0.108	0.062	0.092	0.122	0.096	0.106	0.362	0.103	0.112	0.117
RMSE	0.107	0.062	0.092	0.121	0.096	0.108	0.401	0.103	0.113	0.117
DGP II:										
Mean	-0.506	0.503	0.485	-0.639	0.395	0.242	1.276	-0.500	0.514	0.524
Std. dev.	0.127	0.067	0.096	0.119	0.087	0.106	0.285	0.103	0.115	0.119
RMSE	0.127	0.067	0.097	0.183	0.136	0.279	0.396	0.103	0.115	0.121
DGP III:										
Mean	-0.505	0.502	0.486	-0.627	0.407	0.255	1.295	-0.500	0.513	0.525
Std. dev.	0.117	0.069	0.092	0.125	0.088	0.101	0.297	0.104	0.114	0.120
RMSE	0.116	0.069	0.093	0.178	0.128	0.265	0.418	0.103	0.115	0.122

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

Table 6: Simulations of Dynamic Binary Models

	Infeasible			Conventional				Sieve ML		
N=1000	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP I:										
Mean	-0.496	0.496	0.499	-0.473	0.492	0.476	0.822	-0.514	0.508	0.511
Std. dev.	0.093	0.052	0.067	0.084	0.082	0.084	0.339	0.095	0.094	0.102
RMSE	0.093	0.052	0.067	0.088	0.082	0.088	0.382	0.096	0.094	0.103
DGP II:										
Mean	-0.491	0.496	0.497	-0.621	0.401	0.250	1.299	-0.510	0.509	0.520
Std. dev.	0.097	0.052	0.066	0.089	0.064	0.074	0.260	0.095	0.095	0.105
RMSE	0.097	0.052	0.066	0.150	0.118	0.261	0.396	0.096	0.095	0.107
DGP III:										
Mean	-0.496	0.496	0.499	-0.612	0.403	0.264	1.277	-0.511	0.509	0.519
Std. dev.	0.093	0.052	0.067	0.092	0.063	0.075	0.249	0.095	0.095	0.106
RMSE	0.093	0.052	0.067	0.145	0.116	0.248	0.372	0.095	0.095	0.107

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

Table 7: Simulation of the State Dependence in Dynamic Binary Models

N=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	-0.134	-0.136	-0.138
Std. dev.	0.003	0.035	0.030
RMSE	–	0.035	0.030
DGP II:			
Mean	-0.112	-0.156	-0.136
Std. dev.	0.005	0.035	0.030
RMSE	–	0.056	0.039
DGP III:			
Mean	-0.111	-0.152	-0.137
Std. dev.	0.004	0.037	0.030
RMSE	–	0.055	0.040

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 Binary Models

N=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP I:			
Mean	-0.134	-0.132	-0.141
Std. dev.	0.002	0.026	0.028
RMSE	–	0.026	0.029
DGP II:			
Mean	-0.112	-0.150	-0.140
Std. dev.	0.004	0.027	0.028
RMSE	–	0.047	0.040
DGP III:			
Mean	-0.110	-0.149	-0.141
Std. dev.	0.003	0.027	0.028
RMSE	–	0.047	0.029

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

Table 9: Hausman-type Test for Normality: Empirical Size

	Static Binary Models		Dynamic Binary Models	
	N=500	N=1000	N=500	N=1000
DGP I:	0.080	0.047	0.060	0.027
DGP II:	0.360	0.793	0.480	0.587
DGP III:	0.267	0.587	0.460	0.600

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 10: Simulations of Two-Period Dynamic Count Models

	Infeasible			Conventional				Sieve ML		
N=500	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP IV:										
Mean	-0.508	0.506	0.501	-0.385	0.392	0.205	0.316	-0.555	0.454	0.509
Std. dev.	0.034	0.070	0.059	0.043	0.067	0.053	0.168	0.139	0.169	0.198
RMSE	0.035	0.070	0.059	0.122	0.127	0.299	0.704	0.149	0.174	0.198
DGP V:										
Mean	-0.505	0.487	0.498	-0.381	0.385	0.198	0.308	-0.551	0.467	0.482
Std. dev.	0.033	0.066	0.058	0.040	0.062	0.048	0.169	0.129	0.114	0.182
RMSE	0.033	0.067	0.058	0.126	0.130	0.306	0.712	0.138	0.118	0.182
DGP VI:										
Mean	-0.502	0.493	0.505	-0.372	0.365	0.163	0.308	-0.551	0.462	0.532
Std. dev.	0.035	0.065	0.063	0.040	0.068	0.050	0.198	0.162	0.183	0.224
RMSE	0.035	0.065	0.063	0.134	0.151	0.341	0.720	0.169	0.187	0.226

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

Table 11: Simulations of Two-Period Dynamic Count Models

	Infeasible			Conventional				Sieve ML		
N=500	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP IV:										
Mean	-0.502	0.509	0.503	-0.381	0.377	0.200	0.325	-0.523	0.502	0.519
Std. dev.	0.025	0.045	0.040	0.027	0.048	0.035	0.146	0.121	0.099	0.105
RMSE	0.025	0.046	0.040	0.122	0.132	0.302	0.691	0.123	0.099	0.106
DGP V:										
Mean	-0.500	0.502	0.500	-0.381	0.381	0.198	0.329	-0.502	0.504	0.515
Std. dev.	0.024	0.044	0.044	0.029	0.044	0.034	0.140	0.097	0.091	0.109
RMSE	0.024	0.044	0.044	0.123	0.127	0.304	0.685	0.096	0.091	0.110
DGP VI:										
Mean	-0.504	0.502	0.502	-0.374	0.356	0.162	0.351	-0.504	0.498	0.510
Std. dev.	0.025	0.046	0.044	0.029	0.049	0.036	0.138	0.113	0.096	0.107
RMSE	0.025	0.046	0.044	0.129	0.152	0.340	0.664	0.113	0.096	0.107

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

Table 12: Simulation of the APE($\bar{x}$) in Two-Period Dynamic Count Models

N=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP IV:			
Mean	-0.127	-0.138	-0.161
Std. dev.	0.003	0.014	0.040
RMSE	–	0.019	0.053
DGP V:			
Mean	-0.127	-0.137	-0.161
Std. dev.	0.003	0.013	0.039
RMSE	–	0.019	0.052
DGP VI:			
Mean	-0.117	-0.136	-0.159
Std. dev.	0.003	0.009	0.046
RMSE	–	0.023	0.062

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

Table 13: Simulation of the APE($\bar{x}$) in Two-Period Dynamic Count Models

N=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP IV:			
Mean	-0.117	-0.137	-0.145
Std. dev.	0.002	0.009	0.034
RMSE	–	0.016	0.038
DGP V:			
Mean	-0.126	-0.137	-0.139
Std. dev.	0.002	0.009	0.027
RMSE	–	0.016	0.030
DGP VI:			
Mean	-0.127	-0.136	-0.140
Std. dev.	0.002	0.009	0.032
RMSE	–	0.023	0.039

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

Table 14: Simulations of Four-Period Dynamic Count Models

	Infeasible			Conventional				Sieve ML		
N=500	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP IV:										
Mean	-0.501	0.499	0.500	-0.523	0.491	0.501	0.996	-0.418	0.432	0.348
Std. dev.	0.021	0.023	0.019	0.028	0.036	0.040	0.047	0.105	0.144	0.131
RMSE	0.021	0.023	0.019	0.036	0.037	0.040	0.047	0.133	0.159	0.201
DGP V:										
Mean	-0.500	0.502	0.498	-0.526	0.495	0.501	1.016	-0.420	0.459	0.399
Std. dev.	0.020	0.020	0.017	0.023	0.041	0.035	0.047	0.087	0.116	0.122
RMSE	0.020	0.020	0.018	0.035	0.041	0.035	0.049	0.118	0.123	0.158
DGP VI:										
Mean	-0.498	0.499	0.498	-0.525	0.480	0.501	1.197	-0.409	0.438	0.343
Std. dev.	0.027	0.025	0.017	0.029	0.058	0.037	0.048	0.108	0.138	0.140
RMSE	0.027	0.025	0.017	0.038	0.061	0.037	0.203	0.141	0.151	0.210

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

Table 15: Simulations of Four-Period Dynamic Count Models

	Infeasible			Conventional				Sieve ML		
N=1000	γ	θ	λ	γ	θ	λ	σ	γ	θ	λ
True	-0.5	0.5	0.5	-0.5	0.5	0.5	1	-0.5	0.5	0.5
DGP IV:										
Mean	-0.501	0.500	0.501	-0.524	0.491	0.500	1.007	-0.502	0.481	0.469
Std. dev.	0.015	0.016	0.012	0.018	0.025	0.029	0.032	0.078	0.086	0.090
RMSE	0.015	0.016	0.012	0.030	0.026	0.029	0.032	0.077	0.088	0.095
DGP V:										
Mean	-0.500	0.499	0.502	-0.526	0.493	0.501	1.018	-0.478	0.511	0.496
Std. dev.	0.015	0.014	0.015	0.019	0.026	0.027	0.034	0.085	0.109	0.095
RMSE	0.015	0.014	0.015	0.032	0.027	0.027	0.038	0.087	0.109	0.094
DGP VI:										
Mean	-0.502	0.499	0.501	-0.527	0.485	0.497	1.208	-0.485	0.454	0.434
Std. dev.	0.013	0.017	0.011	0.018	0.032	0.033	0.034	0.084	0.088	0.093
RMSE	0.013	0.016	0.011	0.032	0.036	0.033	0.210	0.085	0.099	0.114

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

Table 16: Simulation of the State Dependence in Four-Period Dynamic Count Models

N=500	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP IV:			
Mean	-0.130	-0.133	-0.119
Std. dev.	0.003	0.008	0.030
RMSE	–	0.010	0.031
DGP V:			
Mean	-0.129	-0.133	-0.120
Std. dev.	0.003	0.007	0.025
RMSE	–	0.009	0.027
DGP VI:			
Mean	-0.119	-0.123	-0.117
Std. dev.	0.003	0.007	0.030
RMSE	–	0.009	0.030

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

Table 17: Simulation of the State Dependence in Four-Period Dynamic Count Models

N=1000	Infeasible Estimator	Conventional Estimator	Sieve ML Estimator
DGP IV:			
Mean	-0.130	-0.132	-0.142
Std. dev.	0.002	0.005	0.022
RMSE	–	0.007	0.025
DGP V:			
Mean	-0.129	-0.132	-0.134
Std. dev.	0.002	0.005	0.024
RMSE	–	0.008	0.024
DGP VI:			
Mean	-0.119	-0.123	-0.137
Std. dev.	0.002	0.005	0.023
RMSE	–	0.008	0.029

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

Table 18: Hausman-type Test for Normality: Empirical Size

	Two-Period Dynamic Count Models		Four-Period Dynamic Count Models	
	N=500	N=1000	N=500	N=1000
DGP IV:	0.253	0.780	0.213	0.060
DGP V:	0.407	0.800	0.260	0.067
DGP VI:	0.387	0.800	0.247	0.087

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.