Econometric Theory, **37**, 2021, 1075–1099.
doi:10.1017/S0266466620000481

NONLINEAR PANEL DATA MODELS WITH DISTRIBUTION-FREE CORRELATED RANDOM EFFECTS*

YU-CHIN HSU
Institute of Economics, Academia Sinica
National Central University
National Chengchi University

JI-LIANG SHIU
Jinan University

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 a Fourier transformation without imposing a distributional assumption on the CRE specification. We subsequently construct a semiparametric 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 CRE specification are identifiable. 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 the CRE. Furthermore, we show that the average partial effects are identifiable and extend our results to dynamic nonlinear panel data models.

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, particularly because of the fundamental difference between linear and nonlinear models. For linear panel data models with additive unobserved heterogeneity, under the fixed

*The authors are grateful to the co-editor, three anonymous referees, Arthur Lewbel, and Matthew Shum for valuable comments and suggestions on previous versions of the paper. The authors are indebted to the editor Peter Phillips for constructive advice and comments, which have considerably improved the presentation of the paper. 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. Ji-Liang Shiu acknowledges support from the China National Science Foundation (Project No. 72073050). Address correspondence to Ji-Liang Shiu, Institute for Economic and Social Research, Jinan University, Guangzhou, China; e-mail: jishiu.econ@gmail.com.

© Cambridge University Press 2021.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

effects formulation, one can apply within transformation to eliminate the unobserved effects, and consistent estimators can be obtained by the generalized method of moments approaches without specifying the distribution of the unobserved heterogeneity. One may 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. Other than this, there are two main approaches in the literature. One treats the unobserved heterogeneity of each individual as a fixed parameter (fixed effect), and the other treats 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.[2] 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.[3]

For the random effects approach, one can first specify the conditional distribution of a parametric nonlinear panel data model, that is, a parametric conditional distribution of the dependent variable, Y_t, conditional on a vector of time-varying explanatory variables, X_t, and individual unobserved heterogeneity, C:

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

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 effect 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. We can then construct an average

[1] Some progress has been made in this direction, including by 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 et al. (2013), Browning and Carro (2014), and Chernozhukov et al. (2015).

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

[3] As discussed in Arellano and Bonhomme (2011), the identification problem is related to situations where information on outcomes is insufficient to identify the unobserved heterogeneity. In the binary response model, the support of outcomes is less rich than the support of the unobserved heterogeneity.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

likelihood of Y given X as

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

The average likelihood in equation (2) is fully parametric in that it depends on a finite number of parameters, and estimation and inference are possible under a standard maximum likelihood estimation (MLE) framework. For example, Chamberlain (1980) proposes this method for static models. With equation (1) and a slightly modified equation (2), one can allow for dynamic models in which X_t can include lagged dependent variables.[4] 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 that 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 effect approach is that the misspecification of $f_{C|X}(c|x;\beta)$ generally results in inconsistent estimates.[5] To avoid this, one can alternatively characterize the identified set of true parameters, but one would lose point identification. For example, Honoré and Tamer (2006) relax the distributional assumption of the initial condition and calculate bounds on parameters in panel dynamic discrete choice models. Chernozhukov et al. (2013) show that bounds for marginal effects in nonlinear panel models can be tightened rapidly as the number of time-series observations grow. Bonhomme (2012) proposes a functional differencing approach to difference out the individual effects and provides restrictions on common parameters alone to obtain consistent estimators when T is fixed. As a result, except for Bonhomme (2012), a general method not subject to the distribution misspecification of the average likelihood in equation (2) for fixed T that also retains point identification of the parameters remains elusive.[6]

To fill this gap in the literature, we provide a correlated random effect (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 equation (1), and we can retain point identification of the parameters. To be more specific, let $\overline{W}$ be a vector of

[4] Please see Section 4.2 for dynamic models.

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

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

time-invariant observed variables.[7] We consider a Mundlak-type specification such that $C = \lambda \overline{W} + V$, so under Assumption 2.2 in the static model, equation (2) can be written as

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

where f_V, the density function of V, is unspecified. Therefore, the average likelihood is the convolution of the conditional distribution of the parametric nonlinear panel data models and that of the unobserved heterogeneity.[8] 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.

We next extend the relation of the Fourier transform to parameters other than the true parameters 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 construct a correctly specified semiparametric average likelihood of observables in that it is equal to the density of observables when evaluated at the true parameters, (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.

We then show that its parameter vector is identifiable by the negative definiteness of the information matrix, which is a standard maximum likelihood condition. Based on the identification result, we propose a sieve maximum likelihood (henceforth sieve ML) estimator of the parameters, (θ_0, λ_0), that is consistent and asymptotically normal.[9] We then propose a Hausman-type test on the parametric specification of f_V in the conventional parametric CRE models. Our method can also identify and estimate the distribution of unobserved effects and in turn, the average partial effects. Note that these two objects that may be of separate interest are sometimes not identified in Bonhomme (2012).[10]

[7] Note that the form of $\overline{W}$ may depend on whether the model is static or dynamic.

[8] Our (3) is similar to Chamberlain (1980) except that we require that $C = \lambda \overline{W} + V$ and λ is not a zero vector.

[9] Note that our estimation and inference results in CRE models can be thought of special cases of the general form of the random effects model in Kiefer and Wolfowitz (1956). However, Kiefer and Wolfowitz (1956) assume the identification condition in their Assumption 4. Kiefer and Wolfowitz (1956) provide examples with conditions such that Assumption 4 holds in these examples, but it is still not straightforward to verify whether Assumption 4 holds in general examples. Laird (1978) and Aitkin (1996) also focus on the estimation of semiparametric mixture models without discussions of identification. A short review of this branch of literature is given in Lindsay and Lesperance (1995). In this sense, we contribute to the literature by providing a set of sufficient conditions such that the identification assumption would hold in a set of CRE models. We thank a referee for bringing this literature to our attention.

[10] Furthermore, the method of Bonhomme (2012) mainly applies to likelihood models with continuous dependent variables, but our method applies to both continuous and discrete dependent variables.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

The key insight of our approach is to utilize the information on 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 is 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; and Hu and Ridder, 2012), panel data models (see Evdokimov, 2011; Arellano and Bonhomme, 2012), etc. Among them, Evdokimov (2011) establishes a 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 obtain an identification of the density of individual effects.[11]

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 et al. (2017). Specifically, Gayle (2013) investigates the 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 provide an identification result of average partial effects.

The remainder of the paper is organized as follows. Section 2 presents the identification results of an internally consistent likelihood function for static models. Section 3 proposes 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. Section 5 concludes the paper. Part of the technical proofs of the results are in the Appendix. We provide an online supplement that includes Monte Carlo simulation results and additional technical proofs.

[11] There are other papers on identification based on other methods. For example, Heckman and Singer (1984) provide identifiability conditions for the Cox proportional hazard model for duration data. Their identification conditions require a restriction on the tail behavior of the true distribution of unobservables, and the identification results are obtained by Laplace transforms. Briesch, Chintagunta, and Matzkin (2010) provide the nonparametric identification and estimation of discrete choice models with unobserved heterogeneity. They impose nonparametric assumptions on the systematic subutility functions and on the distributions of the unobservable random vectors and the heterogeneity parameter to obtain the identification following the method in Matzkin (2003).

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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 $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_T$. Additionally, let C denote an individual unobserved heterogeneity C with support $\mathscr{C}$. We consider the following parametric panel data model:

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

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

The panel data model in equation (4) may be derived from a more primitive econometric model. Suppose that $m(x, c; \theta)$ is a known parametric function, and consider the following binary choice model:

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

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:

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

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 in the functional form $f_{Y_t|X_t,C}$ as

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

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 equation (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 equation (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 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 $\theta \neq \theta_0$ implies $f_{Y|X,C}(y|x, c; \theta) \neq f_{Y|X,C}(y|x, c; \theta_0)$ and $f_{Y_t|X_t,C}(y_t \,|\, x_t, c; \theta_0)$ 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 $\mathbb{R}^{d_\theta}$.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

To control the possible correlation between X_t and C, we use a CREs 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 $\overline{W} = 1/T + 1 \sum_{t=0}^{T}(X_{t1}, ..., X_{tK_1}) = (\overline{X}_1, ..., \overline{X}_{K_1})$ as the time average of the first K_1 explanatory variables of X_t for $t = 0, 1, \ldots, T$.[12]

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 nonzero, where Λ is a compact subset of R^{K_1} such that

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

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; that is, 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 equation (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 that in Wooldridge (2005), because we do not impose a parametric assumption on V; however, ours is less general than that in Wooldridge (2005), because we 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 nonzero 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 unobserved individual 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 $(y_t, x_t, \overline{w}, c) \in \mathcal{Y} \times \mathcal{X} \times \overline{\mathcal{W}} \times \mathcal{C}$;

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

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

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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 controlling for the regressors X_t and C. 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. We 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 $\lambda_{01} \neq 0$. Denote $w_{-1} = (w_2, w_3, \ldots, w_{K_1})$ and $w_{-1} \in \overline{W}_{-1} \equiv \overline{\mathscr{W}}_2 \times \cdots \times \overline{\mathscr{W}}_{K_1}$. 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 $\lambda \in \Lambda$ as well as $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc$ for all $(y,x) \in \mathscr{Y} \times \mathscr{X}$ and $\theta \in \Theta$ are all finite. We note that $f_{Y|X,C}(y|x,c;\theta)$ 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 equation (6) with compact unobserved effects satisfies Assumption 2.4(ii) and $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)dc$ is well defined. When $f_{Y_t|X_t,C}(y_t\,|\,x_t,c;\theta)$ can be written in the form $f(y_t - m(x_t,c;\theta))$, where f is a known density and $|\partial/\partial cm(x_t,c;\theta)| > a > 0$, Assumption 2.4(ii) holds.[13] However, in binary choice models and censored models, Assumption 2.4(ii) will not hold if the support of the unobserved heterogeneity equals the entire real line. The Fourier transforms in equation (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.

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

Let $\alpha = (\theta, \lambda)$, $\alpha_0 = (\theta_0, \lambda_0)$, and $\mathscr{A} = \Theta \times \Lambda$, and consider the following parametric function

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

where $h(\xi, y, x, \overline{w}_{-1}; \lambda) = -\lambda_1 \int_{\overline{\mathscr{W}}_1} e^{-i\xi \sum_{k=1} \lambda_k \overline{w}_k} f_{Y|X,\overline{W}}(y|x, \overline{w})d\overline{w}_1$, $\Omega(y, x, \overline{w}_{-1})$ is a positive weighting function, and $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$.[14] While Assumption 2.4(i) ensures that the term on the top of the quotient in equation (10) is well defined, Assumptions 2.4(ii) and (iii) guarantee that the term on the bottom of the quotient in equation (10) is well defined and nonzero. We can use $\phi_{v;\alpha}$ to construct a semiparametric 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 non-negative 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 sufficient conditions, 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 the Fourier Inversion Formula to the semiparametric family of functions $\{\phi_{V;\alpha}(\xi) : \alpha \in \Theta \times \Lambda\}$ in equation (A.7) to construct a semiparametric family of density functions $\{f_{C|\overline{W}}(c|\overline{w}; \alpha) : \alpha \in \mathscr{A}\}$ in equation (A.11). Because the semiparametric 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 an internally consistent semiparametric distribution of the unobserved heterogeneity.[15] 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)–(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 equation (A.11) is a conditional density function for α in the neighborhood.*

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

[15] See details in equation (A.7).

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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 $f(y|x,\overline{w};\alpha_0) = f(y|x,\overline{w};\widetilde{\alpha}) = f(y|x,\overline{w})$ for all $(y,x,\overline{w}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}$ with probability one at the probability distribution of the random variable $(Y,X,\overline{W})$.

(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.

We next provide sufficient conditions for the identification of α_0. First, we 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 semiparametric density function of observable variables:

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

Equation (14) is called a semiparametric density function, because it also depends on the density of observables $f(y|x,\overline{w})$, which is not parametrically specified. As in Appendix A, 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 semiparametric 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 the 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, defined as

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

where the expectation is taken with respect to $f(y|x,\overline{w};\alpha_0)$.

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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 the 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 equation (16) is negative definite for $(x, \overline{w}) \in \mathscr{X} \times \overline{\mathscr{W}}$ with probability one and that 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 equation (4) and the CREs in Assumption 2.2, θ_0 and λ_0, are identifiable from the population distribution of a panel data sample $\{(y_t, x_t, t = 0, 1, ..., T), \overline{w}\}$. In particular, the density function of the remainder term V in CRE is also identified.*

We prove Theorem 2.1 in four steps and summarize them in Appendix A.1 to make the identification strategy more transparent. We also provide primitive sufficient conditions for Assumption 2.6 and work through a binary choice model in Appendix B of the online supplement.

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 equation (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 semiparametric 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:

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

where $\mathscr{A} \subset R^{d_\theta + K_1}$ is compact with a nonempty 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 that converges very slowly. Thus, we replace $\mathscr{F}_1$ with a finite-dimensional sieve space $\mathscr{F}_{1N}$ that becomes dense in $\mathscr{F}_1$ when the sample size N increases. Specifically, let $\{(Y_i, X_i, \overline{W}_i)\}_{i=1}^N$ be a sample of observed variables. The empirical analogue of the expression in equation (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 maximization problem are

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

The estimation is a standard sieve ML procedure and there is a vast literature on such estimation including Shen (1997), Chen and Shen (1998), Ai and Chen (2003), Chen, Liao, and Sun (2014), and Hahn, Liao, and Ridder (2018). Here, we provide the asymptotic properties of the sieve MLE estimator but leave the regularity conditions to Appendix C of the online supplement.

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 that 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 $v^* = (v_1^*, \ldots, v_{d_\alpha}^*) \in \mathscr{V}^{d_\alpha}$ such that for

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

all $v \in \mathscr{V}$, $d\rho_\gamma(g_0)[v] = \langle \gamma' v^*, v\rangle_\psi$, where $\mathscr{V}$ is the closed linear span of $\mathscr{G} - g_o$, v_j^* satisfies that $d\rho_j(g_0)[v] = \langle v_j^*, v\rangle_\psi$, and $\rho_j(g) = \alpha_j$ for $j = 1, \ldots, d_\alpha$. The following theorem summarizes the limiting distribution of $\widehat{\alpha} = (\hat{\theta}, \hat{\lambda})$.

THEOREM 3.1. *Suppose that Assumptions C.1–C.5 in Appendix C of the online supplement hold. Then,*

$$\sqrt{N}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathscr{N}(0, \Omega),$$

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

In Appendix C of the online supplement, we also present a consistent estimator for the asymptotic covariance matrix for inference. In particular, from Hahn et al. (2018), if the sieve space is generated by a finite number of basis functions such as the Hermite polynomial series, 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 of the online supplement for details.

3.1. Sieve MLE Based on the Hermite Polynomial Series

Theorem 3.1 holds for general sieve space $\mathscr{F}_{1N}$. Here, we discuss how to implement sieve MLE when we consider Hermite polynomial series as our sieve basis functions for the nonparametric nuisance components, $\sqrt{f_1(\cdot)}$. Denote $\phi(\cdot)$ and $H_j(\cdot)$ as the PDF of standard normal and the jth order Hermite polynomial. It follows that $h_j(\cdot) = H_j(\cdot)\phi(\cdot)$ forms 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 equation (17), we obtain:

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

The empirical analogue of the expression in equation (18) is given by

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

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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

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

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

4. DISCUSSION

This section proposes a Hausman-type test for the distribution assumption on V, extends the identification results to dynamic nonlinear panel data models, and provides an identification result on partial effects.

4.1. Specification Test for the Fully Parametric Model

A popular approach in the literature is to impose a distribution assumption on $V \sim f_V(v; \tau_0)$ to obtain a fully parametric likelihood model and to apply the parametric ML method. For example, one may impose a normality assumption on V such that $V \sim f_V(v; \tau_0) = \tau_0^{-1/2}\phi(v \cdot \tau_0^{-1/2})$ for some $\tau_0 > 0$.[16] 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 fully parametric MLE 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 $V \not\sim f_V(v; \tau)$ 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 $\sqrt{N}(\widehat{\alpha} - \widehat{\alpha}_{pa}) \xrightarrow{d} N(0, \Omega_{di})$,

[16] Chamberlain (1980) uses the normality specification in a static probit model. Wooldridge (2005) also employs 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.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

where Ω_{di} stands for the asymptotic covariance of $\sqrt{N}(\widehat{\alpha}-\widehat{\alpha}_{pa})$ under the null. Let $\widehat{\Omega}_{di}$ denote a consistent estimator for Ω_{di}, and define the test statistics as

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

and the null distribution of $\widehat{S}_N$ is a chi-squared distribution with degrees of freedom equal to the dimension of α_0. A formula for $\widehat{\Omega}_{di}$ is given in Appendix D of the online supplement.

4.2. Extension to Dynamic Nonlinear Panel Data Models

Consider a parametric dynamic panel data density function:

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

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 equation (27) are of great practical value. We show that the identification strategy for the static panel data model in Theorem 2.1 can be easily adapted 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 $\gamma_0 \in \Gamma$ 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 nonzero, 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 equation (A.1), we can use Assumptions 4.1 and 4.2(i) & (ii) to obtain

$$f_{Y_2,Y_3,...,Y_T|X,Y_1,\overline{W}}(y_2,y_3,...,y_T|x,y_1,\overline{w})$$
$$= \int_{\mathscr{C}_T}\prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t\,|\,x_t,y_{t-1},\overline{w},c) f_{C|X_t,Y_{t-1},\overline{W}}(c|x,y_1,\overline{w})dc$$

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

$$= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t \,|\, x_t, y_{t-1}, c) f_{C|Y_1,\overline{W}}(c|y_1,\overline{w})dc$$

$$= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t \,|\, x_t, y_{t-1}, c; \theta_0) f_V(c - y_1\gamma_0 - \overline{w}\lambda_0)dc. \tag{29}$$

In other words, we can write the observable density function $f_{Y_2,Y_3,\ldots,Y_T|X,Y_1,\overline{W}}$ 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 also be applied here.

4.3. Identification of Partial Effects

In most empirical applications, researchers are also interested in partial effects, which 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 a continuous X_{tk} on Y_t is the partial derivative of $\mathrm{E}[Y_t|X_t, C]$ with respect to X_{tk}:

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

If X_{tk} is a discrete variable, the 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. Therefore, 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:

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

Here, 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 the APE.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

COROLLARY 4.1. *Under Assumptions 2.1–2.6, the APE as defined in equation (31) is identified from the population distribution of* $\{(y_t, x_t, t = 0, 1, ..., T), \overline{w}\}$.

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

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

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

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

For the parametric dynamic panel data model in equation (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. We can define the magnitude of state dependence as an average partial effect from $Y_{t-1} = 0$ to $Y_{t-1} = 1$ at fixed values of all other variables.

5. 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 on 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 parametric nonlinear panel data models. In other words, 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

We first consider

$$\begin{aligned} f_{Y|X,\overline{W}}(y|x,\overline{w}) &= \int_{\mathscr{C}} f_{Y|X,\overline{W},C}(y|x,\overline{w},c) f_{C|X,\overline{W}}(c|x,\overline{w})dc \\ &= \int_{\mathscr{C}} f_{Y|X,\overline{W},C}(y|x,\overline{w},c) f_{C|\overline{W}}(c|\overline{w})dc \end{aligned}$$

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

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

where we use (a) the law of total probability, (b) Assumptions 2.2 and 2.3(i)–(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 that $\lambda_{01} \neq 0$. Given each (y, x, w_{-1}), 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,

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

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

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

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

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

The intuition behind 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 equation (A.1). For example, if we consider the simplest case,

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

$\lambda_0 = (\lambda_{01}, ..., \lambda_{0K_1})^T = (-1, 0, ..., 0)'$, then equation (A.4) becomes

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

The result in equation (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

$$\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}_{-1}} \left(-\lambda_{0j} \int_{\overline{\mathscr{W}}_1} e^{-i\xi \sum_{k=1} \lambda_{0k}\overline{w}_k} \underbrace{f_{Y|X,\overline{W}}(y|x,\overline{w})}_{\substack{\text{observable}\\ \text{from data}}} d\overline{w}_1 \right) \Omega(y, x, \overline{w}_{-1}) dydxd\overline{w}_{-1}$$

$$= \phi_v(\xi) \int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} \underbrace{f_{Y|X,C}(y|x,c;\theta_0)}_{\substack{\text{population}\\ \text{density}}} \Omega(y, x) dydxdc, \qquad \textbf{(A.5)}$$

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 equation (A.5) are well defined. By Assumption 2.4(iii), dividing both sides of equation (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,

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

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

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

where $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$. Note that the terms in the numerator and denominator of the fraction in equation (A.6) are known. While the term in the numerator can be estimated directly from the data, the term in the denominator can be constructed by the parametric panel data density function $f_{Y_t|X_t,C}(y_t\,|\,x_t, c; \theta)$.

With the correct specification of $f_{Y|X,C}(y|x,c;\theta)$ in Assumption 2.1 and Assumptions 2.2–2.5, we have the well-defined expressions for $\phi_{v;\alpha_0}(\xi)$ and $\phi_{v;\alpha}(\xi)$ in equations (A.6) and (A.7), respectively. Consider the quotient of the two functions

$$\frac{\phi_{v;\alpha}(\xi)}{\phi_{v;\alpha_0}(\xi)} = \frac{Q_1(\xi;\lambda)}{Q_2(\xi;\theta)},$$

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

where

$$Q_1(\xi;\lambda)=\frac{\lambda_1\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}}e^{-i\xi\sum_{k=1}\lambda_k\overline{w}_k}f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}}{\lambda_{01}\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}}e^{-i\xi\sum_{k=1}\lambda_{0k}\overline{w}_k}f_{Y|X,\overline{W}}(y|x,\overline{w})\Omega(y,x,\overline{w}_{-1})dydxd\overline{w}},$$

$$Q_2(\xi;\theta)=\frac{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}}e^{-i\xi c}f_{Y|X,C}(y|x,c;\theta)\Omega(y,x)dydxdc}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}}e^{-i\xi c}f_{Y|X,C}(y|x,c;\theta_0)\Omega(y,x)dydxdc}.$$

Denote $\lambda_{-1}=(\lambda_2,\lambda_3,\ldots,\lambda_{K_1})$. Here, $Q_1(\xi;\lambda)$ represents the relative value of the weighted scaled Fourier transform of the population density of the observed sample $f_{Y|X,\overline{W}}$ with the scale factor λ_1 and the weight factor λ_{-1}, and $Q_2(\xi;\theta)$ indicates the relative value of the weighted Fourier transform of the parametric panel data density function $f_{Y_t|X_t,C}(y_t\,|\,x_t,c;\theta)$.

The Fourier transform gives a frequency domain representation of a function defined on a certain domain, typically the real line and commonly referred to as the time domain. The potential equivalence of $\phi_{v;\alpha}(\xi)$ and $\phi_{v;\alpha_0}(\xi)$ implies that $Q_1(\xi;\lambda)=Q_2(\xi;\theta)$; that is, this refers to the weighted frequency domain representations $Q_1(\xi;\lambda)$, and $Q_2(\xi;\theta)$ being the same. Therefore, the parameter $\alpha=(\theta,\lambda)$ for which there is equivalence of $\phi_{v;\alpha}(\xi)$ and $\phi_{v;\alpha_0}(\xi)$ is determined by an equal change in the weighted frequency domain representation of $f_{Y|X,\overline{W}}$ with the scale factor λ_1 and the weight factor λ_{-1}, as well as the weighted frequency domain representation of $f_{Y_t|X_t,C}(y_t\,|\,x_t,c;\theta)$.

Applying the inverse Fourier transform to $\phi_{v;\alpha}(\xi)$ yields a semiparametric 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 $f_{v;\alpha_0}(v)=1/2\pi\int_{-\infty}^{\infty}e^{-i\xi v}\phi_{v;\alpha_0}(\xi)d\xi=f_V(v)$. This suggests that the PDF of the remainder term V, $f_V(v)$, can be expressed in terms of $\phi_v(\cdot)$ by means of the Fourier inversion formula.[17]

The functions $f_{v;\alpha}$ and $f_{v;\alpha_0}$ are defined through the inverse Fourier transform that synthesizes these functions from their frequency domain representations. Thus, $f_{v;\alpha}$ and $f_{v;\alpha_0}$ are functions of time domains. We can view the question of the potential equivalence in either the time domain or in the frequency domain, sometimes with equal ease and sometimes with one point of view preferred. Here, the potential equivalence of $f_{v;\alpha}$ and $f_{v;\alpha_0}$ is easier to interpret and analyze in their frequency domain representation. As discussed in the previous comment, the parameter $\alpha=(\theta,\lambda)$ for the equivalence of $f_{v;\alpha}$ and $f_{v;\alpha_0}$ is determined by an equal change in the weighted frequency domain representation of $f_{Y|X,\overline{W}}$ with the scale factor λ_1 and the weight factor λ_{-1}, as well as the weighted frequency domain representation of $f_{Y_t|X_t,C}(y_t\,|\,x_t,c;\theta)$.

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 non-negative integers, $a=(a_1,\ldots,a_n)'$, denote $[a]=a_1+\cdots+a_n$, and let D^a denote the differential operator defined by $D^a=\partial^{[a]}/\partial x_1^{a_1}\ldots\partial x_d^{a_d}$. The space $\mathscr{S}(\mathbb{R}^n)$ is a collection of smooth functions $g(x)$ such

[17] Under the integrability conditions of f and $\widehat{f}$, we have $f(x)=(2\pi)^{-n}\int_{\mathbb{R}^n}e^{ix\cdot\xi}\widehat{f}(\xi)d\xi$. This Fourier inversion theorem comes from Chapter 7.5 in Folland (2009)

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

that for all multi-indices a, b, we have

$$\sup_{x\in\mathbb{R}^n} |x^a D^b g(x)| = c_{a,b}(g) < \infty, \tag{A.9}$$

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 such as $e^{-|x|^2}$. Supposing $f \in \mathscr{S}(\mathbb{R}^n)$, then $f(x) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{ix\cdot\xi} \widehat{f}(\xi) d\xi$.[18]

We next attempt to connect this semiparametric 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

$$c_\alpha(\overline{w}) \equiv \int_{\mathscr{C}} f_{v;\alpha}(c - \overline{w}\lambda) dc, \tag{A.10}$$

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

$$f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{c_\alpha(\overline{w})} f_{v;\alpha}(c - \overline{w}\lambda), \tag{A.11}$$

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

Note that when $\mathscr{C} = \mathbb{R}$, the scale factor $c_\alpha(\overline{w})$ has a simpler expression. Using the Fourier transform analytic theory for generalized functions,[19] we can write the last term of the integrand in equation (A.10) as $1/2\pi \int_{\mathscr{C}} e^{-i\xi c} dc = 1/2\pi \int_{-\infty}^{\infty} e^{-i\xi c} dc = \delta(\xi)$, where $\delta(\xi)$ is the Dirac delta function, and it is the Fourier transform of a constant function.[20] The important property of the delta function is that $\int f(\xi)\delta(\xi) dt = f(0)$ for all continuous and compactly supported functions $f(\cdot)$.[21] With this property, if $\phi_{v;\alpha}(\xi)$ is continuous and compactly supported, then equation (A.10) can be further reduced as

$$c_\alpha(\overline{w}) = \int_{-\infty}^{\infty} \underbrace{\left(e^{i\xi\overline{w}\lambda} \phi_{v;\alpha}(\xi)\right)}_{\text{a function of } \xi} \underbrace{\left(\frac{1}{2\pi} \int_{\mathscr{C}} e^{-i\xi c} dc\right)}_{\delta(\xi)} d\xi = \phi_{v;\alpha}(0). \tag{A.12}$$

Recall that $f_{C|\overline{W}}(c|\overline{w};\alpha_0) = f_{v;\alpha_0}(c - \overline{w}\lambda)$, which is the conditional density function of the unobserved heterogeneity. To allow a comparison between $f_{v;\alpha}$ and $f_{v;\alpha_0}$, we need to adjust values measured on different scales of the parameters α to a common scale as a density function. This implies that a sufficient condition for the potential equivalence of $f_{C|\overline{W}}(c|\overline{w};\alpha)$ and $f_{C|\overline{W}}(c|\overline{w};\alpha_0)$ is an equal change in the weighted frequency domain representation of $f_{Y|X,\overline{W}}$ with the scale factor λ_1 and the weight factor λ_{-1}, as well as the weighted frequency domain representation of $f_{Y_t|X_t,C}(y_t|x_t, c;\theta)$ and the density adjustment of the different parameters α.

[18] The result comes from Proposition 1.1 of Chapter X in Torchinsky (2012).

[19] The detailed results for generalized functions are in Chapter 9 of Folland (2009).

[20] The result is in Example 1 of Chapter 9.4 of Folland (2009).

[21] The result is in Example 2 of Chapter 9.1 of Folland (2009).

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

Proof of Lemma 2.1. First, by Assumptions 2.2(i) & (ii), 2.3(i)–(iii), 2.4, and 2.5(ii), $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is well defined. Next, Assumptions 2.5(i) & (ii) imply that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is continuous for all α and is non-negative for α in some open neighborhood of α_0. With the definition of $c_\alpha(\overline{w})$ in equation (A.10), we can obtain that the integration of $f_{C|\overline{W}}(c|\overline{w};\alpha)$ over the domain $\mathscr{C}$ is equal to one. ■

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

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha)dc. \tag{A.13}$$

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

$$\begin{aligned}\int_{\mathscr{Y}} f(y|x,\overline{w};\alpha)dy &= \int_{\mathscr{C}} \left(\int_{\mathscr{Y}} f_{Y|X,C}(y|x,c;\theta)dy\right) f_{C|\overline{W}}(c|\overline{w};\alpha)dc \\ &= \int_{\mathscr{C}} f_{C|\overline{W}}(c|\overline{w};\alpha)dc = 1. \end{aligned} \tag{A.14}$$

Because $f_{C|\overline{W}}(c|\overline{w};\alpha) = 1/c_\alpha(\overline{w}) f_{v;\alpha}(c-\overline{w}\lambda)$, we can write the expression of $f(y|x,\overline{w};\alpha)$ as

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) \frac{1}{c_\alpha(\overline{w})} f_{v;\alpha}(c-\overline{w}\lambda)dc.$$

This shows that the shape of $f(y|x,\overline{w};\alpha)$ is derived from the shape of the parametric nonlinear panel data model $f_{Y|X,C}(y|x,c;\theta)$ modified by the semiparametric family of conditional density functions of the unobserved heterogeneity $1/c_\alpha(\overline{w}) f_{v;\alpha}(c-\overline{w}\lambda)$. In other words, $f(y|x,\overline{w};\alpha)$ is a weighted sum of the function $f_{Y|X,C}(y|x,c;\theta)$, where the weighting function is $1/c_\alpha(\overline{w}) f_{v;\alpha}(c-\overline{w}\lambda)$. The parameter for the observational equivalence of $f(y|x,\overline{w};\alpha)$ and $f(y|x,\overline{w};\alpha_0) = f(y|x,\overline{w})$ is the collection of the parameters α such that a variation of θ in $f_{Y|X,C}(y|x,c;\theta)$ is scaled by a variation of $\alpha = (\theta,\lambda)$ in $1/c_\alpha(\overline{w}) f_{v;\alpha}(v)$ such that a combination of the two variations is equal to the population distribution of the observable, $f(y|x,\overline{w})$. Under a general framework of conditional maximum likelihood estimation, we have the following result.

LEMMA A.1. *If α_0 is identified and $E[\log f(Y|X,\overline{W};\alpha)|X=x,\overline{W}=\overline{w}] < \infty$ 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 $\alpha \neq \alpha_0$,

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

where we use the strict concavity of $\log(\cdot)$. ■

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

Differentiating equation (A.14) 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 equation (15), we have $\partial/\partial\alpha K(\alpha;x,\overline{w})\big|_{\alpha=\alpha_0} = 0$. Similarly, differentiating equation (A.14) 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 the outer product of the gradient of the log likelihood:

$$K''(\alpha_0;x,\overline{w}) = -\mathrm{E}\Big[\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)\big|_{\alpha=\alpha_0}\cdot\frac{\partial}{\partial\alpha}\log f(Y|X,\overline{W};\alpha)'\big|_{\alpha=\alpha_0}\Big|X=x,\overline{W}=\overline{w}\Big]. \tag{A.15}$$

Proof of Theorem 2.1. First, we discussed that the semiparametric density function of observable variables in equation (A.14) is well defined using Assumptions 2.2(i) & (ii), 2.3(i)–(iii), 2.4, and 2.5(i) & (ii). We next proceed to prove the identification result of α_0 using the concavity of the conditional Kullback–Leibler information criterion in Assumption 2.6; that is, the second derivative of $K(\alpha;x,\overline{w})$ in equation (16) is negative definite. Applying the identification result of α_0 to equation (A.8) leads to the identification of the density $f_V(\cdot)$. ■

A.1. Summary of Identification Steps

In this subsection, we present a heuristic sketch of how to utilize the CRE specification and the Fourier transform to construct an internally consistent likelihood function. There are four steps in 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 equation (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 yield the product of the Fourier transforms in equation (A.5) and then extend the relationship to parameters other than the true parameter to obtain the semiparametric function in equation (A.7).

Step 3: Apply the inverse Fourier transform.

Under Assumption 2.5, the inverse Fourier transform is applicable and we can recover the semiparametric distribution of the unobserved heterogeneity in equation (A.11) using the inverse transform and normalization.

Step 4: Construct an internally consistent average likelihood.

We can then combine the semiparametric distribution of the unobserved heterogeneity in Step 3 with the parametric panel data models to obtain the internally consistent average likelihood equation (A.13).

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit: http://doi.org/10.1017/S0266466620000481

REFERENCES

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71(6), 1795–1843.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* 6(3), 251–262.

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

Alvarez, J. & 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. & S. Bonhomme (2009) Robust priors in nonlinear panel data models. *Econometrica* 77(2), 489–536.

Arellano, M. & S. Bonhomme (2011) Nonlinear panel data analysis. *Annual Review of Economics* 3, 395–424.

Arellano, M. & S. Bonhomme (2012) Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies* 79(3), 987–1020.

Arellano, M. & 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*. Wiley.

Bester, C. A. & 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.

Briesch, R. A., P. K. Chintagunta, & R. L. Matzkin (2010) Nonparametric discrete choice models with unobserved heterogeneity. *Journal of Business & Economic Statistics* 28(2), 291–307.

Browning, M. & 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.

Chamberlain, G. (2010) Binary response models for panel data: Identification and information. *Econometrica* 78(1), 159–168.

Chen, S., J. Si, H. Zhang, & Y. Zhou (2017) Root-N consistent estimation of a panel data binary response model with unknown correlated random effects. *Journal of Business & Economic Statistics* 35(4), 559–571.

Chen, X., Z. Liao, & Y. Sun (2014) Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics* 178(1), 639–658.

Chen, X. & X. Shen (1998) Sieve extremum estimates for weakly dependent data. *Econometrica* 66(2), 289–314.

Chernozhukov, V., I. Fernández-Val, J. Hahn, & 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, & 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 Society.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press

Gayle, G.-L. & 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. & S. D. Namoro (2013) Estimation of a nonlinear panel data model with semiparametric individual effects. *Journal of Econometrics* 175(1), 46–59.
Graham, B. & 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, & G. Ridder (2018) Nonparametric two-step sieve M estimation and inference. *Econometric Theory* 34(6), 1281–1324.
Heckman, J. & B. Singer (1984) The identifiability of the proportional hazard model. *The Review of Economic Studies* 51(2), 231–241.
Hoderlein, S. & E. Mammen (2007) Identification of marginal effects in nonseparable models without monotonicity. *Econometrica* 75(5), 1513–1518.
Hoderlein, S. & H. White (2012) Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics* 168(2), 300–314.
Honoré, B. & E. Kyriazidou (2000) Panel data discrete choice models with lagged dependent variables. *Econometrica* 68(4), 839–874.
Honoré, B. & E. Tamer (2006) Bounds on parameters in panel dynamic discrete choice models. *Econometrica* 74(3), 611–629.
Honoré, B. E. & A. Lewbel (2002) Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica* 70(5), 2053–2063.
Hsiao, C. (2015) *Analysis of Panel Data*. Cambridge University Press.
Hu, Y. & G. Ridder (2010) On deconvolution as a first stage nonparametric estimator. *Econometric Reviews* 29(4), 365–396.
Hu, Y. & G. Ridder (2012) Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics* 27(3), 347–385.
Hu, Y. & M. Shum (2012) Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics* 171(1), 32–44.
Kiefer, J. & J. Wolfowitz (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* 27(4), 887–906.
Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* 73(364), 805–811.
Lindsay, B. G. & M. L. Lesperance (1995) A review of semiparametric mixture models. *Journal of Statistical Planning and Inference* 47(1–2), 29–39.
Matzkin, R. (2003) Nonparametric estimation of nonadditive random functions. *Econometrica* 71(5), 1339–1375.
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(6), 2555–2591.
Shiu, J. & 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.
Wooldridge, J. (2010) *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

https://doi.org/10.1017/S0266466620000481 Published online by Cambridge University Press