Internally Consistent Estimation of Nonlinear Panel Data Models with Correlated Random Effects*

Yu-Chin Hsu† Ji-Liang Shiu‡

July 2, 2017

Abstract

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

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

---

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

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

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

1. Introduction

One of the challenges in the panel data literature is how to model the unobserved heterogeneity across individuals. For linear panel data models with an additive unobserved heterogeneity, one can use the fixed-effects formulation so there will be no requirement to postulate the distribution of the unobserved heterogeneity. Consistent estimators can be obtain by generalized method of moments methods after one applies within transformation to eliminate the unobserved effects. We refer to Baltagi (2008), Wooldridge (2010), and Hsiao (2015) for more complete literature reviews. However, in nonlinear models it is not clear how to remove the unobserved heterogeneity so there is a fundamental difference between the linear and nonlinear models.[1]

In nonlinear panel data models, there are two main treatments for the unobserved heterogeneity. One is to treat the unobserved heterogeneity as fixed parameters and the other is to treat it as random variables. When the time dimension is fixed, there are a surprising amount of differences in the identification and estimation of the models between these two treatments. When the unobserved heterogeneity is modeled as fixed effects, the dimension of unknown parameters increases at the same rate as the sample size so estimators are often subject to an incidental parameter problem and a standard maximum likelihood estimator causes bias. Honoré and Kyriazidou (2000) generalized conditional maximum likelihood approaches of Andersen (1970) and Rasch (1993) to estimate the parameters of dynamic discrete choice logit models with strictly exogenous explanatory variables.[2] Chamberlain (2010) considered the identification of binary response models when the time dimension is fixed and the distribution of individual effects is unrestricted. He showed that identification is only possible in the logistic case.[3]

On the other hand, in the random effects approach, one would specify the conditional distri-

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

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

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

bution of the dependent variable, Y_t, conditional on a $1 \times K$-dimensional vector of time-varying explanatory variables, X_t, and an individual unobserved heterogeneity, C,

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

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

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

The average likelihood in Eq. (2) is fully parametric in that it depends on a finite number of parameters, and the estimation and inference is possible under standard parametric framework. Many studies adopt this strategy. For example, Wooldridge (2005) handled the initial conditions problem of dynamic panel data problem by specifying the conditional distribution of the unobserved heterogeneity to be normal distributed with a mean which is a linear combination of the initial value, and exogenous explanatory variables. Alvarez and Arellano (2003) used a similar specification for models with large time and cross-sectional dimensions. Arellano and Bonhomme (2009) focused on estimators that maximize an average likelihood that assigns weights to different values of the unobserved heterogeneity. They provided a characterization of the class of weights that produce first-order unbiased estimators. The disadvantage of this approach is that the misspecification of $f_{C|X}(c|x;\beta)$ generally results in inconsistent estimates.

In addition to the above two approaches, an alternative one is to relax the distributional assumption on the unobserved heterogeneity and then to characterize the identified region containing the true parameter. In other words, one would lose point-identification of the parameters. For example, Honoré and Tamer (2006) relaxed the distributional assumption of the initial condition and calculated bounds on parameters in panel dynamic discrete choice models. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) showed that bounds for marginal effects in nonlinear panel models can be tighten rapidly as the number of time series observations grows.

Despite these contributions, a general way to handle the functional form misspecification of the average likelihood in Eq. (2) for fixed T remains intangible.[4] To address the void in the literature, we provide a data-driven specification of conditional distributions of the unobserved heterogeneity which is internally consistent with the proposed nonlinear panel data models in Eq. (1) and at the same time, to retain point-identification of the parameters of interest. To be specific, we consider a correlated random effects (CRE) approach without fully specifying the distribution of the unobserved heterogeneity conditional on the explanatory variables. Under the Mundlak-type specification, Eq. (2) can be written as an equation containing a time-invariant observed variable $\overline{W}$,

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

where $\alpha = (\theta, \lambda)$. This implies the average likelihood is the convolution of the proposed models and the conditional distribution of the unobserved heterogeneity. Because the Fourier transform of the convolution of the two functions is the product of their individual Fourier transforms, the characteristic function of the conditional distribution of the unobserved heterogeneity can be obtained by the quotient of two characteristic functions related to a density of observables and the proposed known parametric panel data model at the true parameter. Then, we extend the relation of the characteristic function to parameters other than the true parameter and apply Fourier inversion formula to the extended characteristic function to devise a parametric distribution of the unobserved heterogeneity conditional on exogenous variables. Combining the proposed panel data models with the recovered parametric distribution of the unobserved heterogeneity, we can construct a semi-parametric average likelihood of observables and then its parameter vector is identifiable by a standard maximum likelihood condition, the negative definiteness of the information matrix. The semi-parametric average likelihood is correctly specified in the sense: the semi-parametric average likelihood evaluated at the true parameter

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

is equal to a density of observables. Because the recovered semi-parametric distribution of the unobserved heterogeneity is based on the proposed model and the density of observables, we can regard the specification as a data-driven one.

The identification method is constructive in that we can propose a semi-parametric two-step maximum likelihood estimator based on the identification result. The first step is to estimate the density of observables in the quotient relation using kernel density estimator and then we can obtain a semi-parametric distribution of the unobserved heterogeneity conditional on exogenous variables. In the second step, we devise a maximum likelihood estimator that takes the form of an average likelihood estimator with the proposed panel data model and the semi-parametric distribution of the unobserved heterogeneity given covariates. Under standard regularity conditions, the proposed estimator is consistent and asymptotically normal. Because the semi-parametric two-step maximum likelihood estimator relies on the first step nonparametric kernel density estimate for a density of observables, the estimator can be interpreted as a data-driven procedure. The study of a data-driven procedure in this context is novel and enables an internally consistent CRE approach of common nonlinear panel data models with unobserved heterogeneity, such as discrete choice, censored and truncated, sample selection models, etc. In addition, we show that the average partial effects can be identified. Furthermore, we provide an extension of the proposed method to common 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 strategies are also considered in Honoré and Lewbel (2002), Hu and Shum (2012), and Shiu and Hu (2013). Our identification strategy is related to the literature on nonparametric deconvolution including the measurement error models (see Schennach (2004); Schennach (2007); Hu and Ridder (2010); Hu and Ridder (2012)), and panel data models (see Evdokimov (2011); Arellano and Bonhomme (2012)), etc. Evdokimov (2011) established nonparametric identification of a panel data model with nonadditive unobserved heterogeneity and developed a nonparametric estimation procedure. Arellano and Bonhomme (2012) considered random coefficients panel data models where the coefficients can be arbitrarily correlated with the covariates and obtained identification of the density of individual effects.

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

2. Identification

For $t = 1,\ldots,T$, let Y_t denote the dependent variables of interest and X_t denote a $1\times K$-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 semi-parametric panel data models of the following form:

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

where $y_t \in \mathscr{Y}_t$, $x_t \in \mathscr{X}_t$ and $c \in \mathscr{C}$. Also, $\theta \in \Theta$ is a parameter which specifies the exact structure of the model and Θ is the parameter space. We assume that there exists $\theta_0 \in \Theta$ such that $f_{Y_t|X_t,C}(y_t|x_t,c;\theta_0) = f_{Y_t|X_t,C}(y_t|x_t,c)$, where $f_{Y_t|X_t,C}(y_t|x_t,c)$ is the population density function. The panel data model in Eq. (4) may be derived from more primitive econometric model. Throughout this paper, we provide such conditions for the following single-index binary-choice model to illustrate our general method:

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

where $1(\cdot)$ is the 0-1 indicator function, and ε_t is independent of X with a known time-specific distribution function F_{ε_t}. The corresponding conditional distribution is:

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

Note that the identification result of the panel data model in Eq. (4) can be applied to other semi-parametric nonlinear panel data models.

2.1. Assumptions and Results

We assume that the panel data model in Eq. (4) is correctly specified. We make some assumptions.

Assumption 2.1. *Assume that*

(i) For $t = 1, \dots, T$, the density $f_{Y_t|X_t,C}(y_t|x_t, c; \theta)$ is known and uniformly bounded above for all $y_t \in \mathcal{Y}_t$, $x_t \in \mathcal{X}_t$, $c \in \mathcal{C}$ and $\theta \in \Theta$;

(ii) there exists a true parameter $\theta_0 \in \Theta$ such that $f_{Y_t|X_t,C}(y_t|x_t, c; \theta_0)$ is the joint distribution of the observable variables;

(iii) the parameter space, Θ, is a compact subset of $\mathbb{R}^{d_\theta}$.

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

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

Assume that there exists $K_1 \times 1$-dimensional vector of nonzero coefficients $\lambda_0 = (\lambda_{01}, ..., \lambda_{0K_1})'$ in Λ that is a compact subset of R^{K_1} such that

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

where the remainder term V is independent of $\overline{X}_1, ..., \overline{X}_{K_1}$.

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

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

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

(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, Part (ii) of Assumption 2.3 requires that it does not provide any more information on Y_t when the regressors X_t, and C are in presence. Part (iii) of Assumption 2.3 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 $\overline{W}$. This condition implies that the first K_1 explanatory variables in X cannot be time-invariant variables or variables that have exact linear time trends for all cross-section units. Under Assumption 2.2, a sufficient condition for Assumption 2.3(iii) is that the remainder error V is independent of X. Denote the conditional joint density as

$$f_{Y|X,C}(y|x,c) = \prod_{t=1}^{T} f_{Y_t|X_t,C}(y_t|x_t,c). \tag{8}$$

Assumption 2.4. *(Well Defined Characteristic Function)*

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

(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* $\lambda \in \Theta$,

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

The characteristic functions $\sum_{j=1}^{K_1} \left(\lambda_j \int_{\overline{\mathscr{W}}_j} e^{-i\xi \sum_{k=1} \lambda_k \overline{w}_k} f_{Y|X,\overline{W}}(y|x,\overline{w})d\overline{w}_j \right)$ for $(y,x,\overline{w}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}$ and $\lambda \in \Lambda$ and $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)dc$ for all $(y,x) \in \mathscr{Y} \times \mathscr{X}$ and $\theta \in \Theta$ are finite by Assumptions 2.4(i) & (ii), respectively. Because $f_{Y|X,C}(y|x,c;\theta)$ is uniformly bounded for all $\theta \in \Theta$ by Assumption 2.1, if the support of the unobserved heterogeneity $\mathscr{C}$ is compact then Assumption 2.4(ii) holds. Thus, the density $f_{Y|X,C}(y|x,c;\theta)$ for the binary-choice model in Eq. (6) with compact unobserved effects satisfies Assumption 2.4(ii) and $\int_{\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)dc$

is a well defined function. On the other hand, 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 $\frac{\partial}{\partial c}m(x_t,c;\theta) > a > 0$, then Assumption 2.4(ii) holds.[5] Under the CRE condition, the unobserved heterogeneity C is a linear combination of the time average of the explanatory variables in $\overline{W}$ and unobservable V so its support should encompass the supports of these variables. However, requiring the support of the unobserved heterogeneity to be equal to the real line may fail Assumption 2.4(ii) for binary-choice models and censored models. The characteristic functions in Eq. (9) appear as denominators in our identifying formula and Assumption 2.4(iii) rules out zero denominator. Note that all conditions in Assumption 2.4 are testable since they involve the density of observables and the proposed known semi-parametric nonlinear panel data model.

Denote Λ as the parameter spaces containing the population parameters λ_0. Let $\alpha = (\theta, \lambda)$, $\alpha_0 = (\theta_0, \lambda_0)$ and $\mathscr{A} = \Theta \times \Lambda$. Under Assumption 2.4, we can construct a parametric family of functions related to the characteristic function of the remainder term V in the CRE assumption in Eq. (B.10) in Appendix.

Assumption 2.5. *(Continuous Parameter Structure)*
Assume that (i) The semi-parametric panel data density function $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ is continuous at θ for all $\theta \in \Theta$ and $t = 1, ..., T$;
(ii) the parametric family of functions $\{\phi_{V;\alpha}(\xi) : \alpha \in \mathscr{A}\}$ defined in Eq. (B.10) belongs to $L^1(\mathbb{R})$.

Under Assumption 2.5, we can apply Fourier Inversion Formula in Proposition B.1 to the parametric family of functions $\{\phi_{V;\alpha}(\xi) : \alpha \in \Theta \times \Lambda\}$ in Eq. (B.10) to construct the parametric family of density functions $\{f_{C|\overline{W}}(c|\overline{w};\alpha) : \alpha \in \mathscr{A}\}$ in Eq. (B.16).[6] Because the parametric characteristic function $\phi_{V;\alpha}(\xi)$ is derived from the data and the proposed model,[7] $f_{C|\overline{W}}(c|\overline{w};\alpha)$ can be regarded as internally consistent parametric distribution of the unobserved heterogeneity.

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.

[5] Consider $\int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta)dc \le b\sum_{t=1}^T \int_{\mathscr{C}} f(y_t - m(x_t,c;\theta))dc \le b\sum_{t=1}^T \int f(u_t)\frac{du_t}{\frac{\partial m}{\partial c}} < \infty$, where $u_t = y_t - m(x_t,c;\theta)$.

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

[7] See details in Eq. (B.10).

This implies that there is a unique parameter associated with each population structure in the parameter space $\mathscr{A}$.

Definition 2.1. *(i) Two 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 parameter α_0 is said to be identifiable if there exists an open neighborhood of α_0 in $\mathscr{A}$ containing no other parameter observationally equivalent to α_0.

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

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) \underbrace{f_{C|\overline{W}}(c|\overline{w};\alpha)}_{\substack{\text{constructed from} \\ f(y|x,\overline{w}) \text{ and} \\ f_{Y|X,C}(y|x,c;\theta)}} dc. \tag{10}$$

As we will see in Appendix B, we will apply Fourier transformations to combine the parameter structure of $f_{Y|X,C}(y|x,c;\theta)$ with $f(y|x,\overline{w})$ and construct $f_{C|\overline{W}}(c|\overline{w};\alpha)$ under the CRE specification. The parametric density function is correctly specified because $f(y|x,\overline{w};\alpha_0) = f(y|x,\overline{w})$. Next, 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

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

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 $j = 1,\ldots,d_\theta + K_1$, we have the gradient of $K(\alpha;x,\overline{w})$ is a $(d_\theta + K_1) \times 1$-dimensional vector,

$$\frac{\partial}{\partial \alpha} K(\alpha; x, \overline{w}) = \left(\frac{\partial K(\alpha; x, \overline{w})}{\partial \alpha_1}, \ldots, \frac{\partial K(\alpha; x, \overline{w})}{\partial \alpha_{d_\theta + K_1}} \right)'. \tag{11}$$

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:

$$K''(\alpha_0; x, \overline{w}) = -\mathrm{E}\Big[\frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)\big|_{\alpha=\alpha_0} \cdot \frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)'\big|_{\alpha=\alpha_0} \Big| X = x, \overline{W} = \overline{w} \Big]. \tag{12}$$

Assumption 2.6. *(Concave Parameter Structure)*
Assume that the information matrix $K''(\alpha_0; x, \overline{w})$ in Eq. (12) 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 $t = 1, 2, \ldots, T$.*

3. Estimation and Asymptotic Properties

3.1. Semiparametric Two-step Estimators

We consider estimation of the parameter $\alpha = (\theta, \lambda)$ and the proposed model is semi-parametric in the sense that the distribution of the CRE remainder error V in Assumption 2.2 is not specified. The identification results in Section 2 are constructive in that the results suggest an maximum likelihood (MLE) estimator based on the parametric density function of observable variables in Eq. (10).

The fact that the population parameter α_0 is identified and is the unique solution of $K(\alpha; x, \overline{w})$ by Lemma B.2 implies that

$$\mathrm{E}\Big[\log f(Y|X, \overline{W}; \alpha) \Big| X = x, \overline{W} = \overline{w} \Big] < \mathrm{E}\Big[\log f(Y|X, \overline{W}; \alpha_0) \Big| X = x, \overline{W} = \overline{w} \Big], \text{ for all } \alpha \neq \alpha_0. \tag{13}$$

By taking the expected value of the above expression and using iterated expectations, we see that α_0 solves

$$\max_{\alpha} \mathrm{E}\Big[\log f(Y|X,\overline{W};\alpha)\Big]. \tag{14}$$

The identification result guarantees the existence of a solution to the optimization problem in Eq. (14). In Appendix, we show that $\log f(Y_i|X_i,\overline{W}_i;\alpha)$ is identified by

$$\log f(Y_i|X_i,\overline{W}_i;\alpha) = -\log c_\alpha(\overline{W}_i) + \log \rho_\alpha(Y_i,X_i,\overline{W}_i), \tag{15}$$

where

$$c_\alpha(\overline{W}_i) = \int_{-\infty}^{\infty} \frac{e^{i\xi \overline{W}_i \lambda}}{\Delta_1(\theta,\xi)} E\Big[\frac{1}{K_1}\sum_{j=1}^{K_1} \lambda_j e^{-i\xi \overline{W}\lambda} \Omega(Y,X,\overline{W}) \frac{-1}{f_{X\overline{W}}(X,\overline{W})}\Big] \psi_c(\xi) d\xi, \text{ and}$$

$$\rho_\alpha(Y_i,X_i,\overline{W}_i)$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{\mathscr{C}} \frac{f_{Y|X,C}(Y_i|X_i,c;\theta)\cdot e^{-i\xi(c-\overline{W}_i\lambda)}}{\Delta_1(\theta,\xi)} E\Big[\frac{1}{K_1}\sum_{j=1}^{K_1} \lambda_j e^{-i\xi\overline{W}\lambda}\Omega(Y,X,\overline{W}) \frac{-1}{f_{X\overline{W}}(X,\overline{W})}\Big] dc d\xi,$$

$$\psi_c(\xi) = \frac{1}{2\pi}\int_{\mathscr{C}} e^{-i\xi c} dc.$$

We develop a two-step estimator based on the identification result in (15). The first step consists of a nonparametric kernel estimator of the observed density $f_{X\overline{W}}(x,\overline{w})$ which is included in the construction of the parametric density function $f(Y_i|X_i,\overline{W}_i;\alpha)$. The second step is a maximum likelihood (MLE) estimator with the plugged-in kernel density estimator from the first step.

Procedure of MLE Estimator:

Step 1: A kernel density estimator with bandwidth δ for $f_{X\overline{W}}(x,\overline{w})$.

Let $\{Y_i,X_i,\overline{W}_i\}_{1\le i\le N}$ be a sample of observed variables. Construct the following kernel density estimator.

$$\widehat{f}_{X\overline{W}}(X_i,\overline{W}_i) = \frac{1}{N\delta^{KT+K_1}} \sum_{j=1}^{N} K\Big(\frac{X_j - X_i}{\delta}\Big) K\Big(\frac{\overline{W}_j - \overline{W}_i}{\delta}\Big), \tag{16}$$

where δ is a positive smoothing parameter that goes to zero as the sample size increases, and $K(u)$ be a kernel function in the dimension of u.

Step 2: An MLE estimator for α_0.

Based on (15), define

$$\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\widehat{f}_{X\overline{W}}) = \frac{\hat{\rho}_{\alpha,\widehat{f}_{X\overline{W}}}(Y_i,X_i,\overline{W}_i)}{\hat{c}_{\alpha,\widehat{f}_{X\overline{W}}}(\overline{W}_i)}, \tag{17}$$

where

$$\hat{c}_{\alpha,\widehat{f}_{X\overline{W}}(\overline{W}_i)} = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\sum_{j=1}^{K_1}\frac{1}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}\lambda_j e^{-i\xi\overline{W}_j\lambda}\Omega(Y_i|X_i,\overline{W}_i)\frac{-1}{\widehat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi,$$

$$\hat{\rho}_{\alpha,\widehat{f}_{X\overline{W}}}(Y_i,X_i,\overline{W}_i)$$
$$=\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\sum_{j=1}^{K_1}\frac{\lambda_j e^{-i\xi(c-\overline{W}_i\lambda)}f_{Y|X,C}(Y_i|X_i,c;\theta)}{\Delta_1(\theta,\xi)}\Big(e^{-i\xi\overline{W}_j\lambda}\Omega(Y_i,X_i,\overline{W}_i)\frac{-1}{\widehat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi$$

Define $\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\widehat{f}_{X\overline{W}})$ by replacing $f_{X\overline{W}}$ with its kernel estimate $\widehat{f}_{X\overline{W}}$ in the construction of the conditional log likelihood with the plugged-in first step for each observation i, $\log f(Y_i|X_i,\overline{W}_i;\alpha,\widehat{f}_{X\overline{W}})$. Note that it is a random function that depends on the random vector $(Y_i,X_i,\overline{W}_i)$. The empirical analogue of the expression in Eq. (14) with the plugged-in kernel density estimator is given by

$$\widehat{Q}_N(\alpha,\widehat{f}_{X\overline{W}}) = \frac{1}{N}\sum_{i=1}^{N}\log\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\widehat{f}_{X\overline{W}}). \tag{18}$$

The MLE estimator of α_0 from this maximizing problem is

$$\widehat{\alpha} \equiv \arg\max_{\alpha}\widehat{Q}_N(\alpha,\widehat{f}_{X\overline{W}}). \tag{19}$$

3.2. Large-sample Properties

The proposed two-step estimator depends on the preliminary, “first-step" kernel estimator of a function $\widehat{f}_{X\overline{W}}$ and estimates a vector of finite-dimensional parameters α via maximizing the empirical log likelihood with $\widehat{f}_{X\overline{W}}$. The estimation is a semi-parametric two-step procedure and there is a large literature on the estimation such as Newey (1994), Newey and McFadden (1994), Andrews (1994a), Pagan and Ullah (1999) Chen, Linton, and Van Keilegom (2003), and Ichimura and Lee (2010). In this section, we discuss only those important conditions for the consistency and asymptotic normality of our estimator. We leave more technical regularity conditions and the proof in Appendix. We note that the theory for our estimator requires the theory of U-statistic theory of order three.

Assumption 3.1. *Assume that*

(i) the sequence $\{Y_i, X_i\}_{i=1}^N$ *is an independent and identically distributed sequence of random variables;*

(ii) he function $\log f(y|x,\overline{w};\alpha)$ *is continuous over* $\mathscr{A}$ *for all* $(y,x,\overline{w}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}$*;*

(iii) $E[\sup_{\alpha \in \mathscr{A}} |\log f(Y|X,\overline{W};\alpha)|] < \infty$*;*

(iv) α_0 *is the unique maximizer of* $E[\log f(Y|X,\overline{W};\alpha)]$*;*

(v) the function $\log f(y|x,\overline{w};\alpha)$ *is twice continuously differentiable on the interior of* $\mathscr{A}$ *for all* $(y,x,\overline{w}) \in \mathscr{Y} \times \mathscr{X} \times \overline{\mathscr{W}}$*;*

(vi) the second derivative $\frac{\partial^2}{\partial\alpha\partial\alpha'} \log f(Y|X,\overline{W};\alpha)$ *is continuous at* α_0 *with probability one, and* $E[\sup_{\alpha \in \mathscr{A}} |\frac{\partial^2}{\partial\alpha\partial\alpha'} \log f(Y|X,\overline{W};\alpha)|] < \infty$*;*

(vii) $H \equiv E[\frac{\partial^2}{\partial\alpha\partial\alpha'} \log f(Y|X,\overline{W};\alpha_0)]$ *is negative definite at* α_0*.*

Theorem 3.1. *Under Assumption 3.1 and those regularity conditions given in Appendix, we have* $\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \mathscr{V}_\alpha)$ *where* $\mathscr{V}_\alpha$ *the variance and covariance matrix given in Appendix.*

Note that those conditions in Assumption 3.1 are standard for the theory of MLE when $f(Y|X,\overline{W};\alpha)$ is available to econometricians. However, in our case, $f_{X\overline{W}}$ has to be estimated first. Therefore, the additional assumptions in Appendix account for the estimation effect of $\hat{f}_{X\overline{W}}$ on $\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})$. These assumptions and proofs are tedious, so we leave them in Appendix. The expression of $\mathscr{V}_\alpha$ is tedious too, so we leave it in Appendix and suggest to use bootstrap for inference regarding $\widehat{\alpha}$.

4. Discussions

In this section, we discuss how to test the distribution assumption on V in Assumption 2.2, provide identification result on partial effects which is often the objects of interest in empirical studies, and discuss how to extend the results to dynamic nonlinear panel data models.

4.1. Specification Test for Normality Assumption

We provide a specification test for the distribution assumption on V in Assumption 2.2. A popular approach is to impose that $V \sim N(0,\sigma_0^2)$ for some $\sigma_0^2 > 0$.[8] Recall that in our case

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

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

$$\text{(20)} \qquad (\hat{\alpha}_{pa},\hat{\sigma}^2_{pa}) \equiv \arg\max_{\alpha\in\mathscr{A},\sigma^2\le M} \frac{1}{N}\sum_{i=1}^{N} f_{pa}(Y_i|X_i,\overline{W}_i;\alpha,\sigma^2),$$

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

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

Next, we can construct a Hausman-type test for the null hypothesis that $V \sim N(0,\sigma_0^2)$ by comparing $\hat{\alpha}$ and $\hat{\alpha}_{pa}$. To be specific, both our estimator and the fully parametric estimator for α_0 are consistent for α_0 when $V \sim N(0,\sigma_0^2)$. However, if $V \nsim N(0,\sigma^2)$ for any $\sigma^2 \le M$, then our estimator is still consistent, but in general, the fully parametric estimator will converge to a point other than α_0. Therefore, we can construct a test for whether V follows a normal distribution by comparing whether $\hat{\alpha}$ and $\hat{\alpha}_{pa}$ are close or not. Specifically, under the null hypothesis, we can show that $\sqrt{N}(\hat{\alpha}-\hat{\alpha}_{pa}) \xrightarrow{d} N(0,\mathscr{V}_{di})$ where $\mathscr{V}_{di}$ stands for the asymptotic covariance of $\sqrt{N}(\hat{\alpha}-\hat{\alpha}_{pa})$ under null. Let $\widehat{\mathscr{V}}^b_{di}$ denote the bootstrapped estimator for $\mathscr{V}_{di}$. Define the test statistics as

$$\text{(21)} \qquad \widehat{S}_N = N(\hat{\alpha}-\hat{\alpha}_{pa})(\widehat{\mathscr{V}}^b_{di})^{-1}(\hat{\alpha}-\hat{\alpha}_{pa})'$$

and the null distribution of $\widehat{S}_N$ is Chi-squared distribution with degrees of freedom equal to

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

the dimension of α_0.

4.2. Identification of Partial Effects

Theorem 2.1 provides the identification of the parameter vector θ_0 in nonlinear panel data models; however, 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}:

$$\frac{\partial \mathrm{E}[Y_t|X_t, C]}{\partial X_{tk}}. \tag{22}$$

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 values of C are meaningful to plug in. 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. To be specific, multiplying $f_{C|\overline{W}}(c|\overline{w};\alpha)$ by the observed density $f_{\overline{W}}(\overline{w})$ and integrating the product function evaluated at α_0 over $\overline{w}$ yields:

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

Then the APE is given by:

$$\mathrm{APE}(x_{tk}) = \int_{\mathscr{C}} \left(\frac{\partial \mathrm{E}[Y_t|X_t = x_t, C = c]}{\partial X_{tk}}\right) f_C(c) dc = \int_{\mathscr{C}} \left[\int_{\mathscr{Y}_t} y_t \frac{\partial f_{Y_t|X_t,C}(y_t|x_t, c;\theta_0)}{\partial x_{tk}} dy_t\right] f_C(c) dc. \tag{23}$$

For the binary-choice model in Eq. (5), the average partial effect becomes

$$\mathrm{APE}(x_{tk}) = \theta_k \int_{\mathscr{C}} \frac{\partial F_{\varepsilon_t}(x_t\theta + c)}{\partial x_{tk}} f_C(c) dc, \tag{24}$$

where F_{ε_t} is the CDF of the error term in the latent variable model. We now state our identification result for APE.

Corollary 4.1. *Under Assumptions 2.1- 2.6, the APE defined in Eq. (23) is identified from the joint distribution of a panel data sample,* $\{Y_t, X_t\}$ *for* $t = 1, 2, ..., T$.

4.3. Extension to Dynamic Nonlinear Panel Data Models

Consider a parametric dynamic panel data density function:

$$f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c; \theta), \text{ for all } t = 2, ..., T. \tag{25}$$

As discussed in Section 2, a variety of nonlinear panel data models in underlying latent variable formulation can be derived into the density of the dependent variable given a set of covariates. Time series dependence arises naturally in the context of dynamic panel data models that can be used to investigate the effects of lagged outcomes on current outcomes. Thus, the identification and estimation of models in Eq. (25) are of great practical value. We impose the following assumptions for identification in this setting.

Assumption 4.1. *(Movement of the Unobserved Effects)*
Assume that (i) $f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t, y_{t-1}, \overline{w}, c) = f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c)$ *for all* $(y_t, x_t, y_{t-1}, \overline{w}, c) \in \mathscr{Y}_t \times \mathscr{X}_t \times \mathscr{Y}_{t-1} \times \overline{\mathscr{W}} \times \mathscr{C}$*;*
(ii) $f_{C|X,Y_1,\overline{W}}(c|x, y_1, \overline{w}) = f_{C|\overline{W}}(c|\overline{w})$ *for all* $(c, x, y_1, \overline{w}) \in \mathscr{C} \times \mathscr{X} \times \mathscr{Y}_1 \times \overline{\mathscr{W}}$.

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

$$\begin{aligned}
&f_{Y_2,Y_3,...,Y_T|X,Y_1,\overline{W}}(y_2, y_3, ..., y_T|x, y_1, \overline{w}) \\
&= \int_{\mathscr{C}_T} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},\overline{W},C}(y_t|x_t, y_{t-1}, \overline{w}, c) f_{C|X_t,Y_{t-1},\overline{W}}(c|x, y_1, \overline{w}) dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c) f_{C|\overline{W}}(c|\overline{w}) dc \\
&= \int_{\mathscr{C}} \prod_{t=2}^{T} f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c; \theta_0) f_v(c - \overline{w}\lambda_0) dc.
\end{aligned} \tag{26}$$

This implies that we can write the observable density function $f_{Y_2,Y_3,...,Y_T|X,Y_1,\overline{W}}$ 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 f_V. It follows that Eq. (26) can also be handled by using Fourier transforms as before. Therefore, there is a characteristic function of V that leads to identification. Sufficient conditions for identification are similar to those in Theorem 2.1 and the identification result follows.

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

$$C = \gamma_0 Y_1 + \overline{W}\lambda_0 + V, \tag{27}$$

where the error term V is independent of Y_1, and $\overline{W}$. Then, $f_{C|X,Y_1,\overline{W}}(c|x, y_1, \overline{w}) = f_V(C - \gamma_0 Y_1 - \overline{W}\lambda_0)$. Under the specification, our approach still applies. Wooldridge (2005) specified the conditional density $f_{C|X,Y_1,\overline{W}}(c|x, y_1, \overline{w})$ in the same manner, but assumed V is normally distributed.

5. Monte Carlo Simulation

In this section, we present simulation results to illustrate the finite sample performance of the proposed two-step estimation procedure of a panel data probit model in Section 3.1. The main data generating process (DGP) is defined as follows:

$$\begin{aligned}
&Y_t = \mathbf{1}(\theta X_t + C + \varepsilon_t \geq 0), \quad \text{for } t = 1, 2, \\
&C = \lambda\overline{W} + V, \quad \overline{W} = \frac{1}{2}\sum_{t=1}^{2} X_t \\
&(X_1, X_2, \varepsilon_1, \varepsilon_2) \sim N(0, I_4),
\end{aligned}$$

where I_4 is the 4×4 identify matrix and we set $(\theta, \lambda) = (0.5, -0.5)$. We consider various distributions of V. For a random variable Q, we denote the corresponding truncated random variable over interval $[a, b]$ as $Trun(Q, [a, b])$.[9] Let μ_ω be the mean of ω. Three specifications of V are

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

considered:

$$\text{DGP I: } V \sim Trun(N(0,1),[-2,2]),$$

$$\text{DGP II: } V = \omega - \mu_\omega \text{ with } \omega \sim Trun\big(H,[-2,2]\big) \text{ and } \ln H = N(0,5),$$

$$\text{DGP III: } V = \omega - \mu_\omega \text{ with } \omega \sim Trun(Rayleigh(5),[0,10]).$$

The unobserved heterogeneities in all the simulation designs are with bounded supports so Assumption 2.4(ii) is satisfied in all cases. We consider sample sizes 500, and 1,000 and for each case, we consider 150 simulation replications. For comparison, we also consider the other two estimators. The first one is an infeasible estimator that treats V as known. The second one is the conventional random effects estimator which specifies the unobserved heterogeneity to be normally distributed. The simulation results for parameters and APE are presented in Tables 1–2 and 3–4, respectively.

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

Overall, the simulation results show that the proposed two-step 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 and III. The estimation results for APEs in Tables 3–4 have a similar pattern. While the infeasible estimator and the proposed two-step estimator perform well in all simulations, the conventional estimator perform well only in DGP I.

6. Empirical Application

We apply the proposed semiparametric two-step estimator to estimate the persistence effects of union membership using the first two periods of the panel data in Wooldridge (2005). The first period corresponds to year 1981 and the second period corresponds to year 1982. The dynamic behavior of union membership may come from its formulation of behavior maximizing the future utility and mechanical dynamics for membership. We model the membership decisions as the following dynamic probit model:

$$
\begin{aligned}
(28) \qquad & \text{Prob}(\text{Union}_t = 1 | \text{Married}_t, \text{Union}_{t-1}, D_{1982}, C) \\
& = \Phi\left(\beta_0 + \beta_1 \text{Married}_t + \rho \text{Union}_{t-1} + \beta_2 D_{1982} + C\right), \text{ for all } t = 1 \text{ and } 2.
\end{aligned}
$$

where D_{1982} is a year dummy for 1982. The correlated random effect specification for the individual unobserved heterogeneity C follows Eq. (27), where $Y_1 = \text{Union}_0$, and $\overline{W}$ is the 1×2 vector of marital status indicators.

As the identification of the proposed model hinges on assumptions in Section 2.1, it is necessary to discuss the plausibility of the identification conditions in this application. Assumption 2.1 is to specify the distribution of the disturbance in the latent variable formulation for the binary membership decisions to be normally distributed and this results in the probit model in Eq. (28). Because the disturbance is the sum of many unobserved factors after controlling the covariates, we can invoke the central limit theorem to conclude that the disturbance has an approximate normal distribution. If C represents individual union preference, Assumption 2.2 is to replace C with its linear projection onto the time-invariant observed variables, the initial union condition, and the marital status indicators, and the projection error V. Assumption 2.3 states that (a) the union decision is independent of the initial union condition, and the marital status indicators conditional on the current martial status and the previous union decision; (b) given the initial union condition, and the marital status indicators, individual union preference is independent of the time-varying covariates such as the current martial status and the previous union decision. Assumptions 2.4 and 2.5 require that the conditional densities $f_{Y_t|X_t,Y_{t-1},\overline{W}}(y_t|x_t, y_{t-1}, \overline{w})$ and $f_{Y_t|X_t,Y_{t-1},C}(y_t|x_t, y_{t-1}, c;\theta)$ satisfy regular conditions, which are reasonable for this application. Assumption 2.6 is an identification condition for the semiparametric probit model and it makes sure that we have sufficient information to distinguish between alternative population structures.

The estimated coefficients are shown in Table 5. To facilitate comparisons with the conventional approach, we also consider the fully parametric approach which assumes the normality of V. The estimated coefficients in the two methods exhibit the same sign but with different magnitudes. The coefficient for the martial status is 14.9% and 9.5% in the conventional and the two-step methods, respectively. However, both are statistically insignificant. The coefficients on the previous union decision is much lager than the coefficients on the initial union decision. According to Heckman (1978, 1981a,b), the estimation results show that there exists large "true" state dependence and small “spurious" state dependence.

To obtain the magnitude of the APEs for a binary covariate, we adopt the following definition of APEs.

$$APE(\text{Married}_t) = \int_{\mathscr{C}} \Big(\text{Prob}(\text{Union}_t = 1|\text{Married}_t = 1, \text{Union}_{t-1}, D_{1982}, C) \\ - \text{Prob}(\text{Union}_t = 1|\text{Married}_t = 0, \text{Union}_{t-1}, D_{1982}, C)\Big) f_C(c)dc. \tag{29}$$

A similar definition can be applied to Union_{t-1}. The estimation results for the APEs are in Table 6. While the estimated average partial effects for $\text{Married}_t = 0$ are small, the estimated APEs for Union_{t-1} are large. This indicates that union membership has a strong state dependence. The conventional approach gives much higher estimate than the proposed two-step approach (54.4% v.s. 38.6%). Finally, we use the Hausman-type test in Section 4.1 to test the null hypothesis that $V \sim N(0, \sigma_0^2)$. The test statistics is $\widehat{S}_N = 83.402$, which strongly rejects the null hypothesis when we obtain the critical values from $\chi^2(6)$.

7. Conclusion

This paper addresses unsolved issues of the functional form misspecification of the random effect approach for nonlinear panel data models for 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 and utilize the Mundlak-type specification. Then, the average likelihood takes the form of the convolution of the proposed 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 proposed nonlinear panel data models. That is, the average likelihood is correctly specified.

A semi-parametric two-step maximum likelihood estimator is provided. The first step is to estimate a density of observables using kernel density estimator. Based on the identification result, we can use the kernel density estimator to construct the parametric conditional distribution of the unobserved heterogeneity given covariates. In the second step, we devise a maximum likelihood estimator that takes the form of an average likelihood estimator with the proposed known panel data model and the parametric distribution. Under appropriate regularity conditions for the semi-parametric two-step estimator, the estimator is root N consistent and asymptotically normal. Because the semi-parametric two-step maximum likelihood estimator relies on the first step nonparametric kernel density estimate for a density of observables, the estimator can be interpreted as a data-driven procedure. The study of the data-driven procedure in this context is novel and enables an internally consistent random effect approach of common nonlinear panel data models. The identification strategy can also be applied to dynamic nonlinear panel data models.

Appendix

A. Additional Assumptions and Auxiliary Lemmas

In this section, we provide additional assumptions and some auxiliary Lemmas that will be used later in the proofs.

A.1. Asymptotics of $\hat{f}_{X\overline{W}}$

Assumption A.1. *(Kernel) $K(u)$ is a kernel of order s, is symmetric around zero, is equal to zero outside $\prod_{j=1}^{TK+K_1}[-1,1]$, integrates to 1 and is continuously differentiable.*

Assumption A.2. *(Bandwidths) The bandwidth δ satisfies that $\delta \to 0$, $\delta^s = o_p(N^{-1/4})$ and $\log(N)/(N\delta^{TK+K_1}) = o_p(N^{-1/2})$.*

Assumption A.3. *(Smoothness) $f_{X\overline{W}}(x,\overline{w})$ is s-times continuously differentiable on the support of $\mathscr{X} \times \overline{\mathscr{W}}$.*

Lemma A.1. *Suppose that Assumptions A.1, A.2 and A.3 hold. Also, assume that $\mathscr{X}$ and $\overline{\mathscr{W}}$ are both compact sets. Then we have*

$$\sup_{x\in\mathscr{X},\bar{w}\in\mathscr{W}} |\hat{f}_{X\overline{W}}(x,\bar{w}) - f_{X\overline{W}}(x,\bar{w})| = o_p(N^{-1/4}) \tag{A.1}$$

and

$$\left(\frac{-1}{\hat{f}_{xw}(X_j,\overline{W}_j)} - \frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}\right) = \frac{1}{f^2_{X\overline{W}}(X_{tj},\overline{W}_j)} \frac{1}{N}\sum_{\ell=1}^{N}\left(K_h\left(X_\ell - X_j, \overline{W}_\ell - \overline{W}_j\right) - f_{X\overline{W}}(X_j,\overline{W}_j)\right) + o_p(N^{-1/2}). \tag{A.2}$$

The proof for Eq. (A.1) in Lemma A.1 follows the standard arguments for the uniformity of the kernel estimation, so we omit the proof for brevity. The proof for Eq. (A.2) follows by a second order mean expansion and the result in Eq. (A.1), so we omit the proof for brevity too.

We will use U-statistic theory to show our results and we will apply the Theorem A.1 of Ghosal, Sen, and van der Vaart (2000). Here, we give a version of it and we introduce some notations first. Suppose we have i.i.d. observations $X_1,\ldots,X_N$ that are vector-valued and take values in a sample space $\mathscr{X}$. Let f be a function symmetric in its arguments that maps $\mathscr{X}^m$ to real line for some $m \geq 2$. Define

$$\mathscr{V}_{N,m} = \frac{1}{\binom{N}{m}} \sum_{(i_1,\ldots,i_m)\in\mathscr{T}_m^N} f(X_{i_1},\ldots,X_{i_m}),$$

where $\mathscr{T}_m^N$ denotes the set of all combinations of m numbers from $\{1,\ldots,N\}$. Define

$$\mathscr{V}_{N,1} = \frac{m}{N}\sum_{i=1}^{N} E[f(X_i, X_{-i_2}\ldots,X_{-i_m})|X_i]$$

where $(-i_2,\ldots,-i_m)$ denotes a combination of $m-1$ numbers from $\{1,\ldots,i-1,i+1,\ldots,N\}$. Let $\mu_f = E[f(X_1,X_2\ldots,X_m)]$. Let $\mathscr{F}$ denote a class of functions with the envelope function F.

Lemma A.2. *There exists a constant C depending only on m such that*

$$E\Big[\sup_{f\in\mathscr{F}}\big|\mathscr{V}_{N,m} - \mu_f - (\mathscr{V}_{N,1} - m\cdot\mu_f)\big|\Big] \le CN^{-1}\sqrt{E[F^2]}\int_0^1 \sup_{\{Q:\ discrete\ measure\}} \log N(\epsilon\|F\|_{Q,2},\mathscr{F},L_2(Q))d\epsilon,$$

where $N(\cdot)$ denotes the covering number.

Note that Lemma A.2 follows from Theorem A.1 of Ghosal, Sen, and van der Vaart (2000) and the discussions after their Lemma A.2.

Let the class of function $\mathscr{F}\equiv\{f(X,\alpha):\ \alpha\in\mathscr{A}\}$ where $\mathscr{A}$ is a bounded subset of R^k. We say $f(X,\alpha)$ is Lipschitz if

$$|f(X,\alpha_1) - f(X,\alpha_2)| \le B(X)\cdot\|\alpha_1-\alpha_2\| \ \text{ for all } \alpha_1,\alpha_2\in\mathscr{A},$$

for some $B(\cdot):\mathscr{X}\to R$. The following lemma that is based on Theorem 2 of Andrews (1994b) will be used to bound the covering number of a class of functions that is Lipschitz.

Lemma A.3. *Suppose that $f(X,\alpha)$ for $\alpha\in\mathscr{A}$ is Lipschitz and $\mathscr{A}$ is a bounded subset of R^k. Then*

$$\int_0^1 \sup_{\{Q:\ discrete\ measure\}} \log N(\epsilon\|F\|_{Q,2},\mathscr{F},L_2(Q))d\epsilon < \infty$$

with $F(\cdot) = 1\vee\sup_{\alpha\in\mathscr{A}}|f(\cdot,\alpha)|\vee B(\cdot)$ where $a\vee b\equiv\max\{a,b\}$.

We impose further conditions on $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ that will be needed for our proofs based on U-statistics later.

Assumption A.4. *For $t=1,\ldots,T$, for some $0<M<\infty$, and for all $\theta\in\Theta$,*
(i) the density $|f_{Y_t|X_t,C}(y_t|x_t,c;\theta)|\le M$;
(ii) $\|\partial f_{Y_t|X_t,C}(y_t|x_t,c;\theta)/\partial\theta\|\le M$;
(iii) $\|\partial^2 f_{Y_t|X_t,C}(y_t|x_t,c;\theta)/\partial\theta\partial\theta'\|\le M$ for some $M>0$.

B. Proof of Theorem 2.1

Consider

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

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

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

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

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

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

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

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

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

for $j = 1, ..., K_1$,

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

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. (B.1).[10] For example, if we consider the simplest case, $\lambda_0 = (\lambda_{01}, ..., \lambda_{0K_1})^T = (-1, 0, ..., 0)'$, Eq. (B.4) becomes

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

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

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

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

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

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

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

$$\text{(B.8)} \qquad \phi_v(\xi) = \frac{\int_{\mathscr{Y}\times\mathscr{X}\times\overline{\mathscr{W}}} h(\xi, y, x, \overline{w}; \lambda_0) \Omega(y, x, \overline{w}) dy dx d\overline{w}}{\int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0) \Omega(y,x) dy dx dc}$$

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

and (B.8) is equivalent to

$$\text{(B.9)} \qquad \phi_v(\xi) = \frac{\frac{1}{K_1}\sum_{j=1} E\left[\lambda_{0j} e^{-i\xi\overline{W}\lambda_0}\Omega(Y,X,\overline{W})\frac{-1}{f_{X\overline{W}}(X,\overline{W})}\right]}{\Delta_1(\theta_0,\xi)}$$

where $\Delta_1(\theta,\xi) = \int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta_0)\Omega(y,x)dydxdc$. Note that we have used the fact that $\int_{\mathscr{X}} g(x)dx = E[g(X)/f_x(X)]$ where $\mathscr{X}$ is the support of a random variable X and f_x is the pdf of X given that the integral exists.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. (B.9) to all $\alpha \in \mathscr{A}$ to obtain a potential parametric family of functions connected to the characteristic functions of the distribution of the remainder term v in the following form

$$\text{(B.10)} \qquad \phi_{v;\alpha}(\xi) \equiv \frac{\frac{1}{K_1}\sum_{j=1} E\left[\lambda_j e^{-i\xi\overline{W}\lambda}\Omega(Y,X,\overline{W})\frac{-1}{f_{X\overline{W}}(X_t,\overline{W})}\right]}{\Delta_1(\theta,\xi)},$$

where $\phi_{v;\alpha_0}(\xi) = \phi_v(\xi)$. Notice that the terms in the numerator and denominator of the fraction in Eq. (B.9) 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 $f_{Y|X,C}(y|x,c;\theta)$.

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

$$\text{(B.11)} \qquad f_{V;\alpha}(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi v}\phi_{v;\alpha}(\xi)d\xi.$$

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

$$\text{(B.12)} \qquad \sup_{x\in\mathbb{R}^n} |x^a D^b g(x)| = c_{a,b}(g) < \infty,$$

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

$$f(x) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{ix\cdot\xi} \widehat{f}(\xi) d\xi. \tag{B.13}$$

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

$$c_\alpha(\overline{w}) \equiv \int_{\mathscr{C}} f_{v;\alpha}(c - \overline{w}\lambda) dc = \int_{-\infty}^{\infty} \left(e^{i\xi\overline{w}\lambda} \phi_{v;\alpha}(\xi)\right) \left(\frac{1}{2\pi} \int_{\mathscr{C}} e^{-i\xi c} dc\right) d\xi. \tag{B.14}$$

Note that when $\mathscr{C} = \mathbb{R}$, the last term of the integrand becomes $\frac{1}{2\pi} \int_{\mathscr{C}} e^{-i\xi c} dc = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\xi c} dc = \delta(\xi)$ where $\delta(\xi)$ is the Dirac delta function and it is the Fourier transform of a constant function. The important property of the delta function is that $\int f(\xi)\delta(\xi) dt = f(0)$ for all continuous compactly supported functions $f(\cdot)$. With this property, if $\phi_{v;\alpha}(\xi)$ is continuous compactly supported then Eq. (B.14) can be further reduced as

$$c_\alpha(\overline{w}) = \int_{-\infty}^{\infty} \underbrace{\left(e^{i\xi\overline{w}\lambda} \phi_{v;\alpha}(\xi)\right)}_{\text{a function of } \xi} \underbrace{\left(\frac{1}{2\pi} \int_{\mathscr{C}} e^{-i\xi c} dc\right)}_{\psi_c(\xi)} d\xi = \phi_{v;\alpha}(0). \tag{B.15}$$

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

$$f_{C|\overline{W}}(c|\overline{w};\alpha) = \frac{1}{c_\alpha(\overline{w})} f_{v;\alpha}(c - \overline{w}\lambda) \tag{B.16}$$

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

$$\begin{aligned} c_\alpha(\overline{w}) &= \int_{-\infty}^{\infty} \frac{e^{i\xi\overline{w}\lambda}}{\Delta_1(\theta,\xi)} E\Big[\frac{1}{K_1} \sum_{j=1}^{K_1} \lambda_j e^{-i\xi\overline{W}\lambda} \Omega(Y,X,\overline{W}) \frac{-1}{f_{X\overline{W}}(X,\overline{W})}\Big] \psi_c(\xi) d\xi \\ f_{v;\alpha}(c - \overline{w}\lambda) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{i\xi(\overline{w}\lambda - c)}}{\Delta_1(\theta,\xi)} E\Big[\frac{1}{K_1} \sum_{j=1}^{K_1} \lambda_j e^{-i\xi\overline{W}\lambda} \Omega(Y,X,\overline{W}) \frac{-1}{f_{X\overline{W}}(X,\overline{W})}\Big] d\xi, \end{aligned} \tag{B.17}$$

where

$$\begin{aligned} \Delta_1(\theta,\xi) &= \int_{\mathscr{Y}\times\mathscr{X}\times\mathscr{C}} e^{-i\xi c} f_{Y|X,C}(y|x,c;\theta)\Omega(y,x) dy_t dx_t dc, \\ \psi_c(\xi) &= \left(\frac{1}{2\pi} \int_{\mathscr{C}} e^{-i\xi c} dc\right). \end{aligned}$$

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

Proof: First, by Assumptions 2.2(i)&(ii), 2.3(i)(ii)&(iii), 2.4, and 2.5(ii), $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is well defined. Then, Assumption 2.5(i) & (ii) implies that $f_{C|\overline{W}}(c|\overline{w};\alpha)$ is continuous for all α and is nonnegative for α in some open neighborhood of α_0. With the definition of $c_\alpha(\overline{w})$ in Eq. (B.14), we can obtain the integration of $f_{C|\overline{W}}(c|\overline{w};\alpha)$ over the domain $\mathscr{C}$ is equal to one. *Q.E.D.*

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

$$f(y|x,\overline{w};\alpha) = \int_{\mathscr{C}} f_{Y|X,C}(y|x,c;\theta) f_{C|\overline{W}}(c|\overline{w};\alpha)dc. \tag{B.18}$$

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

$$\begin{aligned}\int_{\mathscr{Y}} f(y|x,\overline{w};\alpha)dy_t &= \int_{\mathscr{C}} \left(\int_{\mathscr{Y}_t} f_{Y|X,C}(y|x,c;\theta)dy_t \right) f_{C|\overline{W}}(c|\overline{w};\alpha)dc \\ &= \int_{\mathscr{C}} f_{C|\overline{W}}(c|\overline{w};\alpha)dc = 1.\end{aligned} \tag{B.19}$$

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

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

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

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

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

Differentiating Eq. (B.19) 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. (11), we have $\frac{\partial}{\partial \alpha} K(\alpha; x, \overline{w})\big|_{\alpha=\alpha_0} = 0$. Similarly, differentiating Eq. (B.19) twice and applying the result to the second derivative of $K(\alpha; x, \overline{w})$, the matrix of the second derivative can be written as minus outer product of the gradient of the log likelihood:

$$K''(\alpha_0; x, \overline{w}) = -\mathrm{E}\Big[\frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)\big|_{\alpha=\alpha_0} \cdot \frac{\partial}{\partial \alpha} \log f(Y|X, \overline{W}; \alpha)'\big|_{\alpha=\alpha_0} \Big| X = x, \overline{W} = \overline{w}\Big]. \tag{B.20}$$

Proof of Theorem 2.1: First we have discussed that the parametric density function of observable variables in Eq. (B.19) is well defined using Assumptions 2.2(i)&(ii), 2.3(i)(ii)&(iii), 2.4, and 2.5(i)&(ii). We next proceed to prove the result using concavity of conditional Kullback-Leibler information criterion, i.e., the second derivative of $K(\alpha; x, \overline{w})$ in Eq. (12) is negative definite. *Q.E.D.*

B.1. Summary of Identification Steps

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

There are four steps toward the construction of the internally consistent average likelihood function and we start with the proposed 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. (B.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. (B.7) and then extend to relationship to obtain the parametric function in Eq. (B.10).

Step 3: Apply the inverse Fourier transform.

Under Assumption 2.5, the inverse Fourier transform is applicable and we can recover the parametric distribution of the unobserved heterogeneity in Eq. (B.16) using the inverse transform and normalization.

Step 4: Construct an internally consistent average likelihood.

We can then combine the parametric distribution of the unobserved heterogeneity in Step 3 with the proposed panel data models to obtain the internally consistent average likelihood Eq. (B.18).

C. Proof of Theorem 3.1

In this section, we derive the asymptotic normality of the proposed estimators. Without loss of generality, we assume that $K = 1$, so $\lambda = \lambda_1$ is a scalar. We also assume that θ is a scalar. For notational simplicity, we write $D_i = (Y_i, X_i, \overline{W}_i)$ when there will be no confusion. We also assume that $\Omega(y,x,\bar{w}) = g_y(y)g_x(x)\cdot \Pi_{k=1}^{K} g_{\bar{w}_k}(\bar{w}_k)$, where g_y, g_x and $g_{\bar{w}_k}$ are PDFs.

C.1. Log-likelihood Function

We first derive the log-likelihood function that we use to estimate. Recall that $f(Y_i|X_i,C_i;\theta) = \Pi_{t=1}^{T} f(Y_{it}|X_{it},C_i;\theta)$. Note that

$$
\begin{aligned}
\log f(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}}) &= \log \int_{\mathscr{C}} f_{Y|X,C}(Y_i|X_i,c;\theta) f_{C|\overline{W}}(c|\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})dc \\
&= \log\Big(\frac{1}{c_{\alpha,\hat{f}_{y|xw}}(\overline{W}_i)} \int_{\mathscr{C}} f_{Y|X,C}(Y_i|X_i,c;\theta)\hat{f}_{v;\alpha,\hat{f}_{X\overline{W}}}(c-\bar{w}\lambda)dc\Big) \\
&= -\log c_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) + \log \int_{\mathscr{C}} f_{Y|X,C}(Y_i|X_i,c;\theta) f_{v;\alpha,\hat{f}_{X\overline{W}}}(c-\bar{w}\lambda)dc \\
&= -\log c_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) + \rho_{\alpha,\hat{f}_{X\overline{W}}}(D_i).
\end{aligned}
$$

We then approximate $c_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)$ and $\rho_{\alpha,\hat{f}_{X\overline{W}}}(D_i)$ by:

$$
\begin{aligned}
\hat{c}_{\alpha,\hat{f}_{X\overline{W}}(\overline{W}_i)} &= \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_{tj},\overline{W}_j)}\Big)\psi_c(\xi)d\xi, \\
\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i) &= \frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{\lambda e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}f_{Y|X,C}(Y_i|X_i,c;\theta)}{\Delta_1(\theta,\xi)}\Big(e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi.
\end{aligned}
$$

C.2. Consistency of $\hat{\alpha}$ for α_0

Asymptotics for $N^{-1}\sum_{i=1}^{N}\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)$:

Note that it is straightforward to see that $\sup_{W_i}\left|\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) - c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)\right| = o_p(n^{-1/4})$, so by a second order mean expansion at $c_{\alpha,f_{y|xw}}(\overline{W}_i)$, we have the following

$$
\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) = \log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i) + \frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) - c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)) + o_p(N^{-1/2}).
$$

Let

$$
\Delta_2(\lambda,\xi) = E\Big[\lambda e^{-i\xi\overline{W}\lambda}\Omega(Y,X,\overline{W})\frac{-1}{f_{X\overline{W}}(X,\overline{W})}\Big].
$$

Then we have

$$
\begin{aligned}
&\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) - c_{\alpha,f_{X\overline{W}}}(\overline{W}_i) \\
&= \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\Big(\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi \\
&\quad + \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}e^{i\xi\overline{W}_i\lambda}\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi.
\end{aligned}
$$

Let $\mu_c(\alpha) = E[\log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)]$ and we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i) - \mu_c(\alpha) \\
=&\frac{1}{N}\sum_{i=1}^{N}\log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i) - \mu_c(\alpha) \\
&+ \frac{1}{n^2}\sum_{i,j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\Big(\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi \\
&+ \frac{1}{n^2}\sum_{i,j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}e^{i\xi\overline{W}_i\lambda}\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi + o_p(N^{-1/2}) \\
=&A_{1N}(\alpha) + A_{2N}(\alpha) + A_{3N}(\alpha) + o_p(N^{-1/2}).
\end{aligned}
$$

We consider $A_{2n}(\alpha)$ term first. Note that by Lemma A.1, we have

$$
\begin{aligned}
&A_{2N}(\alpha) \\
=&\frac{1}{N^3}\sum_{i,j,\ell=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{e^{i\xi\overline{W}_i\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{f^2_{X\overline{W}}(X_j,\overline{W}_j)}\Big(K_h\big(X_\ell - X_j, \overline{W}_\ell - \overline{W}_j\big) - f_{X\overline{W}}(X_j,\overline{W}_j)\Big)\psi_c(\xi)d\xi \\
&\qquad + o_p(N^{-1/2}).
\end{aligned}
$$

Because Assumption A.4 imposes the regularity conditions on the proposed conditional distribution $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ with respect to the parameter θ, an integration of the product of a continuous function of $f_{Y_t|X_t,C}(y_t|x_t,c;\theta)$ and a continuous function not related to θ can be Lipschitz continuous with respect to θ if the integrand is absolutely integrable and this justifies the interchange of integration and differentiation. This implies that

$$
\Big\{\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{e^{i\xi\overline{W}_i\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{f^2_{X\overline{W}}(X_j,\overline{W}_j)}\Big(K_h\big(X_\ell - X_j, \overline{W}_\ell - \overline{W}_j\big) - f_{X\overline{W}}(X_j,\overline{W}_j)\Big)\psi_c(\xi)d\xi\Big|\alpha\in\mathscr{A}\Big\}
$$

is Lipschitz and with an envelope function G_2 satisfying that $N^{-1}\sqrt{E[G_2^2]} = o_p(1)$. Then by Lemma A.2, we have

$$A_{2N}(\alpha) = \frac{1}{N}\sum_{i}^{N}\frac{1}{\Delta_1(\theta,\xi)}E\Big[\frac{e^{i\xi\overline{W}\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W})}\Big]\lambda e^{-i\xi\overline{W}_i\lambda}E[g_y(Y_\ell)|X_\ell,\overline{W}_\ell]\cdot\Omega(X_\ell,\overline{W}_\ell)\frac{1}{f_{X\overline{W}}(X_\ell,\overline{W}_\ell)}\psi_c(\xi)d\xi+\mu_c(\alpha)+o_p(N^{-1/2})$$

because

$$\begin{aligned}
&E\Big[\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{e^{i\xi\overline{W}_i\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{f^2_{X\overline{W}}(X_j,\overline{W}_j)}\Big(K_h\big(X_\ell-X_j,\overline{W}_\ell-\overline{W}_j\big)-f_{X\overline{W}}(X_j,\overline{W}_j)\Big)\psi_c(\xi)d\xi\Big|D_i\Big]=o_p(1),\\
&E\Big[\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{e^{i\xi\overline{W}_i\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{f^2_{X\overline{W}}(X_j,\overline{W}_j)}\Big(K_h\big(X_\ell-X_j,\overline{W}_\ell-\overline{W}_j\big)-f_{X\overline{W}}(X_j,\overline{W}_j)\Big)\psi_c(\xi)d\xi\Big|D_j\Big]=o_p(1),\\
&E\Big[\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{e^{i\xi\overline{W}_i\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{f^2_{X\overline{W}}(X_j,\overline{W}_j)}\Big(K_h\big(X_\ell-X_j,\overline{W}_\ell-\overline{W}_j\big)-f_{X\overline{W}}(X_j,\overline{W}_j)\Big)\psi_c(\xi)d\xi\Big|D_\ell\Big],\\
&=\frac{1}{\Delta_1(\theta,\xi)}E\Big[\frac{e^{i\xi\overline{W}\lambda}}{c_{\alpha,f_{X\overline{W}}}(\overline{W})}\Big]\lambda e^{-i\xi\overline{W}_i\lambda}E[g_y(Y_\ell)|X_\ell,\overline{W}_\ell]\cdot\Omega(X_\ell,\overline{W}_\ell)\frac{1}{f_{X\overline{W}}(X_\ell,\overline{W}_\ell)}\psi_c(\xi)d\xi+\mu_c(\alpha)+o_p(1).
\end{aligned}$$

We consider $A_{3N}(\alpha)$ term now. Similarly, we have

$$\Big\{\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}e^{i\xi\overline{W}_i\lambda}\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_{tj},\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi\Big|\alpha\in\mathscr{A}\Big\}$$

is Lipschitz with an envelop function G_3 satisfying that $N^{-1}\sqrt{E[G_3^2]}=o_p(1)$. Then by Lemma A.2, we have

$$A_{3N}(\alpha)=\frac{1}{N}\sum_{i}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}E\Big[\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W})}e^{i\xi\overline{W}\lambda}\Big]\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_{tj},\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi+o_p(N^{-1/2}),$$

because

$$\begin{aligned}
&E\Big[\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}e^{i\xi\overline{W}_i\lambda}\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi\Big|D_i\Big]=0\\
&E\Big[\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}e^{i\xi\overline{W}_i\lambda}\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi\Big|D_j\Big]\\
&=\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}E\Big[\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W})}e^{i\xi\overline{W}\lambda}\Big]\Big[\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]\psi_c(\xi)d\xi.
\end{aligned}$$

Therefore, it is sufficient to show that

$$\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^{N}\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)-\mu_c(\alpha)\Big|=o_p(1).$$

Asymptotics for $N^{-1}\sum_{i=1}^{N}\log\hat{\rho}_{\alpha\hat{f}_{X\overline{W}}}(D_i)$:

Similar to $\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}$ case, we have

$$\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)=\log\rho_{\alpha,f_{X\overline{W}}}(D_i)+\frac{1}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}(\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)-\rho_{\alpha,f_{X\overline{W}}}(D_i))+o_p(N^{-1/2}).$$

Also, let $\mu_\rho(\alpha)=E[\log\rho_{\alpha,f_{X\overline{W}}}(D_i)]$ and we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)-\mu_\rho(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^N\log\rho_{\alpha,f_{X\overline{W}}}(D_i)-\mu_\rho(\alpha)\\
&+\frac{1}{2\pi N^2}\sum_{i,j=1}^N\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{\lambda e^{-i\xi(c-\overline{w}_i\lambda)}f_{Y|X,C}(y_i|x_i,c;\theta)}{\Delta_1(\theta,\xi)\rho_{\alpha,f_{X\overline{W}}}(D_i)}\Big(e^{-i\xi\overline{w}_j\lambda}\Omega(D_j)\Big(\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\Big)dcd\xi\\
&+\frac{1}{2\pi N^2}\sum_{i,j=1}^N\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{\lambda e^{-i\xi(c-\overline{w}_i\lambda)}f_{Y|X,C}(y_i|x_i,c;\theta)}{\Delta_1(\theta,\xi)\rho_{\alpha,f_{X\overline{W}}}(D_i)}\Big[e^{-i\xi\overline{w}_j\lambda}\Omega(D_j)\frac{-1}{f_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big]dcd\xi.
\end{aligned}$$

Similar to $\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}$ case, we have, we have $\sup_{\alpha\in\mathscr{A}}|N^{-1}\sum_{i=1}^N\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)-\mu_\rho(\alpha)|=o_p(1)$.

Consistency of $\hat{\alpha}$:

Let $Q(\alpha)=E[\log f(Y|X,\overline{W};\alpha,f_{X\overline{W}}))]$. Consistency of $\hat{\alpha}$ for α_0 holds following the fact that

$$\begin{aligned}
&\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N\log\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})-Q(\alpha)\Big|\\
\le&\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N\log\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})-\log f(Y_i|X_i,\overline{W}_i;\alpha,f_{X\overline{W}})\Big|\\
&+\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N\log\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})-Q(\alpha)\Big|\\
\le&\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N-\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)+\log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)\Big|+\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)-\log\rho_{\alpha,f_{X\overline{W}}}(D_i)\Big|\\
&+\sup_{\alpha\in\mathscr{A}}\Big|\frac{1}{N}\sum_{i=1}^N\log\hat{f}(Y_i|X_i,\overline{W}_i;\alpha,\hat{f}_{X\overline{W}})-Q(\alpha)\Big|=o_p(1),
\end{aligned}$$

and α_0 is the unique minimizer for $\arg\max_{\alpha\in\mathscr{A}}Q(\alpha)$. Let $E\big[\int_{-\infty}^{\infty}g_{\partial c,\theta}(\overline{W}_i,\xi)\Delta_2(\lambda,\xi)\psi_c(\xi)d\xi\big]=\mu_{g,\partial c,\theta}(\alpha)$.

C.3. Asymptotic Normality

Asymptotics for $n^{-1}\sum_{i=1}^{N}\partial\log(\hat{c}_{\alpha,\hat{f}_{y|xw}}(\overline{W}_i))/\partial\theta$:

Taking derivative of $\log\left(\hat{c}_{\alpha,\hat{f}_{y|xw}}(\overline{W}_i)\right)$ with respect to θ gives:

$$\frac{\partial\log\left(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)\right)}{\partial\theta}=\frac{1}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}\frac{\partial\left(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)\right)}{\partial\theta}$$
$$=\frac{1}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{-\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)}\left(e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\right)\psi_c(\xi)d\xi.$$

Note that

$$\frac{1}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}$$
$$=\frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}-\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)}\left(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\right)\psi_c(\xi)d\xi+o_p(N^{-1/2}),$$
$$\frac{\partial\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}{\partial\theta}$$
$$=\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}{\partial\theta}+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}e^{i\xi\overline{W}_i\lambda}\frac{-\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)}\left(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\right)\psi_c(\xi)d\xi+o_p(N^{-1/2}),$$

so

$$\frac{\partial\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}$$
$$=\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}$$
$$+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}e^{i\xi\overline{W}_i\lambda}\left(-\frac{(\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta)^2}{c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)\Delta_1(\theta,\xi)}-\frac{\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\right)\left(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\right)\psi_c(\xi)d\xi$$
$$+o_p(N^{-1/2})$$
$$\equiv\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}g_{\partial c,\theta}(\overline{W}_i,\xi)\left(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\right)\psi_c(\xi)d\xi+o_p(N^{-1/2})$$

where

$$g_{\partial c,\theta}(\overline{W}_i,\xi)=e^{i\xi\overline{W}_i\lambda}\left(-\frac{(\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta)^2}{c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)\Delta_1(\theta,\xi)}-\frac{\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\right).$$

Let $\mu_{\partial \log c,\theta}(\alpha) = E\left[\partial \log(c_{\alpha,f_{X\overline{W}}}(\overline{W}))/\partial\theta\right]$. Then

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \log(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i))}{\partial\theta} - \mu_{\partial\log c,\theta}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)} - \mu_{\partial\log c,\theta}(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)} - \mu_{\partial\log c,\theta}(\alpha)\Big)\\
&+\frac{1}{N^2}\sum_{i,j=1}^{N}\int_{-\infty}^{\infty} g_{\partial c,\theta}(\overline{W}_i,\xi)\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi + o_p(N^{-1/2}).
\end{aligned}
$$

Let $E\left[\int_{-\infty}^{\infty} g_{\partial c,\theta}(\overline{W}_i,\xi)\Delta_2(\lambda,\xi)\psi_c(\xi)d\xi\right] = \mu_{g,\partial c,\theta}(\alpha)$. By the same arguments for $A_{2n}(\alpha)$ and $A_{3n}(\alpha)$, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \log(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i))}{\partial\theta} - \mu_{\log c,\theta}(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\theta}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)} - \mu_{\partial\log c,\theta}(\alpha)\\
&+\frac{1}{N}\sum_{i=1}^{N}\Big(\int_{-\infty}^{\infty} E[g_{\partial c,\theta}(\overline{W}_i,\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i) - E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{-1}{f_{X\overline{W}}(X_i,\overline{W}_i)}\psi_c(\xi)d\xi\Big)\\
&+\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} g_{\partial c,\theta}(\overline{W}_i,\xi)\Delta_2(\lambda,\xi)\psi_c(\xi)d\xi - \mu_{g,\partial c,\theta}(\alpha) + o_p(N^{-1/2})\\
\equiv&\frac{1}{N}\sum_{i=1}^{N}\psi_{\partial\log c,\theta}(D_i) + o_p(N^{-1/2}).
\end{aligned}
$$

Asymptotics for $N^{-1}\sum_{i=1}^{N}\partial\log(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i))/\partial\lambda$:

Taking derivative of $\hat{c}_{\alpha,\hat{f}_{y|xw}}(\overline{W}_i)$ with respect to λ gives:

$$
\begin{aligned}
&\frac{\partial\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda}\\
=&\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big(i\xi\overline{W}_i e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi\\
=&\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big((i\xi\overline{W}_i+\lambda^{-1})e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)\psi_c(\xi)d\xi.
\end{aligned}
$$

Note that

$$\frac{\partial \hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda} = \frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda} + \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}(i\xi\overline{W}_i+\lambda^{-1})e^{i\xi\overline{W}_i\lambda}\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi$$
$$+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{1}{\Delta_1(\theta,\xi)}e^{i\xi\overline{W}_i\lambda}\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_3(\lambda,\xi)\Big)\psi_c(\xi)d\xi.$$

where

$$\Delta_3(\lambda,\xi) = E\Big[i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}\lambda}\Omega(Y_t,X_t,\overline{W})\frac{1}{f_{X\overline{W}}(X,\overline{W})}\Big].$$

Note that

$$\frac{1}{\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)} = \frac{1}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}$$
$$-\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\lambda}{c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)}\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi + o_p(N^{-1/2}).$$

Therefore,

$$\begin{aligned}
&\frac{\partial\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda}\\
=&\frac{\partial\log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda} + \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{(i\xi\overline{W}_i+\lambda^{-1})e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{xw}(X_j,\overline{W}_j)} - \Delta_3(\lambda,\xi)\Big)\psi_c(\xi)d\xi\\
&-\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{i\xi\overline{W}_i\lambda}(\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\lambda)^2}{\Delta_1(\theta,\xi)c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}\Big(e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi + o_p(N^{-1/2})\\
=&\frac{\partial\log c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}{\partial\lambda} + \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}g_{2,\partial c,\lambda}(\overline{W}_i,\xi)\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}g_{3,\partial c,\lambda}(\overline{W}_i,\xi)\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)} - \Delta_3(\lambda,\xi)\Big)\psi_c(\xi)d\xi + o_p(N^{-1/2})
\end{aligned}$$

where

$$g_{2,\partial c,\lambda}(\overline{W}_i,\xi) = \frac{(i\xi\overline{W}_i+\lambda^{-1})e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)} - \frac{\lambda^{-1}e^{i\xi\overline{W}_i\lambda}\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\lambda}{\Delta_1(\theta,\xi)c^2_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}$$
$$g_{3,\partial c,\lambda}(\overline{W}_i,\xi) = \frac{e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)}.$$

Define

$$\mu_{\partial \log c,\lambda}(\alpha) = E\big[\partial \log(c_{\alpha,f_{X\overline{W}}}(\overline{W}))/\partial\lambda\big],$$
$$\mu_{g_2,\partial c,\lambda}(\alpha) = E\Big[\int_{-\infty}^{\infty} g_{2,\partial c,\lambda}(\overline{W},\xi)\Delta_2(\lambda,\xi)\psi_c(\xi)d\xi\Big],$$
$$\mu_{g_3,\partial c,\lambda}(\alpha) = E\Big[\int_{-\infty}^{\infty} g_{3,\partial c,\lambda}(\overline{W},\xi)\Delta_3(\lambda,\xi)\psi_c(\xi)d\xi\Big].$$

By the same arguments before we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \log(\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i))}{\partial\lambda} - \mu_{\partial\log c,\lambda}(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\partial c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)/\partial\lambda}{c_{\alpha,f_{X\overline{W}}}(\overline{W}_i)} - \mu_{\partial\log c,\lambda}(\alpha)\Big)\\
&+\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} E[g_{2,\partial c,\lambda}(\overline{W},\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i)-E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{-1}{f_{X\overline{W}}(X_i,\overline{W}_i)}\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} g_{2,\partial c,\lambda}(\overline{W}_i,\xi)\Delta_2(\lambda,\xi)\psi_c(\xi)d\xi - \mu_{g_2,\partial c,\lambda}(\alpha)\\
&+\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} E[g_{3,\partial c,\lambda}(\overline{W},\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i)-E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{1}{f_{X\overline{W}}(X_i,\overline{W}_i)}\psi_c(\xi)d\xi\\
&+\frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} g_{3,\partial c,\lambda}(\overline{W}_i,\xi)\Delta_3(\lambda,\xi)\psi_c(\xi)d\xi - \mu_{g_3,\partial c,\lambda}(\alpha) + o_p(N^{-1/2})\\
\equiv&\frac{1}{N}\sum_{i=1}^{N}\psi_{\partial\log c,\lambda}(D_i) + o_p(N^{-1/2}).
\end{aligned}$$

Asymptotics for $n^{-1}\sum_{i=1}^{n}\partial\log(\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i))/\partial\theta$:

Recall that

$$\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i) = \frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)}\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi.$$

First,

$$\begin{aligned}
&\frac{\partial\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial\theta}\\
=&\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}} e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}\Big(\frac{\partial f_{Y|X,C}(Y_i|X_i,c;\theta)/\partial\theta}{\Delta_1(\theta,\xi)} - f_{Y|X,C}(Y_i|X_i,c;\theta)\frac{\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)}\Big)\\
&\cdot\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi.
\end{aligned}$$

Note that

$$
\begin{aligned}
&\frac{1}{\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}\\
=&\frac{1}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}-\frac{\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\theta}{\rho^2_{\alpha,f_{X\overline{W}}}(D_i)}\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)}{\Delta_1(\theta,\xi)}\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_i)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big)dcd\xi,\\
&\frac{\partial\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial\theta}\\
=&\frac{\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)}{\partial\theta}+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}\Big(\frac{\partial f_{Y|X,C}(Y_i|X_i,c;\theta)/\partial\theta}{\Delta_1(\theta,\xi)}-f_{Y|X,C}(Y_i|X_i,c;\theta)\frac{\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)}\Big)\\
&\qquad\cdot\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big)dcd\xi.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{\partial\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)/\partial\theta}{\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}\\
=&\frac{\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\theta}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{\partial\rho,\theta}(D_i,c,\xi)\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big)dcd\xi+o_p(N^{-1/2})
\end{aligned}
$$

where

$$
\begin{aligned}
&g_{\partial\rho,\theta}(D_i,c,\xi)\\
=&e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}\Big(\frac{1}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}\Big(\frac{\partial f_{Y|X,C}(Y_i|X_i,c;\theta)/\partial\theta}{\Delta_1(\theta,\xi)}-f_{Y|X,C}(Y_i|X_i,c;\theta)\frac{\partial\Delta_1(\theta,\xi)/\partial\theta}{\Delta_1^2(\theta,\xi)}\Big)\\
&-\frac{(\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\theta)^2}{\rho^2_{\alpha,f_{X\overline{W}}}(D_i)}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)}{\Delta_1(\theta,\xi)}\Big).
\end{aligned}
$$

Let $\mu_{\log\rho,\theta}(\alpha)=E\big[\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\theta\big]$ and $\mu_{g,\partial\rho,\theta}(\alpha)=E\big[\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{\partial\rho,\theta}(D_i,c,\xi)\Delta_2(\lambda,\xi)dcd\xi\big]$. Then

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\log(\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial\theta}-\mu_{\log\rho,\theta}(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\theta}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}-\mu_{\log\rho,\theta}(\alpha)\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\Big(\int_{-\infty}^{\infty}\int_{\mathscr{C}}E[g_{\partial\rho,\theta}(D_i,c,\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i)-E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{-1}{f_{X\overline{W}}(X_i,\overline{W}_i)}dcd\xi\Big)\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{\partial\rho,\theta}(D_i,c,\xi)\Delta_2(\lambda,\xi)dcd\xi-\mu_{g,\partial\rho,\theta}(\alpha)+o_p(N^{-1/2})\\
\equiv&\frac{1}{N}\sum_{i=1}^{N}\psi_{\partial\log\rho,\theta}(D_i)+o_p(N^{-1/2}).
\end{aligned}
$$

Asymptotics for $N^{-1}\sum_{i=1}^{N}\partial\log(\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i))/\partial\lambda$:

Taking derivative of $\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)$ with respect to λ gives:

$$
\begin{aligned}
&\frac{\partial \hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial \lambda} \\
=&\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{-i\xi c}(i\xi\overline{W}_i+\lambda^{-1})}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi \\
&+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)}\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}\Big)dcd\xi.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\frac{\partial \hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial \lambda}=&\frac{\partial \rho_{\alpha,f_{X\overline{W}}}(D_i)}{\partial \lambda} \\
&+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{-i\xi c}(i\xi\overline{W}_i+\lambda^{-1})}{\Delta_1(\theta,\xi)}\Big(e^{i\xi\overline{W}_i\lambda}\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big)dcd\xi \\
&+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{-i\xi c}e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)}\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_3(\lambda,\xi)\Big)dcd\xi.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{\partial \log \hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)}{\partial \lambda} \\
=&\frac{\partial \log \rho_{\alpha,f_{X\overline{W}}}(D_i)}{\lambda}+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{2,\partial\rho,\lambda}(\overline{W}_i,c,\xi)\Big(\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{-1}{\hat{f}_{X\overline{W}}(X_{tj},\overline{W}_j)}-\Delta_2(\lambda,\xi)\Big)\psi_c(\xi)d\xi \\
&+\frac{1}{2\pi N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{3,\partial\rho,\lambda}(\overline{W}_i,c,\xi)\Big(i\xi\overline{W}_j\lambda e^{-i\xi\overline{W}_j\lambda}\Omega(D_j)\frac{1}{\hat{f}_{X\overline{W}}(X_j,\overline{W}_j)}-\Delta_3(\lambda,\xi)\Big)\psi_c(\xi)d\xi
\end{aligned}
$$

where

$$
\begin{aligned}
g_{2,\partial\rho,\lambda}(D_i,c,\xi)&=\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)(i\xi\overline{W}_i+\lambda^{-1})e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)\rho_{\alpha,f_{X\overline{W}}}(D_i)}-\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)\lambda^{-1}e^{i\xi\overline{W}_i\lambda}\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\lambda}{\Delta_1(\theta,\xi)\rho^2_{\alpha,f_{X\overline{W}}}(D_i)} \\
g_{3,\partial\rho,\lambda}(D_i,c,\xi)&=\frac{f_{Y|X,C}(Y_i|X_i,c;\theta)e^{i\xi\overline{W}_i\lambda}}{\Delta_1(\theta,\xi)\rho_{\alpha,f_{X\overline{W}}}(D_i)}.
\end{aligned}
$$

Define

$$
\begin{aligned}
\mu_{\log\rho,\lambda}(\alpha)&=E\big[\partial\log(\rho_{\alpha,f_{X\overline{W}}}(D))/\partial\lambda\big], \\
\mu_{g_2,\partial\rho,\lambda}(\alpha)&=E\Big[\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{2,\partial\rho,\lambda}(D,c,\xi)\Delta_2(\lambda,\xi)dcd\xi\Big], \\
\mu_{g_3,\partial\rho,\lambda}(\alpha)&=E\Big[\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{3,\partial\rho,\lambda}(D,c,\xi)\Delta_3(\lambda,\xi)dcd\xi\Big].
\end{aligned}
$$

By the same arguments before we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \log(\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i))}{\partial\lambda}-\mu_{\log\rho,\lambda}(\alpha)\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\partial\rho_{\alpha,f_{X\overline{W}}}(D_i)/\partial\lambda}{\rho_{\alpha,f_{X\overline{W}}}(D_i)}-\mu_{\log\rho,\lambda}(\alpha)\Big)\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}E[g_{2,\partial c,\lambda}(D,c,\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i)-E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{-1}{f_{X\overline{W}}(X_i,\overline{W}_i)}dcd\xi\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{2,\partial\rho,\lambda}(D_i,c,\xi)\Delta_2(\lambda,\xi)dcd\xi-\mu_{g_{2,\partial c,\lambda}}(\alpha)\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}E[g_{3,\partial\rho,\lambda}(D,c,\xi)]\lambda e^{-i\xi\overline{W}_i\lambda}\big(g_y(Y_i)-E[g_y(Y_i)|X_i,\overline{W}_i]\big)\cdot\Omega(X_i,\overline{W}_i)\frac{1}{f_{X\overline{W}}(X_i,\overline{W}_i)}dcd\xi\\
&+\frac{1}{2\pi N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{\mathscr{C}}g_{3,\partial\rho,\lambda}(D_i,c,\xi)\Delta_3(\lambda,\xi)dcd\xi-\mu_{g_{3,\partial c,\lambda}}(\alpha)+o_p(N^{-1/2})\\
\equiv&\frac{1}{N}\sum_{i=1}^{N}\psi_{\partial\log\rho,\lambda}(D_i)+o_p(N^{-1/2}).
\end{aligned}
$$

Asymptotics of $\sqrt{N}(\hat{\alpha}-\alpha_0)$:

Define

$$
\widehat{H}(\alpha)=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\alpha\partial\alpha'}\log\hat{f}(y_i|x_i,\overline{w}_i;\alpha,\hat{f}_{X\overline{W}})
$$

Following standard arguments, we have

$$
\sqrt{N}(\hat{\alpha}-\alpha_0)=-\widehat{H}^{-1}(\ddot{\alpha})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\alpha}\log\hat{f}(y_i|x_i,\overline{w}_i;\alpha_0,\hat{f}_{X\overline{W}})
$$

where $\ddot{\alpha}$ is between $\hat{\alpha}$ and α_0. It is straightforward to show that $\widehat{H}^{-1}(\ddot{\alpha})\xrightarrow{p}H^{-1}$ where H is defined in Assumption 3.1. Also, we have

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\alpha}\log\hat{f}(y_i|x_i,\overline{w}_i;\hat{\alpha},\hat{f}_{X\overline{W}})&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix}\frac{\partial}{\partial\theta}\log\hat{f}(y_i|x_i,\overline{w}_i;\hat{\alpha},\hat{f}_{X\overline{W}})\\ \frac{\partial}{\partial\lambda}\log\hat{f}(y_i|x_i,\overline{w}_i;\hat{\alpha},\hat{f}_{X\overline{W}})\end{pmatrix}\\
&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix}-\frac{\partial}{\partial\theta}\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)+\frac{\partial}{\partial\theta}\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)\\ -\frac{\partial}{\partial\lambda}\log\hat{c}_{\alpha,\hat{f}_{X\overline{W}}}(\overline{W}_i)+\frac{\partial}{\partial\lambda}\log\hat{\rho}_{\alpha,\hat{f}_{X\overline{W}}}(D_i)\end{pmatrix}\\
&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix}-\psi_{\partial\log c,\theta}(\overline{W}_i)+\psi_{\partial\log\rho,\theta}(D_i)\\ -\psi_{\partial\log c,\lambda}(\overline{W}_i)+\psi_{\partial\log\rho,\lambda}(D_i)\end{pmatrix}+o_p(1).
\end{aligned}
$$

It follows that

$$\sqrt{N}(\hat{\alpha}-\alpha_0) = -H^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix} -\psi_{\partial\log c,\theta}(\overline{W}_i)+\psi_{\partial\log\rho,\theta}(D_i) \\ -\psi_{\partial\log c,\lambda}(\overline{W}_i)+\psi_{\partial\log\rho,\lambda}(D_i)\end{pmatrix} + o_p(1)$$
$$\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_\alpha(D_i)+o_p(1) \xrightarrow{d} \mathcal{N}(0,\mathcal{V})$$

where $\mathcal{V}$ is the covariance matrix generated by $\psi_\alpha(D_i)$.

References

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.

ANDREWS, D. (1994a): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

ANDREWS, D. W. K. (1994b): "Empirical Process Methods in Econometrics," *in Handbook of Econometrics, ed. by R. F. Engle and D. L. McFadden*, 4, 2247–2294.

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

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

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

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

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

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

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

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

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

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): “Estimation of Semiparametric Models When the Criterion Function is Not Smooth,” *Econometrica*, 71(5), 1591–1608.

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*.

GHOSAL, S., A. SEN, AND A. VAN DER VAART (2000): “Testing Monotonicity of Regression,” *Annals of Statistics*, 28(4), 1054–1082.

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

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

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

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

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

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

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

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

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

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

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

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

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

ICHIMURA, H., AND S. LEE (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-estimators," *Journal of Econometrics*, 159(2), 252–266.

NEWEY, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," *in Handbook of Econometrics, ed. by R. F. Engle and D. L. McFadden*, 4, 2111–2245.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press.

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.

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: Monte Carlo Simulations

	Infeasible Estimator		Conventional Estimator			Two-step Estimator	
Sample size=500	θ	λ	θ	λ	σ	θ	λ
True	0.5	-0.5	0.5	-0.5	1	0.5	-0.5
DGP I:							
Mean	0.500	-0.503	0.501	-0.504	0.915	0.523	-0.531
Median	0.499	-0.512	0.507	-0.519	0.894	0.517	-0.522
RMSE	0.062	0.092	0.062	0.090	0.245	0.143	0.122
DGP II:							
Mean	0.505	-0.510	0.307	-0.301	1.534	0.538	-0.461
Median	0.504	-0.507	0.304	-0.316	1.520	0.519	-0.460
RMSE	0.068	0.093	0.201	0.235	0.598	0.157	0.169
DGP III:							
Mean	0.499	-0.499	0.308	-0.303	1.541	0.451	-0.459
Median	0.499	-0.503	0.302	-0.313	1.522	0.451	-0.458
RMSE	0.099	0.128	0.200	0.235	0.602	0.164	0.154

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

Table 2: Monte Carlo Simulations

	Infeasible Estimator		Conventional Estimator			Two-step Estimator	
Sample size=1000	θ	λ	θ	λ	σ	θ	λ
True	0.5	-0.5	0.5	-0.5	1	0.5	-0.5
DGP I:							
Mean	0.504	-0.502	0.506	-0.501	0.915	0.508	-0.507
Median	0.506	-0.504	0.507	-0.499	0.931	0.508	-0.496
RMSE	0.041	0.062	0.050	0.073	0.215	0.149	0.224
DGP II:							
Mean	0.504	-0.499	0.389	-0.379	1.308	0.523	-0.434
Median	0.498	-0.496	0.388	-0.378	1.323	0.518	-0.436
RMSE	0.045	0.069	0.119	0.142	0.362	0.135	0.208
DGP III:							
Mean	0.504	-0.500	0.315	-0.303	1.539	0.461	-0.382
Median	0.502	-0.504	0.314	-0.301	1.528	0.472	-0.395
RMSE	0.065	0.088	0.189	0.217	0.584	0.203	0.301

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

Table 3: Monte Carlo Simulation of the Average Partial Effects APE($\bar{x}$)

Sample size=500	Infeasible Estimator	Conventional Estimator	Two-step Estimator
DGP I:			
Mean	0.143	0.136	0.136
Std. dev.	0.020	0.017	0.042
RMSE	–	0.019	0.043
DGP II:			
Mean	0.131	0.066	0.134
Std. dev.	0.021	0.014	0.042
RMSE	–	0.067	0.042
DGP III:			
Mean	0.120	0.066	0.112
Std. dev.	0.019	0.014	0.042
RMSE	–	0.056	0.042

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

Table 4: Monte Carlo Simulation of the Average Partial Effects APE($\bar{x}$)

Sample size=1000	Infeasible Estimator	Conventional Estimator	Two-step Estimator
DGP I:			
Mean	0.143	0.142	0.131
Std. dev.	0.014	0.011	0.043
RMSE	–	0.011	0.045
DGP II:			
Mean	0.136	0.093	0.127
Std. dev.	0.016	0.017	0.010
RMSE	–	0.045	0.012
DGP III:			
Mean	0.118	0.068	0.132
Std. dev.	0.015	0.004	0.040
RMSE	–	0.050	0.038

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

Table 5: Estimates of the Persistent Effects of Union Membership

Explanatory Variables	Conventional Estimator	Two-step Estimator
Married$_t$	0.149	0.095
	(0.198)	(0.628)
Union$_{t-1}$	2.332	1.855
	(1.122)	(0.846)
Union$_0$	0.511	0.168
	(0.034)	(0.440)
Married$_{81}$	0.116	0.122
	(0.030)	(0.423)
Married$_{82}$	-0.156	-0.341
	(0.021)	(0.813)
Constant	-1.890	-1.283
	(0.068)	(0.983)
Variance, σ	1.011	–
	(0.097)	–

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 100 simulations.

Table 6: Estimates of the Average Partial Effects

Explanatory Variables	Conventional Estimator	Two-step Estimator
Married$_t$	0.046	0.036
	(0.002)	(0.089)
Union$_{t-1}$	0.544	0.386
	(0.184)	(0.088)

Note: Standard deviations of the parameters are computed by the standard deviation of the estimates across 100 simulations. The definition of average partial effect is in Eq. (29).