Indirect Inference for Nonlinear Panel Models with Fixed Effects*

Shuowen Chen†

November 1, 2021

Abstract

Fixed effect estimates of nonlinear panel data models suffers from the incidental parameter problem. This leads to two undesirable consequences in applied research: (1) point estimates are subject to large bias, and (2) confidence intervals have incorrect coverage. This paper proposes a simulation–based method for bias reduction. The method simulates data using the model with estimated individual effects, and finds values of parameters by equating fixed effect estimators obtained from observed and simulated data. The asymptotic framework provides consistency, bias correction and asymptotic normality results. An application and simulations to labor force participation illustrates the finite–sample performance of the method.

*Conversations with Chuqing Jin inspired and improved this project. I am indebted to Iván Fernández-Val, Jean–Jacques Forneron and Hiroaki Kaido for patience, guidance and encouragement. For helpful discussions and suggestions, I thank Karun Adusumilli, Eric Hardy, Dennis Kristensen, Yan Liu, Pierre Perron, Zhongjun Qu, Marc Rysman, Xiaoxia Shi, Martin Shi Yuen Shu, Shuang Wang, Guang Zhang and participants in BU econometrics seminars and the econometrics reading group. All errors are mine.

†Boston University. Email: swchen@bu.edu.

1 Introduction

Panel data refer to data for multiple entities (e.g., individuals, firms, etc.) observed at two or more time periods. Unobserved heterogeneity across entities often accounts for a large fraction of the variation in panel data. When this heterogeneity is correlated with the explanatory variables in the regression specifications, the resulting omitted variable bias renders point estimates inconsistency.

Adding individual fixed effects α_i's is the main approach to control for this time–invariant unobserved heterogeneity. Compared to other approaches like random effects and correlated random effects, the fixed effect method does not impose distributional assumptions on α_i's or restrict their relationships with other explanatory variables. Instead, each α_i is treated as a parameter to be estimated. However, because the number of α_i's increases with the sample size and each α_i is estimated using only entity i's time series observations, adding fixed effects introduces the incidental parameter problem in estimating the vector of parameters of interest θ_0. This problem has two consequences for applied research: (1) point estimates are subject to a large bias, and (2) confidence intervals have incorrect coverage.

This paper proposes a new method to debias fixed effect estimators in a class of nonlinear panel models. The method is named *indirect fixed effect estimator* and features two main steps: the first one is to simulate data by using estimated individual effects $\widehat{\alpha}_i$'s from the observed data. The second step is to find the vector of parameters that matches the fixed effect estimators using observed and simulated data.

The method has two advantages: first, it does not require an explicit characterization of the bias term, which can be hard to derive for complex models. Instead, the method finds the solution by automatically correcting the bias because the vector of parameter values that is the closest to θ_0 renders the same bias in fixed effects estimations. Second, standard errors can be derived using the delta method, so there is no need to use the bootstrap, which is computationally intensive.

The two properties are inherited from a precedent simulation–based estimation approach called indirect inference, which was first developed by Gouriéroux et al. (1993) and Smith (1993). In a nutshell, indirect inference uses an auxiliary model to summarize the statistical properties of the observed data and simulated data, and finds values of model parameters that match the parameters of the auxiliary model, estimated using the observed and simulated data, in terms of a minimum–distance criterion function. Because the same regression is run on observed and simulated data, matched estimators have the same bias structure and thus the bias gets cancelled.

The theory of indirect inference, however, is not directly applicable to nonlinear panel models, which are widely used in various fields of economics like industrial organization and labor. Because the individual effects cannot be differenced out, data simulation seem infeasible without imposing a functional form specification on them, and the bias term is a complicated function of θ_0 and α_i's.

To simulate data, this paper proposes using the estimated individual effects $\widehat{\alpha}_i$'s. These are informative proxies for the unknown individual effects α_i's because they become more accurate estimates when each individual's number of time series observations T grows large. Intuitively speaking, although data simulated using $\widehat{\alpha}_i$'s do not perfectly mimic the observed data, such a difference vanishes when T increases.

The indirect fixed effect estimator then debiases by matching the fixed effect estimators using observed and simulated data. This brings two advantages for the implementation and theoretical analysis of the new estimator. First, the minimum–distance criterion function for matching is just–identified because the dimensions of the fixed effect estimators are identical. Therefore, there is no need to consider an estimation of the optimal weighting matrix. It further implies that the matching can be made as exact as machine precision permits. The second advantage is with respect to the relationship between the vector of parameters of interest θ_0 and the unique maximizer of the limiting log likelihood function for fixed effects estimation. To back out point estimators of θ_0 from fixed effect estimators using simulated data, this relationship should be invertible. Because the unique maximizer is essentially θ_0, the relation turns out to be an identity function. Therefore, invertibility is satisfied trivially.

This paper presents consistency, bias correction and asymptotic normality results for the indirect fixed effect estimator. As usual in the indirect inference literature, consistency requires that the fixed effect estimators using observed and simulated data converge to the unique maximizer of the limiting log likelihood. Although the pointwise convergence of $\widehat{\theta}$ to θ_0 is a standard result in the large–T panel literature, three important differences arise in the analysis of fixed effect estimators using simulated data and pose technical challenges for establishing the uniform convergence.

First, the simulated data are generated using $\widehat{\alpha}_i$'s instead of the true ones. To justify this practice, the corresponding log likelihood function should uniformly well–approximate the one rendered by data simulated using the true individual effects. Otherwise, simulated fixed effect estimator is not even pointwise convergent. The proof of this statement, however, is complicated by the fact that the log likelihood function using simulated data can be nonsmooth for important types of nonlinear panel models, with binary choice models as leading examples. Intuitively speaking, when the dependent variable is discrete, a small change in the parameter values can lead to discrete changes in the simulated data. As a result, the sample log likelihood function using simulated data is discontinuous.

Simulations often generate discontinuous objective functions (e.g., McFadden, 1989; Pakes and Pollard, 1989), but this paper confronts a second difference: the fixed effect estimator using simulated data is nonsmooth with respect to the data generating process (DGP). Therefore, standard proof strategies in the panel literature (e.g., Hahn and Newey, 2004; Hahn and Kuersteiner, 2011) cannot be directly applied to characterize its limiting behavior.

Empirical process theory provides ample tools to handle nonsmoothness functions and moments in econometrics (Andrews, 1994), but the analysis of a nonsmooth fixed effect estimator is further complicated by the third difference: the presence of incidental parameters, whose number increases with the sample size n.

To prove uniform convergence with nonsmoothness, this paper follows Newey (1991) by establishing pointwise convergence and stochastic equicontinuity of fixed effect estimator in the simulation world. Intuitively speaking, pointwise convergence is equivalent to uniform convergence for any finite number of grid points, but without smoothness, the gap between any two grids can behave rather erratically. The stochastic equicontinuity condition is hence required to restrict such behaviors in probability.

The theoretical analysis of the indirect fixed effects estimator relies on some key structures of the panel data and the log likelihood function. Under the assumption that panel data are independent along the cross section dimension, this paper first justifies data simulation with $\widehat{\alpha}_i$'s by proving that the corresponding log likelihood function uniformly approximates the one from simulated data generated by true individual effects. As such, a uniform law of large number can be established and pointwise convergence in the simulation world follows from the standard consistency argument (Newey and McFadden, 1994). To verify the stochastic equicontinuity condition of fixed effect estimators using simulated data, this paper uses the concavity property of the profiled log likelihood to verfity one of the primitive conditions for stochastic equicontinuity in Andrews (1994). The proof strategy might be of independent interest for future research.

Regarding the asymptotic normality of the new estimator, because fixed effect estimators using simulated data are not smooth in θ and $\widehat{\alpha}_i$'s, the conventional strategy in indirect inference that relies on implicit function theorem (e.g., Gouriéroux et al., 1993) is not directly applicable. A regularity condition is thus imposed, which, combined with consistency, allows to explore bias correction and uniform central limit theorem through the lens of fixed effect estimators. More specifically, the fixed effect estimators in both worlds feature the same influence functions and bias structures, and it suffices to show the bias and CLT terms in the simulation world uniformly converge to the true asymptotic bias and variance respectively.

The main step of proving bias correction is to show that the bias term from simulated data is uniformly close to an infeasible counterpart from data simulated using the true individual effects with asymptotically negligible approximation errors. Because the latter term converges to the same probability limit as does the bias term from observed data, the theory of indirect inference ensures bias cancellation.

Applying this strategy to establish the asymptotic normality faces an extra technical challenge: the individual approximation errors of the CLT term for each i do not vanish at a rate fast enough so that the accumulated errors are asymptotically negligible. Using empirical process method, this paper bounds each individual CLT term by a uniform envelope and invoke the Donsker theorem.

Related Literature

The indirect fixed effect estimator presented in this paper combines four strands of literature, and this section provides a non–exhaustive of them. The incidental parameter problem is first discussed by Neyman and Scott (1948). When T is fixed, fixed effect estimators of nonlinear models are in general inconsistent because $\widehat{\theta}$ is not separable from $\widehat{\alpha}_i$'s and estimation errors do not vanish, even when number of cross–section n is very large (Chamberlain, 1984; Lancaster, 2000). Only some special models like static linear and conditional Logit specifications feature fixed–T consistent estimators (Andersen, 1970). A key insight of the large–T panel data literature is that the incidental parameter problem becomes an asymptotic bias problem when T grow with the sample size n. When n and T grow at the same rate, fixed effect estimators are consistent and asymptotically normal, but they still have a bias comparable to standard errors.

In the search for asymptotically unbiased estimators, there are two leading approaches. For certain types of models, the bias terms have been characterized analytically and corrected using a plugged–in approach (Hahn and Kuersteiner, 2002; Hahn and Newey, 2004; Fernández-Val, 2009; Hahn and Kuersteiner, 2011). However, such terms can be hard to derive for complicated models. Under further sampling and regularity conditions, bias terms can be automatically corrected using jackknife. For example, Hahn and Newey (2004) proposed delete–one panel jackknife for data that do not have dependencies among observations of the same unit. Dhaene and Jochmans (2015) relaxed the assumption to stationarity along the time series, and proposed a split–panel method. Under an unconditional homogeneity assumption, Fernández-Val and Weidner (2016) allowed for two–way fixed effects and propose a jackknife method that corrects biases from both dimensions. See Arellano and Hahn (2007) and Fernández-Val and Weidner (2018) for recent surveys. Standard errors are typically obtained by panel bootstrap, which can be computationally intensive[1]. Kim

[1]To obtain one debiased point estimate, fixed effects estimations are run three times: one for the whole sample, and twice for the two split samples. If the number of bootstraps is set to be 500, then the total number of fixed effects estimation becomes 1500. In addition, in practice it is often recommended to use multiple sample splits to improve the finite–sample performance.

and Sun (2016) proposed a parametric bootstrap bias correction method for the nonlinear panel models considered in this paper. Similar to the indirect fixed effect estimator, parametric bootstrap corrects the bias term implicitly using simulations, but it is computationally intensive to obtain the standard errors, and the proof strategies are totally different.

Second, this paper extends the existing theory and practice of indirect inference. Since the introduction of the method, its asymptotic theory has mainly been focused on times series data (Gouriéroux et al., 1993; Smith, 1993; Gallant and Tauchen, 1996). Some recent papers explore asymptotic properties in panel data with discrete dependent variables, but there are two key differences with this paper. First, their settings hold time series dimension fixed and study different types of models. For example, Bruins et al. (2018) did not consider models with fixed effects, Frazier et al. (2019) imposed normality on individual effects, and Taber and Sauer (2021) assumed a bivarite normal distribution on the types of individuals. Second, they deal with nonsmoothness by either smoothing discontinuous parts or constructing a differentiable criterion function that is asymptotically equivalent to the original one.

Gouriéroux et al. (2010) is the first paper that establishes theoretical properties of indirect inference for a class of large–T panel models. They applied indirect inference to dynamic panel linear models, whose fixed effect estimators are known to be biased (Nickell, 1981). The linear structure allows them to eliminate individual fixed effects α_i's by first–difference. As such, α_i's do not show up in the bias term, and data can be simulated without information on them. However, first difference does not apply to nonlinear panel models, and this paper fills the gap by extending the theory to handle the presence of α_i's in data simulation and the bias term.

Indirect inference is popular in various fields of economics, including empirical industrial organization (Collard-Wexler, 2013), labor economics (Altonji et al., 2013) and macroeconomics (Guvenen and Smith, 2014; Berger and Vavra, 2019). However, finding an informative auxiliary model is not a trivial task, and researchers often have to assume the invertibility of the limiting relationship between auxiliary parameters and parameters of interest. This paper provides an alternative choice, namely the log likelihood function from the nonlinear panel model, for researchers that employ panel data with fixed effects. The estimation procedures are simple to implement as fixed effect estimation schemes are available in free software like R and Julia.

Nonsmooth objective functions are common in econometrics, and empirical process methods are standard tools for asymptotic analysis (Andrews, 1994; Newey and McFadden, 1994; van der Vaart and Wellner, 1996). The seminal work on simulation–based methods by Pakes and Pollard (1989) is predicated on the independence assumption of cross section data and therefore is not suitable for panel data, which feature dependence for each individual time series. Dedecker and Louhichi (2002) provided an overview of maximal inequalities for empirical central limit theorems for dependent data. Kato et al. (2012) provided new stochastic inequalities for mixing sequences and also established stochastic equicontinuity in the presence of nuisance parameters, but their analysis focused on a different class of nonlinear models, namely panel quantile regression models.

Finally, this paper contributes to a burgeoning literature that considers simulations from semiparametric models. Simulation–based methods like simulated method of moments (McFadden, 1989; Pakes and Pollard, 1989; Lee and Ingram, 1991; Duffie and Singleton, 1993) and indirect inference are widely used to estimate models that do not render tractable moments or likelihood functions. See Gouriéroux and Monfort (1997) for an overview. These methods typically require models to be fully specified, but economic theory does not always provide guidance on functional forms, distributions of shocks or measurement error of observed data. Therefore, the resultant estimators can be subject to misspecification.

Dridi and Renault (2000) and Dridi et al. (2007) embedded the semiparametric structural model into a full model for data simulation, and proposed an encompassing principles where parameters of interest are consistently estimated even though nuisance parameters are inconsistently estimated due to misspecification of the full model. Schennach (2014) considered parameters estimation in moment conditions that contain unobservable variables, and proposes a simulation–based method that constructs equivalent moments involving only observable variables. Gospodinov et al. (2017) considered parameter estimation of autoregressive distributed lag models in which covariates are contaminated by serially correlated measurement errors. They proposed a method such that simulated covariates preserve the dependence structure observed in the data even though the dynamics of latent covariates or measurement errors are not specified. Forneron (2020) approximated the distribution of shocks by sieves and proposes a sieve–SMM estimator that jointly estimates structural parameters and the distribution of shocks.

Structure of the Paper

The rest of the paper proceeds as follows: Section 2 introduces the model and describes the fixed effect estimator and incidental parameter problem. Section 3 provides an overview of the indirect fixed effect estimator and its implementation. Section 4 presents the theoretical properties of the estimator. Section 5 applies the method to dynamic labor force participation to illustrate the finite–sample properties of the estimator. Section 6 uses Monte Carlo simulations to compare the new estimator with other bias correction methods. Section 7 concludes and discusses open questions. Appendices A, B and C consist of proofs and computation details.

2 Nonlinear Panel Model and fixed effect estimator

This section starts with a description of nonlinear panel models with fixed effects. Let the data observations be denoted by $\{z_{it} = (y_{it}, x_{it})\colon i = 1, \ldots, n; t = 1, \ldots, T\}$, where y_{it} is dependent variable and x_{it} is a $p \times 1$ vector of explanatory variable. The observations are independent across individual i and weakly dependent across time t. The data generation process (DGP) of outcome y_{it} takes the following form:

$$y_{it} \mid x_i^T, \boldsymbol{\alpha} \sim f(\cdot \mid x_{it}; \theta, \alpha_i), \tag{1}$$

where $x_i^T := (x_{i1}, \ldots, x_{iT})$, θ is a $p \times 1$ vector of model parameters, α_i is a scalar individual effect and $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)$. The explanatory variable x_{it} is strictly exogenous, but can be potentially extended to include lagged dependent varaible as well. The model is semiparametric in that neither the distribution of α_i's nor their relationships with explanatory variables x_{it} is specified. The conditional density f denotes the parametric part of the model. Depending on the assumptions on f, this type of models have been used to study many different questions of economic interest.

Example 1 (Discrete Choice Model)**.** Let y_{it} denote a binary variable and F_u a cumulative distribution function (CDF), e.g., the standard normal or logistic distribution. Suppose the binary variable is generated by the following single index process with additive individual effects:

$$y_{it} = \mathbf{1}\{x_{it}'\theta + \alpha_i \geq u_{it}\}, \quad u_{it} \mid x_i^t, \alpha \sim F_u,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Then the conditional distribution of y_{it} is expressed as

$$f(y_{it} \mid x_{it}, \alpha_i; \theta) = F_u(x_{it}'\theta + \alpha_i)^{y_{it}}(1 - F_u(x_{it}'\theta + \alpha_i))^{1-y_{it}}.$$

Helpman et al. (2008) modeled a country's export decision as Probit and estimate the gravity equation with country fixed effects. Fernández-Val (2009) used a Probit specification to estimate determinants of females' labor force participation decisions in the presence of time–invariant heterogeneity such as willingness to work. Collard-Wexler (2013) used a binary Logit specification with market–fixed effects to study whether a ready–mix concrete plant decides to be active in a market.

Example 2 (Poisson Regression Model)**.** The Poisson distribution is useful in modeling count data. Let y_{it} denote arrivals of new events within a certain time interval for firm i in year t. For $\lambda_{it} = \exp(x_{it}'\theta + \alpha_i)$, the conditional density becomes

$$f(y_{it} \mid x_{it}, \alpha_i; \theta) = \frac{\lambda_{it}^{y_{it}} \exp(-\lambda_{it})}{y_{it}!}\mathbf{1}\{y_{it} \in \{0, 1, \ldots\}\}.$$

Using the number of citation–weighted patents as a proxy for innovation, Aghion et al. (2005) employed this specification to study the relationship between innovation and competition with industry fixed effects.

Model (1) admits a log likelihood function. The true values of the parameters, denoted by θ_0 and $\boldsymbol{\alpha}_0 := (\alpha_{10}, \ldots, \alpha_{n0})$, are the unique solution to the population conditional maximum likelihood problem

$$(\theta_0, \alpha_{10}, \ldots, \alpha_{n0}) = \underset{(\theta, \alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{p+n}}{\arg\max} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E} \ln f(y_{it} \mid x_{it}; \theta, \alpha_i), \tag{2}$$

where the expectation is with respect to the distribution of the observed data, conditional on the unobserved effects and initial conditions. Section 4 discusses assumptions under which the log likelihood function is concave in all parameters and the solution uniquely exists. The indirect fixed effect estimator relies on the uniqueness condition for consistency.

2.1 The Fixed Effect Estimator

The fixed effect estimator of θ is obtained by doing maximum likelihood on sample analog of the population problem (2), treating each α_i as a parameter to be estimated.

$$(\widehat{\theta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_n) \in \underset{(\theta, \alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{p+n}}{\arg\max} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \theta, \alpha_i).$$

To facilitate theoretical analysis, this equation is rewritten such that the individual effects are profiled out. More specifically, given θ, the optimal $\widehat{\alpha}_i(\theta)$ for each i is defined as

$$\widehat{\alpha}_i(\theta) \in \underset{\alpha \in \mathbb{R}}{\arg\max} \frac{1}{T} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \theta, \alpha).$$

The estimators $\widehat{\theta}$ and $\widehat{\alpha}_i$ are then

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\max} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \theta, \widehat{\alpha}_i(\theta)), \quad \widehat{\alpha}_i = \widehat{\alpha}_i(\widehat{\theta}). \tag{3}$$

Section 4 discusses assumptions under which these estimators exist and are unique with probability approaching one as n and T become large.

2.2 The Incidental Parameter Problem

In panel models, the individual effects are incidental parameters, i.e., nuisance parameters whose dimension grows with the number of cross sectional observations n. As equation (3) shows, the fixed effect estimator $\widehat{\theta}$ cannot generally be separated from the estimator of individual effects $\widehat{\alpha}_i$'s. Because each $\widehat{\alpha}_i$ is only estimated using the T observations for i, its estimation error does not vanish if T fixed, even as n approaches infinity. These estimation errors in turn contaminate $\widehat{\theta}$. This is the incidental parameter problem for fixed effects estimation. Mathematically,

$$\widehat{\theta} \xrightarrow{p} \theta_T := \underset{\theta\in\mathbb{R}^p}{\arg\max} \underset{n\to\infty}{\text{plim}} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\ln f(y_{it} \mid x_{it}, \theta, \widehat{\alpha}_i(\theta))\right).$$

In contrast, the population problem (2) uses the true individual effects α_{i0}'s as follows:

$$\theta_0 := \underset{\theta\in\mathbb{R}^p}{\arg\max} \underset{n\to\infty}{\text{plim}} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\ln f(y_{it} \mid x_{it}, \theta, \alpha_{i0})\right).$$

Therefore, $\theta_T \neq \theta_0$ in general.

To illustrate the problem, suppose y_{it} has the normal distribution $\mathcal{N}(\alpha_{i0}, \theta_0)$, and the goal is to estimate the variance θ_0 in the presence of unknown individual–specific means α_{i0}'s. The fixed effect estimator is $\widehat{\theta} = \frac{1}{nT-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \widehat{\alpha}_i)^2$, where $\widehat{\alpha}_i = \frac{1}{T-1}\sum_{t=1}^{T} X_{it}$. When T is fixed and n approaches infinity, Neyman and Scott (1948) show that

$$\widehat{\theta} \xrightarrow{p} \theta_0 - \frac{\theta_0}{T}.$$

On the other hand, when T also grows to infinity, the bias term $-\frac{\theta_0}{T}$ approaches zero. The large–T panel literature generalizes this insight and shows that the incidental parameter problem becomes an asymptotic bias problem when n and T grow at the same rate.

3 An indirect fixed effect Estimator

The key feature of the indirect fixed effect estimator is to match $\widehat{\theta}$ with a fixed effect estimator from simulated data generated by $\widehat{\alpha}_i$'s and a given θ. To avoid confusion, it is necessary to introduce notations to distinguish parameters in the simulation world from those in Section 2. More specifically, this paper uses β and γ_i to denote the vector of parameters of interest and individual effects in the log likelihood function using simulated data.

To clarify the notations and introduce the implementation of indirect fixed effect estimator, this section first revisits the Neyman–Scott example. Using the panel Probit model as a concrete example, this section then illustrates the presence of nonsmoothness and discusses the general estimation procedures.

3.1 Neyman–Scott Example Revisited

That $y_{it} \mid \alpha_{i0} \sim \mathcal{N}(\alpha_{i0}, \theta_0)$ i.i.d over n and t implies the following data generating process (DGP) of the observed data:

$$y_{it}(\alpha_{i0}, \theta) = \alpha_{i0} + \sqrt{\theta} u_{it}, \quad u_{it} \sim \mathcal{N}(0,1).$$

This equation cannot be simulated without information on the distribution of α_{i0}'s. The indirect fixed effect estimator uses $\widehat{\alpha}_i$'s instead, and the simulated data have the following representation:

$$y_{it}^h(\widehat{\alpha}_i, \theta) = \widehat{\alpha}_i + \sqrt{\theta} u_{it}^h, \quad u_{it}^h \sim \mathcal{N}(0,1),$$

where the superscript h denotes a simulation path. The fixed effect estimator using $\{y_{it}^h(\widehat{\alpha}_i, \theta)\}$ is

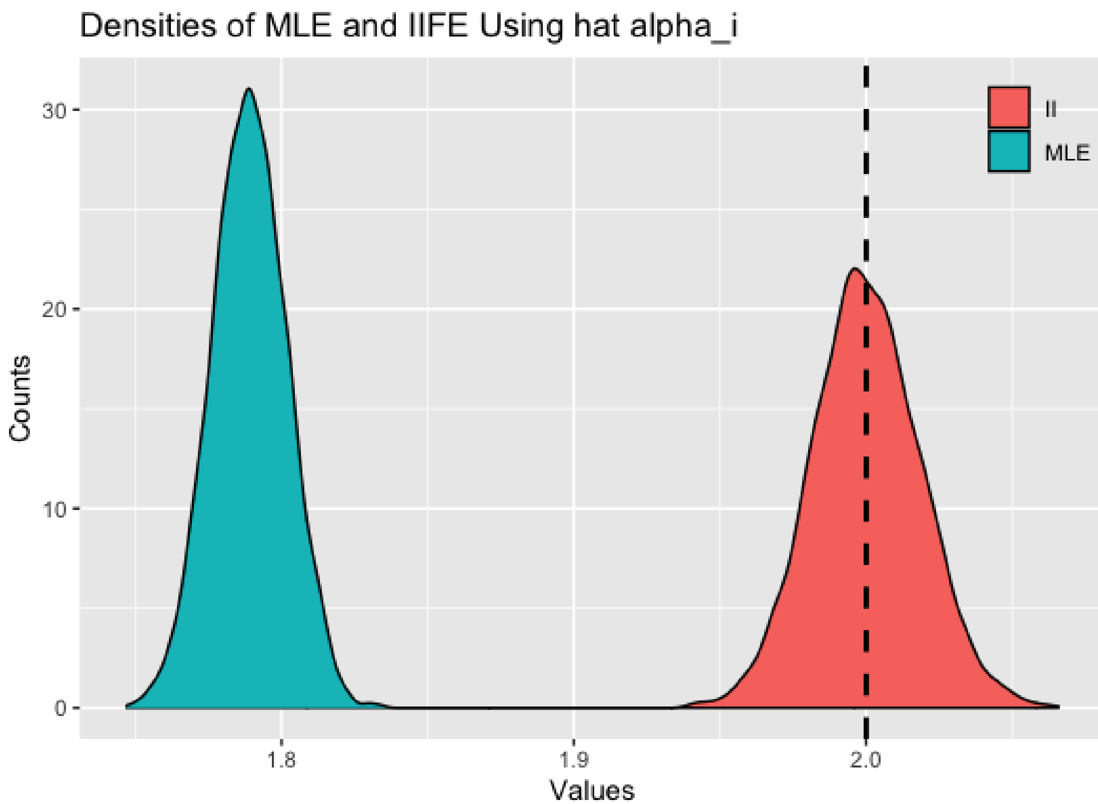
$$\widehat{\beta}^h(\theta) := \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it}^h(\widehat{\alpha}_i, \theta) - \widehat{\gamma}_i)^2 = \frac{\theta}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (u_{it}^h - \frac{1}{T-1} \sum_{t=1}^{T} u_{it}^h),$$

where $\widehat{\gamma}_i := \frac{1}{T-1} \sum_{t=1}^{T} y_{it}^h(\widehat{\alpha}_i, \theta)$. The interpretation of $\widehat{\beta}^h(\theta)$ is that the estimator changes if a different value of θ is used to simulate the data. Because $\widehat{\alpha}_i$'s are fixed throughout the simulation process, the dependence on $\widehat{\alpha}_i$ is suppressed. The indirect fixed effect estimator $\widetilde{\theta}$ is the solution to

$$\widehat{\theta} = \widehat{\beta}^h(\widetilde{\theta}).$$

Figure (1) illustrates the issues with $\widehat{\theta}$ and the performance of $\widetilde{\theta}$ in this example. The green region plots the density distribution of the fixed effect estimator $\widehat{\theta}$ and it conveys two messages: (1) fixed effect estimator is subject to a large bias and (2) the interval around $\widehat{\theta}$ does not have the correct coverage. The red region, on the other hand, plots the density distribution of $\widetilde{\theta}$ and two observations arise: (1) the new estimator corrects the bias significantly and (2) the interval is wider than that for $\widehat{\theta}$. This reflects a general trade–off between bias and variance: an unbiased estimator has a larger variance.

Figure 1: Comparison of FE and IFE



Note: Density plots of fixed effects and indirect fixed effect estimator for θ_0. The DGP is $y_{it} = \alpha_{i0} + \sqrt{\theta_0} u_{it}$, where $u_{it} \sim \mathcal{N}(0,1)$. The true value $\theta_0 = 2$ is depicted by the dashed line and $\alpha_{i0} = i$ for $i = 1, \ldots, n$. The sample size is $n = 2500, T = 5$ and number of simulation H is set to be 1. The simulations are conducted 5000 times.

Remark 1 (Caveat of this example)**.** Due to the simple structure of the data, $\widehat{\theta}$ and $\widehat{\beta}^h(\theta)$ have closed–form expressions, and the bias term does not contain α_i's. However, it is not the case in model (1). In addition, the discrete nature of binary dependent variables leads to a nonsmooth log likelihood function in the simulation world.

3.2 Challenges of Nonsmoothness

Consider the binary choice panel Probit model as a concrete example. Given θ, $\widehat{\alpha}_i$'s and x_{it}, the simulated dependent variable is

$$y_{it}^h(\theta, \widehat{\alpha}_i) = \mathbf{1}(x_{it}'\theta + \widehat{\alpha}_i > u_{it}^h), \quad u_{it}^h \sim \mathcal{N}(0,1),$$

where u_{it}^h are simulation draws from the standard normal distribution. The corresponding log likelihood function is

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}^h(\theta,\widehat{\alpha}_i)\log\left(\Phi(x_{it}'\beta+\gamma_i)\right)+(1-y_{it}^h(\theta,\widehat{\alpha}_i))\log\left(1-\Phi(x_{it}'\beta+\gamma_i)\right). \tag{4}$$

This equation illustrates the three different aspects of simulated fixed effect estimator $\widehat{\beta}^h(\theta)$. Because $\{y_{it}^h(\theta,\widehat{\alpha}_i)\}$ are discontinuous in θ and $\widehat{\alpha}_i$'s, equation (4) is discontinuous, which carries over to its maximizer $\widehat{\beta}^h(\theta)$. In addition, estimating $\widehat{\beta}^h(\theta)$ involves incidental parameters γ_i's. The probability limit of equation (4) is

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big[y_{it}(\theta,\alpha_{i0})\log\left(\Phi(x_{it}'\beta+\gamma_i)\right)+(1-y_{it}(\theta,\alpha_{i0}))\log\left(1-\Phi(x_{it}'\beta+\gamma_i)\right)\Big], \tag{5}$$

where the expectation is over u_{it}^h and x_{it}, and $\widehat{\alpha}_i$'s are replaced by α_{i0}'s.

Remark 2 (A comparison with panel quantile regression (QR) models)**.** One important type of nonlinear panel models that is not included in model (1) but also features nonsmoothness is panel QR models[2]. Kato et al. (2012) consider the following setup with individual effects:

$$Q_\tau(y_{it}\mid x_{it},\gamma_{i0}(\tau))=\gamma_{i0}(\tau)+x_{it}'\beta_0(\tau),$$

where $\tau\in(0,1)$ is a quantile index, and $Q_\tau(y_{it}\mid x_{it},\gamma_{i0}(\tau))$ is the conditional τ–quantile of y_{it} given $(x_{it},\gamma_{i0}(\tau))$. The fixed effects quantile regression (FE–QR) estimator for this model is

$$(\widehat{\gamma}_{\text{FE-QR}},\widehat{\beta}_{\text{FE-QR}})\in\arg\min\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(y_{it}-\gamma_i-x_{it}'\beta),$$

where $\gamma:=(\gamma_1,\ldots,\gamma_n)'$ and $\rho_\tau(u):=\{\tau-\mathbf{1}\{u\le 0\}\}u$ is the check function. Because the check function is non–differentiable, the FE–QR estimator is also non–smooth. However, what makes it different from this paper is that the criterion function (4) is still smooth in β and γ_i's, and this is the key component for the asymptotic properties of the indirect fixed effect estimator[3].

[2] See Galvao and Kato (2018) for a recent survey.

[3] More specifically, the CLT and bias terms for smooth likelihood are combinations of high–order stochastic expansion terms of score functions, which is hard to derive for panel QR models.

3.3 General Estimation Procedure

From the known distribution F_u, the simulated unobservables $\{u_{it}^h\}$ are independently drawn for $h = 1, \ldots, H$, where H denotes the number of simulated panel data sets. For a given value of θ, let $y_{it}^h(\theta, \widehat{\alpha}_i)$ denote the simulated dependent variable for simulation path h, then the sample log likelihood function using the h–th simulated data is

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it}^h(\theta, \widehat{\alpha}_i) \mid x_{it}; \beta, \gamma_i), \tag{6}$$

where β and γ_i respectively denote the finite–dimensional parameter and incidental parameter in the simulation world. The fixed effect estimator to this problem is

$$\widehat{\beta}^h(\theta) = \arg\max_{\beta \in \mathbb{R}^p} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it}^h(\theta, \widehat{\alpha}_i) \mid x_{it}; \beta, \widehat{\gamma}_i(\beta, \theta)),$$

where

$$\widehat{\gamma}_i(\beta, \theta) = \arg\max_{\gamma \in \mathbb{R}} \frac{1}{T} \sum_{t=1}^{T} \ln f(y_{it}^h(\theta, \widehat{\alpha}_i) \mid x_{it}; \beta, \gamma).$$

Repeating this estimation for all simulated panel, the following average can be computed:

$$\widehat{\beta}^H(\theta) := \frac{1}{H} \sum_{h=1}^{H} \widehat{\beta}^h(\theta),$$

The indirect fixed effect estimator $\widetilde{\theta}^H$ is the solution to

$$\widehat{\theta} = \widehat{\beta}^H(\widetilde{\theta}^H), \tag{7}$$

where the superscript H stresses that the finite–sample performance depends on the number of simulations conducted. The box below summarizes the steps required to compute the estimator.

Algorithm: Computing the indirect fixed effect estimator

(i) Obtain $(\widehat{\theta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$ using the observed data.

(ii) Set a random seed and H. For each i and t, draw unobservables $\{u_{it}^h\}_{h=1}^H$ from F_u.

(iii) Given θ and $\widehat{\alpha}_i$'s, use model (1) and $\{u_{it}^h\}$ to simulate dependent variable $\{y_{it}^h(\theta, \widehat{\alpha}_i)\}$; construct data $\{y_{it}^h(\theta, \widehat{\alpha}_i), x_{it}\}$, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

(iv) Obtain $\widehat{\beta}^h(\theta)$ using the simulated data in Step (iii).

(v) Repeat steps (ii) and (iii) for all $h = 1, \ldots, H$ and solve for equation (7).

Remark 3 (Common random number)**.** Step (ii) follows the standard practice for simulations (Glasserman and Yao, 1992) by drawing unobserved shocks only once at the beginning of the algorithm. It implies that $\widehat{\beta}^h(\theta)$ and $\widehat{\beta}^{h'}(\theta)$ are independent for $h \neq h'$ conditional on x_{it}.

Remark 4 (The role of H)**.** The number of H affects the finite–sample performance of the estimator, and increasing H reduces the asymptotic variance. Just like SMM and indirect inference, there is a trade off between precision of the estimator and intensity of computation. The estimation method, however, is different from the simulated maximum likelihood (Manski and Lerman, 1981), which is inconsistent for fixed H due to a nonlinear transformation of simulated choice probabilities.

Remark 5 (Choices of optimization algorithms)**.** When computing fixed effect estimators in the simulation world, the discontinuity nature makes gradient–based optimization methods unsuitable, so simplex–based methods like Nelder–Mead are used instead.

4 Asymptotic Properties

This section starts with a discussion of the main assumptions that lead to theoretical properties of $\widehat{\theta}$ and $\widehat{\alpha}_i$'s. These assumptions are standard in large–T panel data models (Hahn and Kuersteiner, 2011), and they also impose certain structures that help establish the asymptotic properties of the indirect fixed effect estimator. Additional assumptions are imposed to ensure simulations do not affect the panel data structure.

Assumption 1 (Large T asymptotics)**.** $n, T \to \infty$ such that $nT^{-1} \to \kappa \in (0, \infty)$.

Assumption 1 requires that time series dimension grows at the same rate as the cross section dimension. The assumption defines the large–T asymptotics framework and is a necessary condition for consistency of fixed effect estimators using observed data. This assumptions allows to transform the incidental parameter problem from a consistency to a bias problem, the latter of which can be quantified.

Assumption 2 (Sampling of observed data)**.** (i) $\{z_{it}\}_{t=1}^{\infty}$ are independent across i; (ii) For each i, $\{z_{it}\}_{t=1}^{\infty}$ is a stationary α–mixing sequence with mixing coefficient $\alpha_i(m)$, and $\sup_i |\alpha_i(m)| \leq Ka^m$ for some a such that $0 < a < 1$ and some $K > 0$.

Assumption 2(i) imposes independence along the cross–section dimension. This assumption is crucial for the theoretical analysis of fixed effect estimation because it allows to decompose the aggregate log likelihood to individual contributions, each of which then only contains a fixed number of parameters. Assumption 2(ii) imposes temporal dependence on each individual time series. The quantity $\alpha_i(m)$ measures for each i how much dependence exists between data separated by at least m time periods, and a uniform bound is imposed so as to bound covariances and moments when using law of large numbers (LLN) and central limit theorem (CLT). Interested readers can refer to Section 3.4 in White (2000) for definitions and properties. Note that Assumption 2 rules out nonstationarity such as time effects and linear trends.

Assumption 3 (Identification)**.** Denote $G_{(i)}(\theta, \alpha_i) \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E} \ln f(y_{it} \mid x_{it}, \theta, \alpha_i)$. For each $\eta > 0$,

$$\inf_i \left[G_{(i)}(\theta_0, \alpha_{i0}) - \sup_{(\theta,\alpha):\|(\theta,\alpha)-(\theta_0,\alpha_{i0})\| \geq \eta} G_{(i)}(\theta, \alpha) \right] > 0.$$

Assumption 3 is a sufficient condition that ensures the log likelihood function admits a unique maximizer based on time series variation. This assumption allows to prove the consistency of fixed effect estimators under large−T asymptotics. The indirect fixed effect estimator also requires this assumption for consistency.

Assumption 4 (Envelope condition)**.** (*i*) The parameter $\varphi_i := (\theta, \alpha_i) \in \operatorname{int} \Theta \times \Gamma_\alpha$, where Θ and Γ_α are compact, convex subset of $\mathbb{R}^p$ and $\mathbb{R}$ respectively. (*ii*) There exists an envelope function $M(z_{it})$ such that

$$\|D^\nu G_{(i)}(\varphi_1; \theta) - D^\nu G_{(i)}(\varphi_2; \theta)\| \leq M(z_{it})\|\varphi_1 - \varphi_2\|,$$

where $D^\nu G_{(i)}(\varphi_1; \theta) := \partial^{|\nu|} G_{(i)}(\varphi)/(\partial \varphi_1^{\nu_1} \ldots \partial \varphi_p^{\nu_p})$ and $|\nu| \leq 5$.

Assumption 4(*i*) imposes compactness of parameter space, which is standard for establishing asymptotic properties of extremum estimators (Newey and McFadden, 1994). Compactness is indispensable for proving uniform convergence with nonsmooth criterion functions (Newey, 1991). Assumption 4(*ii*) imposes a Lipschitz condition on the log likelihood function and a moment condition on the envelope function. This allows to establish uniform law of large number (ULLN) of sample log likelihood function and hence the pointwise consistency of $\widehat{\theta}$.

Under these assumptions and some regularity conditions on the Hessian matrix, Hahn and Kuersteiner (2011) established the following two results:

$$\max_{1 \leq i \leq n} |\widehat{\alpha}_i - \alpha_{i0}| = o_p(1), \tag{8}$$

$$\widehat{\theta} = \theta_0 + \frac{A(\theta_0, \alpha_0)}{\sqrt{nT}} + \frac{B(\theta_0, \alpha_0)}{T} + o_p\left(\frac{1}{T}\right). \tag{9}$$

Equation (8) states that the maximal deviation of $\widehat{\alpha}_i$ from α_{i0} converges to zero. This uniform consistency result is crucial for the theory of indirect fixed effect estimator because it justifies the usage of $\widehat{\alpha}_i$'s for data simulations. Equation (9) characterizes the asymptotic relationship between $\widehat{\theta}$ and θ_0. The term $A(\theta_0, \alpha_0)$ is the influence function that satisfies the central limit theorem (CLT) with zero mean. The term $B(\theta_0, \alpha_0)$ converges to its expected value. Therefore, $\widehat{\theta}$ is consistent, asymptotically normal, but biased. Hahn and Kuersteiner (2011) derived the analytical forms of both terms, which are complicated functions of θ_0 and α_0.

Because the new estimator involves simulations, the following regularity condition is required so that the simulated data still maintain the mixing properties. Another regularity condition is that the parameter space in the simulation world is compact. Because (β, γ) is just a change of notation from (θ, α), this assumption is natural.

Assumption 5 (Simulation)**.** (*i*) Assumption 2 holds for the simulated process for all $\theta \in \Theta$. (*ii*) The parameter spaces for β and γ_i are Θ_β and Γ_γ are compact.

In sum, Assumption 5 allows for an asymptotic representation of simulated fixed effect estimator $\widehat{\beta}^h(\theta)$ that resembles the one for $\widehat{\theta}$, i.e., equation (9).

Remark 6 (Inferring $\widetilde{\theta}^H$ from fixed effect estimators)**.** Backing out $\widetilde{\theta}^H$ from $\widehat{\beta}^H(\widetilde{\theta}^H)$ requires an invertible relationship between θ and $\beta(\theta)$, which is the maximizer of the limiting function for equation (6)

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\ln f(y_{it}(\theta, \alpha_{i0}) \mid x_{it}; \beta, \gamma_i).$$

The expectation is taken over simulation draws and sampling of observed data, and $\widehat{\alpha}_i$ is replaced with α_{i0}. This function is essentially the population function for the fixed effects estimation problem, except that β and γ_i are used to denote parameters for estimation in the simulation world. Assumption 3 thus ensures the uniqueness of $\beta(\theta)$, which is θ. As such, invertibility is satisfied trivially[4]. Although $\beta(\theta)$ is an identity function, for the rest of the paper this notation is kept to avoid the confusion between maximum of the limit and a parameter for data generation.

4.1 Consistency

In order for the indirect inference–type estimator to be consistent, three conditions should be satisfied (Gouriéroux et al., 1993): an invertible relationship between θ and $\beta(\theta)$, pointwise convergence of $\widehat{\theta}$ to $\beta(\theta_0)$, and uniform convergence of $\widehat{\beta}^h(\theta)$ to $\beta(\theta)$ over the compact parameter

[4]For readers who are familiar with indirect inference, the relationship essentially means that the binding function is an identity. This is because the auxiliary model is identical to the structural model, and thus the parameters in the two models coincide. Many papers that employ indirect inference often have to assume invertibility of the binding function (Collard-Wexler, 2013; Gospodinov et al., 2017), but this assumption is guaranteed in this paper.

space Θ. The first condition satisfies because $\beta(\theta)$ is an identity, and equation (9) gives the second condition. The following proposition formally states the uniform convergence condition.

Proposition 1 (Uniform convergence of fixed effect estimator using simulated data)**.** *Under Assumptions 1–5,*

$$\sup_{\theta\in\Theta}\|\widehat{\beta}^h(\theta) - \beta(\theta)\| \xrightarrow{p} 0.$$

The current proof specializes in panel Probit models, but it is generalizable to other models that features concavity and smoothness in (β, γ_i). Details are available in Appendix B, and here the main ideas are discussed.

Proving the uniform convergence condition with nonsmoothness requires two steps: pointwise convergence of $\widehat{\beta}^h(\theta)$ to $\beta(\theta)$, and a stochastic equicontinuity condition as follows:

$$\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\|\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)\|\Big) \le C\delta, \tag{10}$$

where C is a constant and δ is a positive scalar.

Following the standard argument in Newey and McFadden (1994), pointwise convergence requires a ULLN result of log likelihood function using simulated data (4) to the limiting log likelihood (5). The log likelihood (4) has two sources of randomness: the first source comes from sampling variation of observed data, and the other is from simulations of unobservables. The non–standard part, however, is that data are simulated using $\widehat{\alpha}_i$'s. Therefore, it is necessary to first show that (4) uniformly well approximates the log likelihood using data generated by α_{i0}'s:

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int_U \Big[y_{it}^h(\theta,\alpha_{i0})\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big) + (1-y_{it}^h(\theta,\alpha_{i0}))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)\Big]dF_u, \tag{11}$$

where the integration is with respect to the distribution of simulation draws to eliminate randomness from simulations. The details are available in Lemma 1, and intuition is provided here. Because panel data are independent along cross section, it suffices to show that each individual's log likelihood

$$\frac{1}{T}\sum_{t=1}^{T} y_{it}^h(\theta,\widehat{\alpha}_i)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big) + (1-y_{it}^h(\theta,\widehat{\alpha}_i))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)$$

satisfies this property. Given θ, this individual log likelihood is an additive and multiplicative combinations of indicator functions of scalar $\widehat{\alpha}_i$ and smooth functions of (β, γ_i), which belongs to classes of functions that satisfy stochastic equicontinuity (van der Vaart and Wellner, 1996).

Therefore, its empirical process

$$\nu_T(\tau) = \frac{1}{T}\sum_{t=1}^{T}\Big[y_{it}^h(\theta,\tau)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big) + (1-y_{it}^h(\theta,\tau))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)$$
$$- \int_U y_{it}^h(\theta,\tau)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big) + (1-y_{it}^h(\theta,\tau))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)dF_u\Big]$$

is stochastic equicontinuous. Combined with uniform consistency result of $\widehat{\alpha}_i$'s and LLN of $\nu_T(\alpha_{i0})$, an application of the triangular inequality leads to the uniform approximation result. Now that (11) only has randomness from observed data, its uniform convergence to the limiting log likelihood (5) follows the argument as in Hahn and Kuersteiner (2011). As such, the pointwise convergence of $\widehat{\beta}^h(\theta)$ follows through[5]. To verify the stochastic equicontinuity condition (10), note that the profiled log likelihood

$$\widehat{Q}(\beta;\theta) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}^h(\theta,\widehat{\alpha}_i)\log\Big(\Phi(x_{it}'\beta+\widehat{\gamma}_i)\Big) + (1-y_{it}^h(\theta,\widehat{\alpha}_i))\log\Big(1-\Phi(x_{it}'\beta+\widehat{\gamma}_i)\Big)$$

is concave in β. By definition, $\widehat{\beta}^h(\theta_1)$ satisfies the first–order condition $\partial\widehat{Q}(\widehat{\beta}^h(\theta_1);\theta_1)/\partial\beta = 0$. A first–order Taylor expansion with respect to $\widehat{\beta}^h(\theta_1)$ around $\widehat{\beta}^h(\theta_2)$ shows that $\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)$ is bounded by $\left|\frac{\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} - \frac{\partial\widehat{Q}(\widehat{\beta}(\theta_2);\theta_1)}{\partial\beta}\right|$, which, by the Cauchy–Schwarz inequality, is bounded by the product of two terms: a smooth function of (β,γ_i) and

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}^h(\theta_1,\widehat{\alpha}_i) - y_{it}^h(\theta_2,\widehat{\alpha}_i)). \tag{12}$$

Therefore, it suffices to bound the two terms in expectation. The technical challenge mainly comes from proving this for equation (12). Although indicator functions are well–known to have controlled complexities (Andrews, 1994), and a similar result on the difference of indicator functions with univariate variable is given in Chen et al. (2003), here θ's are multi–dimensional. It turns out that the expectation of (12) satisfies the L^2–smoothness regularity condition[6] in Andrews (1994). Therefore, the stochastic equicontinuity condition for $\widehat{\beta}^h(\theta)$ is verified.

Armed with Proposition 1, the consistency of $\widetilde{\theta}^H$ follows the arguments as in Gouriéroux et al. (1993). The proof is straightforward because there is no need to consider the weighting matrix.

Theorem 1 (Consistency of indirect fixed effect estimator)**.** *Under Assumptions 1–5,*

$$\widetilde{\theta}^H \xrightarrow{p} \theta_0.$$

[5]Details are available in Lemma 2.

[6]See proof of Proposition 1 for details.

4.2 Bias Correction and Asymptotic Normality

Recall that the indirect fixed effect estimator using H simulations $\widetilde{\theta}^H$ is the solution to $\widehat{\theta} = \widehat{\beta}^H(\widetilde{\theta}^H)$. Non–differentiability of $\theta \mapsto \widehat{\beta}^H(\theta)$ means that the techniques in the indirect inference literature (e.g., Gouriéroux et al., 1993) is not applicable. The following smoothness assumption is imposed.

Assumption 6. For all positive random sequences $\delta_{nT} \to 0$,

$$\sup_{\|\theta_1-\theta_2\|\leq\delta_{nT}} \sqrt{nT}\|\widehat{\beta}^H(\theta_1) - \widehat{\beta}^H(\theta_2) - \mathbb{E}(\widehat{\beta}^H(\theta_1) - \widehat{\beta}^H(\theta_2))\| \xrightarrow{p} 0.$$

Assumption 6 requires that the difference between $\widehat{\beta}^H(\theta_1)$ and $\widehat{\beta}^H(\theta_2)$ can be approximated by its expectation at a $\sqrt{nT}$ rate. Combined with consistency of $\widetilde{\theta}^H$ and the mean value theorem, it allows to analyze the asymptotic normality of $\widetilde{\theta}^H$ through the lens of fixed effect estimators as follows:

$$\sqrt{nT}(\widetilde{\theta}^H - \theta_0) = \sqrt{nT}(\widehat{\theta} - \widetilde{\beta}^H(\theta_0)) + o_p(1). \tag{13}$$

Recall that equation (9) characterizes the representation of $\widehat{\theta}-\theta_0$. Because the same regression is run on simulated data h and the likelihood is smooth in (β, γ_i), the same structure of representation arises, namely that

$$\widehat{\beta}^h(\theta_0) - \theta_0 = \frac{A^h(\theta_0, \widehat{\alpha})}{\sqrt{nT}} + \frac{B^h(\theta_0, \widehat{\alpha})}{T} + o_p\Big(\frac{1}{T}\Big). \tag{14}$$

The terms $A^h(\theta_0, \widehat{\alpha})$ and $B^h(\theta_0, \widehat{\alpha})$ reflect that the data are generated using θ_0, $\widehat{\alpha}$ and simulated unobservables $\{u_{it}^h\}$. A combination of (9), (13) and (14) therefore leads to

$$\begin{aligned}\sqrt{nT}(\widetilde{\theta}^H - \theta_0) = &\Big(A(\theta_0, \alpha_0) - \frac{1}{H}\sum_{h=1}^{H} A^h(\theta_0, \widehat{\alpha})\Big) \\ &+ \sqrt{\frac{n}{T}}\Big(B(\theta_0, \alpha_0) - \frac{1}{H}\sum_{h=1}^{H} B^h(\theta_0, \widehat{\alpha})\Big) + o_p(1).\end{aligned} \tag{15}$$

This equation reflects two observations. First, $\widehat{\theta}$ is unbiased if $B(\theta_0, \alpha_0)$ and $B^h(\theta_0, \widehat{\alpha})$ both converge to $\mathbb{E}(B(\theta_0, \alpha_0))$. Second, $\widehat{\theta}$ is asymptotically normal if $A(\theta_0, \alpha_0)$ and $A^h(\theta_0, \widehat{\alpha})$ converge to the same limiting distribution, but the variance is inflated by a factor of $\frac{1}{H}$. The rest of the section provides the main ideas of the proof.

Bias Correction

The intuition can be gained by setting $H = 1$ and considering an infeasible fixed effect estimator $\widehat{\beta}^h(\theta_0, \alpha_0)$, which is obtained from data simulated by (θ_0, α_0). Then the representation of $\widehat{\beta}^h(\theta_0, \alpha_0) - \theta_0$ takes the form

$$\sqrt{nT}(\widehat{\beta}^h(\theta_0, \alpha_0) - \theta_0) = A^h(\theta_0, \alpha_0) + \sqrt{\frac{n}{T}} B^h(\theta_0, \alpha_0) + o_p(1).$$

The theory of indirect inference implies that $B(\theta_0, \alpha_0)$ and $B^h(\theta_0, \alpha_0)$ converge to the same probability limit $\mathbb{E}(B(\theta_0, \alpha_0))$. Because the actual simulated data are generated by $\widehat{\alpha}$, it suffices to show that $B^h(\theta_0, \widehat{\alpha})$ uniformly well approximates $B^h(\theta_0, \alpha_0)$ such that the approximation error is asymptotically negligible. More specifically, the bias term using simulated data takes the following form,

$$B^h(\theta_0, \widehat{\alpha}) = -\Big[\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big]^{-1} \frac{1}{n}\sum_{i=1}^{n} B_i^h(\theta_0, \widehat{\alpha}_i),$$

where $\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)$ is individual i's information matrix, and it is a smooth function of all its arguments. Therefore, $\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i(\theta_0, \widehat{\alpha}_i) \xrightarrow{p} \mathcal{I}_i(\theta_0, \alpha_{i0})$ for each i. Although each $B_i^h(\theta_0, \widehat{\alpha}_i)$ is nonsmooth in $\widehat{\alpha}_i$, uniform consistency of $\widehat{\alpha}_i$'s allows to establish another stochastic equicontinuity condition such that $B^h(\theta_0; \widehat{\alpha})$ replaces $B^h(\theta_0, \alpha_0)$ with negligible errors.

Proposition 2 (Bias correction of $\widetilde{\theta}^H$)**.** *Under Assumptions 1–6,*

$$|B^h(\theta_0, \widehat{\alpha}) - B^h(\theta_0, \alpha_0)| \xrightarrow{p} 0.$$

As such, the indirect fixed effect estimator corrects the bias.

Asymptotic Normality

The same intuition could be applied to establish the asymptotic normality: if $A^h(\theta_0, \widehat{\alpha}_i)$ can uniformly well approximate $A^h(\theta_0, \alpha_0)$ with negligible errors, then the asymptotic normality result in indirect inference literature follows through (Gouriéroux et al., 1993, Proposition 5). The technical challenge, however, is that the approximation error is not decreasing at a sufficiently fast rate. More specifically, the independence assumption along the cross section implies that

$$A^h(\theta_0, \widehat{\alpha}) = \Big[\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i^h(\theta_0, \widehat{\alpha}_i)\Big]^{-1} \frac{1}{\sqrt{nT}} \sum_{i=1}^{n}\sum_{t=1}^{T} A_{it}^h(\theta_0, \widehat{\alpha}_i),$$

where $A_{it}^h(\theta_0, \widehat{\alpha}_i)$ is a combination of high–order derivatives of the log likelihood. Intuitively speaking, $A_h^h(\theta_0, \widehat{\alpha}_i)$ is a plug–in estimate of $A^h(\theta_0, \alpha_{i0})$, but because the number of individual effects increases with the sample size n and each $\widehat{\alpha}_i$ does not converge to α_{i0} fast enough, a multiplication of factor $1/\sqrt{nT}$ magnifies the approximation errors.

Establishing the functional asymptotic normality involves the following main steps: for each i, invoke a coupling lemma (e.g., Dedecker and Louhichi, 2002, Lemma 4.1) to construct independent sequences that approximate its simulated time series with a controlled remainder, and find an envelope for its simulated CLT term. As such, the expectation of each i's simulated CLT term $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} A_{it}^h(\theta_0, \widehat{\alpha}_i)$ can be bounded by a finite uniform entropy integral (van der Vaart, 2000, Corollary 19.35). These individual bounds are then aggregated to form a measure of approximation error that is asymptotically negligible. This in turn leads to the following theorem on the asymptotic distribution of the indirect fixed effects estimator.

Theorem 2. *Under Assumptions 1–6,*

$$\sqrt{nT}\Big(\widetilde{\theta}^H - \theta_0\Big) \xrightarrow{d} \mathcal{N}\Big(0, \Big(1 + \frac{1}{H}\Big)\Omega\Big),$$

where $\Omega := \mathbb{E}(A(\theta_0, \alpha_0)A(\theta_0, \alpha_0)')$.

As previously discussed in Remark 4, the number of simulations H shows up as a factor that inflates the asymptotic variance. There are two interpretations. The first is in line with other simulation–based methods: using simulations introduces an additional source of uncertainty and it is manifested through an increase in variance. The other interpretation is related to the trade–off between bias and variance. Because the indirect inference estimator debiases fixed effects estimator, the variance is larger, and it is quantified by the number of panel data simulated.

5 Application to Labor Force Participation

Research on the relationship between fertility on female labor force participation is complicated by the presence of unobserved factors that affect both decisions. Following Hyslop (1999), this paper addresses the omitted variable issue by including individual fixed effects into the binary response panel probit model for the female labor force participation.

The data come from the Panel Study of Income Dynamics (PSID) and constitute a nine–year longitudinal sample spanning from 1979 to 1988. The sample includes 664 women aged 18—60 in 1985 who were continuously married with husbands in the labor force in each of the sample periods and changed their labor force participation statuses. Consider the following static specification:

$$y_{it} = \mathbf{1}\{x_{it}'\theta + \alpha_i > u_{it}\}, \quad u_{it} \sim \mathcal{N}(0,1),$$

where y_{it} denotes the labor force participation indicator for woman i at time t, and x_{it} denotes a vector of time–varying covariates. These covariates include numbers of children of at most 2 years of age, between 3 and 5 years of age, between 6 and 17 years of age; log of the husband's income[7], age and age squared. The individual effects α_i's are included to control for time–invariant unobserved heterogeneity such as unwillingness to work or ability.

TABLE 1: PARAMETER ESTIMATES FOR STATIC LFP

	kids0_2	kids3_5	kids6_17	loghusinc	age	age2
FE	-0.71	-0.41	-0.13	-0.24	2.32	-0.29
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
IFE−1	-0.65	-0.36	-0.08	-0.17	2.24	-0.29
	(0.08)	(0.07)	(0.06)	(0.08)	(0.53)	(0.07)
IFE−10	-0.60	-0.32	-0.10	-0.30	2.08	-0.27
	(0.06)	(0.05)	(0.04)	(0.06)	(0.39)	(0.05)
IFE−20	-0.60	-0.32	-0.10	-0.30	2.08	-0.27
	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)
ABC	-0.63	-0.37	-0.11	-0.22	2.39	-0.25
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
JBC	-0.62	-0.36	-0.10	-0.21	1.73	-0.22
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
SBC	-0.92	-0.58	-0.26	-0.30	2.28	-0.26
	(0.09)	(0.09)	(0.08)	(0.07)	(0.89)	(0.12)

Notes: Standard errors are stored in the parenthesis and are computed based on the Hessian matrix of profiled log likelihood. The SBC estimates and standard errors computation follow page 1025 in Dhaene and Jochmans (2015).

Table (1) reports estimates of index coefficients using different methods. The standard errors, which are computed from the Hessian of the profiled log likelihood, are tabulated in the parenthesis. IFE−1, IFE−10 and IFE−20 denote indirect fixed effect estimators with H being 1, 10 and 20 respectively. For comparisons, the table includes results using split–panel jackknife method (SBC) (Dhaene and Jochmans, 2015) and analytical bias correction (ABC) (Fernández-Val, 2009). Because the specification is static, the leave–one–out jackknife method (JBC) in Hahn and Newey (2004) is applicable and serves as an additional comparison. The results show that indirect fixed effect estimators are closely comparable to ABC and JBC, while SBC produces estimates that are larger

[7]This variable serves as a proxy for permanent nonlabor income (Hyslop, 1999).

in magnitude. Because SBC achieves bias correction through sample splitting, the standard errors are computed by averaging the weighted Hessian matrices and thus become larger.

6 Monte Carlo Simulations

This section considers Monte Carlo simulations calibrated to the same PSID data. The details of calibration procedures are available in Appendix C.1. The indirect inference fixed effect estimator is compared with the fixed effect estimation, the analytical bias correction and two jackknife bias correction methods. All simulations are done 1000 times and H is set to 10. All statistics are relative to the true parameters and multiplied by 100.

Table 2: Simulation Results for Static LFP

	FE			IFE−10			IFE−20		
	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE
kids0_2	14.75	9.62	17.61	-4.78	8.18	9.47	-4.21	8.80	9.75
kids3_5	14.74	14.27	20.51	-6.83	13.12	14.79	-6.43	13.23	14.70
kids6_17	14.49	36.58	39.33	-18.06	38.81	42.78	-16.98	39.83	43.26
loghusinc	14.87	25.83	29.79	-3.34	26.76	26.96	-4.66	24.97	25.38
age	13.53	19.34	23.60	0.24	7.79	7.79	-0.87	9.98	10.01
age2	13.47	20.61	24.62	-2.25	22.51	22.61	-4.00	23.00	23.63

Notes: FE denotes fixed effects estimates. IFE−10 and IFE−20 denote indirect fixed effect estiamtes with H being 10 and 20. Simulations are conducted 1000 times, and all relative statistics are multiplied by 100.

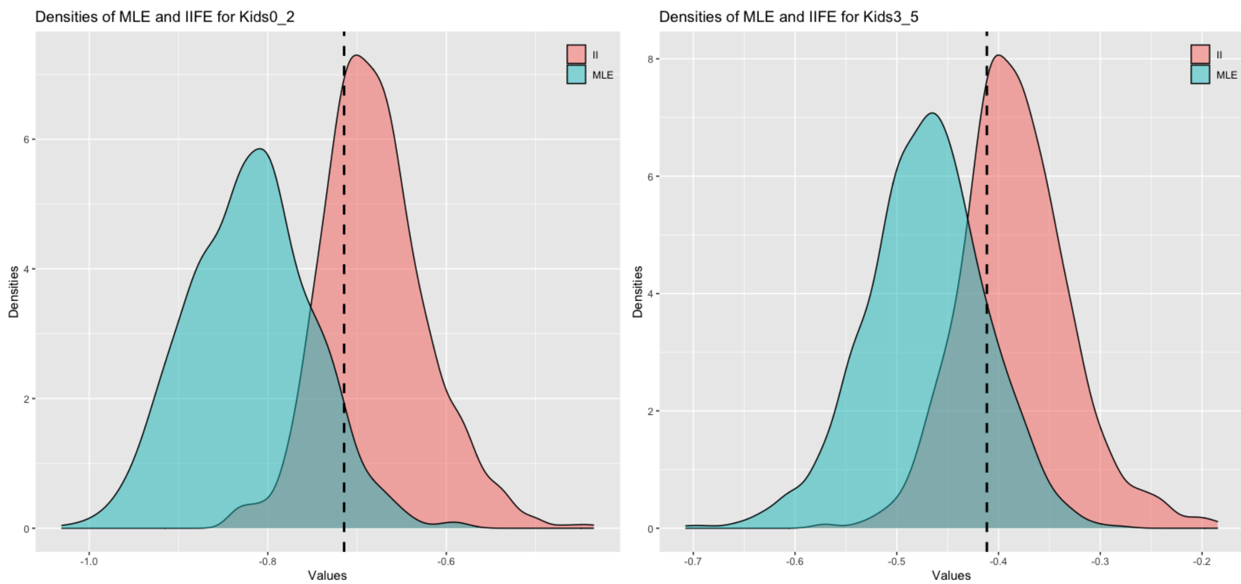
Table 3: Simulation Results for Static LFP

	ABC			JBC			SBC		
	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE
kids0_2	1.18	8.39	8.47	-3.46	8.09	8.79	-5.22	12.47	13.52
kids3_5	1.35	12.59	12.65	-3.33	12.17	12.61	-4.70	21.12	21.63
kids6_17	1.54	32.39	32.40	-3.50	31.12	31.30	-4.16	56.48	56.61
loghusinc	1.55	22.75	22.79	-3.43	21.87	22.12	-6.11	28.27	28.91
age	0.38	27.50	27.49	4.27	16.67	17.21	-4.02	34.81	35.03
age2	0.48	18.41	18.41	-4.36	17.78	18.30	-3.95	37.07	37.27

Notes: ABC denotes analytical bias correction in Fernández-Val (2009). JBC denotes leave−one−out jackknife bias correction in Hahn and Newey (2004) SBC denotes split−sample jackknife method. SBC denotes split−panel bias correction in Dhaene and Jochmans (2015). Simulations are conducted 1000 times, and all relative statistics are multiplied by 100.

Table (2) reports the simulation results of fixed effects and indirect fixed effect estimators. Fixed effect estimators are subject to a bias that is of the same order of magnitude as the standard deviation. This leads to severe under–coverage of the confidence intervals. The indirect fixed effect estimators, on the other hand, reduce bias by a margin without much inflation in the standard deviation. As an illustration, Figure (2) plots the densities of FE and IFE–10 estimates of $kids$0_2 and $kids$3_5, where the dashed lines denote the true values in the calibration exercise. The graphs show that fixed effect estimators are subject to a big bias, and thus the confidence intervals around them are not informative about the true coverage of the confidence intervals. On the other hand, because IFE–10 removes a large fraction of the bias, the recentering of confidence intervals admits a better coverage.

FIGURE 2: DENSITIES OF TWO FE AND IFE–10 POINT ESTIMATES



(A) VARIABLE: KIDS0_2 (B) VARIABLE: KIDS3_5

Table (3) tabulates the simulation results of analytical bias correction (ABC) and two jackknife bias correction methods. Compared with IFE, ABC unsurprisingly features smaller bias, but standard deviations are comparable. Turning to the other two methods that automatically correct bias, first note that JBC admits smaller bias and standard deviations than SBC. The simulation results are in line with Hughes and Hahn (2020), who theoretically showed that SBC has a larger higher–order variance and remaining bias than JBC[8]. On the other hand, IFE is comparable with JBC in terms of both bias and standard deviation.

[8] Therefore, in practice it is recommended to use panel bootstrap to obtain standard errors for SBC.

7 Conclusion

Fixed effect estimation of nonlinear panel models is subject to a large bias of point estimates and incorrect coverage of confidence intervals. This paper proposes an indirect fixed effect estimator that reduces the bias and obtains standard errors without bootstrap. The current theory is restricted to strictly exogenous explanatory variables, but Monte Carlo simulations in Appendix C.2 shows that the method can accommodate lagged dependent variables as well. Naturally, the next step is to extend the current theory to allow for dynamics in the DGP.

There are at least three other questions for further explorations. First, average partial effects are often the quantities of interest in nonlinear models. This paper establishes theoretical properties of finite dimensional parameters, and it could be interesting to explore if they can be extended to handle average partial effects, which is a function of explanatory variables, parameters of interest and incidental parameters.

Second, this paper directly works with non–smooth log likelihood function and establishes the asymptotic properties of the estimator. However, a practical concern of non–smoothness is that gradient–based optimization schemes cannot be used for estimation, and gradient–free schemes like Nelder–Mead face computational difficulty in high–dimensional problems. The indirect fixed effect estimator might benefit from smoothing approaches like kernel smoothing, but the theoretical justification can be nontrivial as smoothing can introduce an additional bias.

Finally, incorporating unobserved heterogeneity into dynamic discrete choice (DDC) models is an active area of research. One popular approach treats unobserved heterogeneity as an unobserved state variable and assumes individuals can be categorized into a finite number of types (Kasahara and Shimotsu, 2009; Arcidiacono and Miller, 2011). Introducing fixed effects circumvents the need to take a stand on the number of types, but can potentially complicate identification and estimation: the individual effects show up in both the current payoff and the continuation value, the latter of which has to be solved using a fixed–point algorithm. Duflo et al. (2012) It would be exciting to investigate whether some of the ideas in this paper can be applied to incorporate fixed effects into DDC models.

References

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005): "Competition and Innovation: an Inverted-U Relationship," *The Quarterly Journal of Economics*, 120, 701–728.

Altonji, J. G., A. A. Smith, and I. Vidangos (2013): "Modeling Earnings Dynamics," *Econometrica*, 81, 1395–1454.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum–Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 32, 283–301.

Andrews, D. W. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 4, chap. 37, 2247–2294.

Arcidiacono, P. and R. A. Miller (2011): "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity," *Econometrica*, 79, 1823–1867.

Arellano, M. and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, Cambridge University Press, chap. 12, 381–409.

Berger, D. and J. Vavra (2019): "Shocks versus Responsiveness: What Drives Time-Varying Dispersion?" *Journal of Political Economy*, 127, 2104–2142.

Bruins, M., J. A. Duffy, M. P. Keane, and A. A. Smith (2018): "Generalized Indirect Inference for Discrete Choice Models," *Journal of Econometrics*, 205, 177–203.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by J. Griliches and M. Intriligator, Elsevier, vol. 2, chap. 22, 1247–1318.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608.

Collard-Wexler, A. (2013): "Demand Fluctuations in the Ready-Mix Concrete Industry," *Econometrica*, 81, 1003–1037.

Dedecker, J. and S. Louhichi (2002): "Maximal Inequalities and Empirical Central Limit Theorem," in *Empirical Process Techniques for Dependent Data*, ed. by H. Dehling, T. Mikosch, and M. Sørensen, Springer, chap. 3, 137–160.

Dhaene, G. and K. Jochmans (2015): “Split–Panel Jackknife Estimation of Fixed-effect Models,” *The Review of Economic Studies*, 82, 991—-1030.

Dridi, R., A. Guay, and E. Renault (2007): “Indirect Inference and Calibration of Dynamic Stochastic General Equilibrium Models,” *Journal of Econometrics*, 136, 397–430.

Dridi, R. and E. Renault (2000): “Semi–Parametric Indirect Inference,” Unpublished manuscript.

Duffie, D. and K. J. Singleton (1993): “Simulated Moments Estimation of Markov Models of Asset Prices,” *Econometrica*, 61, 929–952.

Duflo, E., R. Hanna, and S. P. Ryan (2012): “Incentives Work: Getting Teachers to Come to School,” *American Economic Review*, 102, 1241–1278.

Fernández-Val, I. (2009): “Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models,” *Journal of Econometrics*, 150, 71–85.

Fernández-Val, I. and M. Weidner (2016): “Individual and time effects in nonlinear panel models with large N, T,” *Journal of Economics*, 192, 291–312.

——— (2018): “Fixed Effects Estimation of Large–T Panel Data Models,” *Annual Review of Economics*, 10, 109–138.

Forneron, J.-J. (2020): “A Sieve–SMM Estimator for Dynamic Models,” Working paper.

Frazier, D. T., T. Oka, and D. Zhu (2019): “Indirect Inference with a Non–Smooth Criterion Function,” *Journal of Econometrics*, 212, 623–645.

Gallant, R. A. and G. Tauchen (1996): “Which Moments to Match?” *Econometric Theory*, 12, 657–681.

Galvao, A. F. and K. Kato (2018): “Quantile Regression Methods for Longitudinal Data,” in *Handbook of Quantile Regression*, ed. by R. Koenker, V. Chernozhukov, X. He, and L. Peng, Chapman and Hall/CRC, chap. 19, 363–380.

Glasserman, P. and D. D. Yao (1992): “Some Guidelines and Guarantees for Common Random Numbers,” *Management Science*, 38, 884–908.

Gospodinov, N., I. Komunjer, and S. Ng (2017): “Simulated Minimum Distance Estimation of Dynamic Models with Errors-in-Variables,” *Journal of Econometrics*, 200, 181–193.

Gouriéroux, C. and A. Monfort (1997): *Simulation-based Econometric Methods*, Oxford University Press.

Gouriéroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, 85–118.

Gouriéroux, C., P. Phillips, and J. Yu (2010): "Indirect Inference for Dynamic Panel Models," *Journal of Econometrics*, 157, 68–77.

Guvenen, F. and A. A. Smith (2014): "Inferring Labor Income Risk and Partial Insurance from Economic Choices," *Econometrica*, 82, 2085–2129.

Hahn, J. and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T Are Large," *Econometrica*, 70, 1639–1657.

——— (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27, 1152 – 1191.

Hahn, J. and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," *The Quarterly Journal of Economics*, 123, 441–487.

Hughes, D. W. and J. Hahn (2020): "The Higher–Order Efficiency of Jackknife Bias Corrections in Panel Models," Working Paper.

Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294.

Kasahara, H. and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

Kato, K., A. F. Galvao Jr., and G. V. Montes-Rojas (2012): "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics*, 170, 76–91.

Kim, M. S. and Y. Sun (2016): "Bootstrap and k-step Bootstrap Bias Corrections for the Fixed Effects Estimator in Nonlinear Panel Data Models," *Econometric Theory*, 32, 1523–1568.

Lancaster, T. (2000): "The Incidental Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413.

Lee, B.-S. and B. F. Ingram (1991): "Simulation estimation of time-series models," *Journal of Econometrics*, 47, 197–205.

Manski, C. and S. Lerman (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, MIT Press, chap. 7, 2–50.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995–1026.

Newey, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1167.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 4, chap. 36, 2111–2245.

Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 5, 999–1033.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 874–897.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Schennach, S. M. (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–385.

Smith, A. A. (1993): "Estimating Nonlinear Time–Series Models Using Simulated Vector Autoregressions," *Journal of Applied Econometrics*, 8, 63–84.

Taber, C. and R. M. Sauer (2021): "Understanding Women's Wage Growth using Indirect Inference with Importance Sampling," *Journal of Applied Econometrics*.

van der Vaart, A. W. (2000): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series of Statistics.

White, H. (2000): *Asymptotic Theory for Econometricians: Revised Edition*, Emerald Publishing Limited.

A Auxiliary Results

A.1 Proof of Lemma 1

Lemma 1 (Uniform Convergence of Sample Criterion Function using Simulated Data)**.**

$$\max_{1\le i\le n}\sup_{(\beta,\gamma)}\left|\widehat{G}^h_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\xrightarrow{p}0,$$

where

$$\widehat{G}^h_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\ln f(y^h_{it}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma);$$
$$G_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\ln f(y_{it}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma).$$

Proof. The proof consists of two main steps. The first step deals with $\widehat{\alpha}_i$'s in data simulation and shows that $\widehat{G}^h_{(i)}$ is uniformly close to a criterion that uses α_{i0} to simulate the data, i.e.,

$$\widetilde{G}_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\int_U \ln f(y_{it}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF(u).$$

The second step is a uniform law of large number results showing that $\widetilde{G}_{(i)}(\beta,\gamma)$ uniformly converges to $G_{(i)}(\beta,\gamma)$.

Step 1: Given θ and a scalar τ, note that

$$\frac{1}{T}\sum_{t=1}^{T}\ln f(y^h_{it}(\theta,\tau)\mid x_{it};\beta,\gamma):=\frac{1}{T}\sum_{t=1}^{T}y^h_{it}(\theta,\tau)\ln\Phi(x_{it}'\beta+\gamma)+(1-y^h_{it}(\theta,\tau))\ln(1-\Phi(x_{it}'\beta+\gamma))$$

consists of two components: (1) an indicator function of scalar τ and (2) a smooth, bounded and monotone function of (β,γ). The indicator function $y^h_{it}(\theta,\widehat{\alpha}_i)$ belongs to type I class of Andrews (1994), which satisfies Pollard's entropy condition. The second component belongs to a class of functions satisfying bracketing entropy condition (van der Vaart and Wellner, 1996, Section 2.7.2).

Because $\frac{1}{T}\sum_{t=1}^{T}\ln f(y^h_{it}(\theta,\tau)\mid x_{it};\beta,\gamma)$ is an additive and multiplicative combination of the two classes of components, its function class also satisfies the entropy condition (Andrews, 1994), which is the primitive condition for stochastic equicontinuity. More specifically, define the following empirical process:

$$\nu_T(\tau) = \frac{1}{T}\sum_{t=1}^{T}\Big[\ln f(y_{it}^h(\theta,\tau) \mid x_{it};\beta,\gamma) - \int_U \ln f(y_{it}^h(\theta,\tau) \mid x_{it};\beta,\gamma)dF_u\Big],$$

where the integration is over the known distribution of simulation draws. By one of the equivalent definitions of stochastic equicontinuity (i.e., Andrews, 1994, p.2252), the following condition holds: for every sequence of constants $\{\delta_T\}$ that converges to zero,

$$\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma, |\tau_1-\tau_2|\leq\delta_T} \sqrt{T}|\nu_T(\tau_1) - \nu_T(\tau_2)| \xrightarrow{p} 0. \tag{A.1}$$

A first–order Taylor expansion on $\int_U \ln f(y_{it}^h(\theta,\alpha_{i0}) \mid x_{it};\beta,\gamma)$ with respect to α_{i0} around $\widehat{\alpha}_i$ yields

$$\begin{aligned}\int_U \ln f(y_{it}^h(\theta,\alpha_{i0}) \mid x_{it};\beta,\gamma)dF_u = &\int_U \ln f(y_{it}^h(\theta,\widehat{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u \\ &+ \frac{\partial \int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i - \alpha_{i0}).\end{aligned}$$

Combined with condition (A.1),

$$\begin{aligned}&\Big|\frac{1}{T}\sum_{t=1}^{T}\Big[\ln f(y_{it}^h(\theta,\widehat{\alpha}_i) \mid x_{it};\beta,\gamma) - \int_U \ln f(y_{it}^h(\theta,\alpha_{i0}) \mid x_{it};\beta,\gamma)dF_u\Big]\Big| \\ &= \Big|\nu_T(\widehat{\alpha}_i) - \frac{\partial \int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i - \alpha_{i0})]\Big| \\ &\leq |\nu_T(\widehat{\alpha}_i)| + \sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i - \alpha_{i0})\Big| \\ &= |\nu_T(\alpha_{i0}) + \nu_T(\widehat{\alpha}_i) - \nu_T(\alpha_{i0})| + \sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i - \alpha_{i0})\Big| \\ &\leq |\nu_T(\alpha_{i0})| + \sqrt{T}|\nu_T(\widehat{\alpha}_i) - \nu_T(\alpha_{i0})| + \Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i) \mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}\Big| \cdot |\widehat{\alpha}_i - \alpha_{i0}|.\end{aligned}$$

where the third and last lines are due to triangular inequality. Because $\nu_T(\alpha_{i0})$ is a normalized sum of mean zero random variables, $\nu_T(\alpha_{i0}) \xrightarrow{p} 0$ by LLN. The second term is the stochastic equicontinuity condition in Eq. (A.1). Because the derivative is bounded by Assumption 4 and $\max_{1\leq i\leq n}|\widehat{\alpha}_i - \alpha_{i0}| = o_p(1)$ (Hahn and Kuersteiner, 2011, Theorem 4), the third term is thus $o_p(1)$.

Therefore

$$\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma}\left|\frac{1}{T}\sum_{t=1}^{T}[\ln f(y_{it}^h(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma)-\int_U \ln f(y_{it}^h(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF_u]\right|\xrightarrow{p}0.$$

Step 2: The second part of the proof shows that

$$\max_{1\le i\le n}\sup_{(\beta,\gamma)}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\xrightarrow{p}0.$$

Following the the proof structure of Lemma 4 in Hahn and Kuersteiner (2011), note that

$$P\Big[\max_{1\le i\le n}\sup_{(\beta,\gamma)}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\ge\eta\Big]\le\sum_{i=1}^{n}P\Big[\sup_{(\beta,\gamma)}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\ge\eta\Big].$$

Since the parameter space is compact, it suffices to show that

$$\sup_{\Gamma_j}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\to0,$$

where Γ_j is a subset of $\mathcal{B}\times\Gamma_\gamma$ such that $\|\beta-\beta'\|\le\varepsilon$ and $|\gamma-\gamma'|\le\varepsilon$ for (β,γ) and $(\beta',\gamma')\in\Gamma_j$. By Assumption 4 on $G_{(i)}$,

$$\begin{aligned}
&\left|G_{(i)}(\beta,\gamma)-G_{(i)}^h(\beta',\gamma')\right|\le\mathbb{E}M(z_{it})|(\beta,\gamma)-(\beta',\gamma')|<\varepsilon\mathbb{E}M(z_{it}),\\
&\left|\widetilde{G}_{(i)}(\beta,\gamma)-\widetilde{G}_{(i)}(\beta',\gamma')\right|\le\frac{1}{T}\sum_{t=1}^{T}M(z_{it})|(\beta,\gamma)-(\beta',\gamma')|<\frac{\varepsilon}{T}\sum_{t=1}^{T}M(z_{it}).
\end{aligned}$$

By the triangular inequality,

$$\begin{aligned}
\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|&-\left|\widetilde{G}_{(i)}(\beta',\gamma')-G_{(i)}(\beta',\gamma')\right|\\
&\le\left|\Big(\widetilde{G}_{(i)}(\beta,\gamma)-\widetilde{G}_{(i)}(\beta',\gamma')\Big)-\Big(G_{(i)}(\beta,\gamma)-G_{(i)}(\beta',\gamma')\Big)\right|\\
&\le\left|\widetilde{G}_{(i)}(\beta,\gamma)-\widetilde{G}_{(i)}(\beta',\gamma')\right|+\left|G_{(i)}(\beta,\gamma)-G_{(i)}(\beta',\gamma')\right|\\
&<\varepsilon\mathbb{E}M(z_{it})+\frac{\varepsilon}{T}\sum_{t=1}^{T}M(z_{it})\\
&=\frac{\varepsilon}{T}\Big(\sum_{t=1}^{T}M(z_{it})-\mathbb{E}M(z_{it})\Big)+\frac{\varepsilon}{T}\mathbb{E}M(z_{it})+\varepsilon\mathbb{E}M(z_{it})\\
&<\frac{\varepsilon}{T}\Big|\sum_{t=1}^{T}M(z_{it})-\mathbb{E}M(z_{it})\Big|+2\varepsilon\mathbb{E}M(z_{it}).
\end{aligned}$$

Therefore by a rearrangement of the terms,

$$\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right| \leq \left|\widetilde{G}_{(i)}(\beta',\gamma')-G_{(i)}(\beta',\gamma')\right| + \frac{\varepsilon}{T}\left|\sum_{t=1}^{T} M(x_{it})-\mathbb{E}M(x_{it})\right| + 2\varepsilon\mathbb{E}M(x_{it}).$$

Let ε be such that $2\varepsilon \max_i \mathbb{E}M(z_{it}) < \frac{\eta}{3}$, then

$$\begin{aligned}
&P\Big[\sup_{\Gamma_j}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right| > \eta\Big]\\
\leq &P\Big[\left|\widetilde{G}_{(i)}(\beta',\gamma')-G_{(i)}(\beta',\gamma')\right| > \frac{\eta}{3}\Big] + P\Big[\frac{1}{T}\left|\sum_{t=1}^{T} M(x_{it})-\mathbb{E}M(x_{it})\right| > \frac{\eta}{3\varepsilon}\Big] + P\Big[2\varepsilon\mathbb{E}M(x_{it}) > \frac{\eta}{3}\Big]\\
=&o(T^{-2}),
\end{aligned}$$

where the last line follows as the first two terms on the right–hand side are $o(T^{-2})$ by Lemma 1 in Hahn and Kuersteiner (2011) and the last term is of probability zero by construction. Since $n = O(T)$,

$$\begin{aligned}
&P\Big[\max_{1\leq i\leq n}\sup_{(\beta,\gamma)}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right| \geq \eta\Big]\\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{m(\varepsilon)} P\Big[\sup_{\Gamma_j}\left|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right| \geq \eta\Big]\\
&= o(T^{-1})
\end{aligned}$$

□

A.2 Proof of Lemma 2

Lemma 2 (Pointwise Consistency of Auxiliary Estimator in the Simulation World)**.** $\forall \theta \in \Theta$,

$$\widehat{\beta}^h(\theta) \xrightarrow{p} \beta(\theta) = \theta.$$

Proof. The proof structure follows from that for Theorem 3 in Hahn and Kuersteiner (2011), with minor modification of notations. Fix $\eta > 0$ and set

$$\varepsilon = \inf_i \left[G_{(i)}(\theta, \alpha_{i0}) - \sup_{\{(\beta,\gamma): \|(\beta,\gamma)-(\theta,\alpha_{i0})\| > \eta\}} G_{(i)}(\beta, \gamma) \right] > 0$$

With probability $1 - o(T^{-1})$,

$$\begin{aligned}
\max_{\|\beta-\theta\|>\eta,\gamma_1,\ldots,\gamma_n} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta, \gamma_i) &\leq \max_{\|(\beta,\gamma_i)-(\theta,\alpha_{i0})\|>\eta} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta, \gamma_i) \\
&\leq \max_{\|(\beta,\gamma_i)-(\theta,\alpha_{i0})\|>\eta} \frac{1}{n}\sum_{i=1}^{n} G_{(i)}(\beta, \gamma_i) + \frac{1}{3}\varepsilon \\
&< \frac{1}{n}\sum_{i=1}^{n} G_{(i)}(\theta, \alpha_{i0}) - \frac{2}{3}\varepsilon \\
&< \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\theta, \alpha_{i0}) - \frac{1}{3}\varepsilon,
\end{aligned}$$

where the second and last inequalities are due to Lemma 1. By definition

$$\max_{\beta,\gamma_1,\ldots,\gamma_n} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta, \gamma_i) \geq \frac{1}{n}\sum_{i=1}^{n} G^h_{(i)}(\theta, \alpha_{i0}).$$

Hence

$$P\left[\|\widehat{\beta}^h(\theta) - \beta(\theta)\| \geq \eta \right] = o(T^{-1}).$$

□

B Proofs of Main Results

B.1 Proof of Proposition 1 (Uniform Consistency)

Proof. The main structure of the proof follows Theorem 1 in Newey (1991). The parameter space Θ is compact by assumption. The limiting function $\beta(\theta)$ is continuous since it is an identity function. Lemma 2 establishes the pointwise convergence result using simulated data: $\forall\theta \in \Theta$, $\widehat{\beta}^h(\theta) \xrightarrow{p} \beta(\theta)$. Therefore, it suffices to prove that $\widehat{\beta}^h(\theta)$ is stochastic equicontinuous.

By Markov inequality, $\forall\eta > 0$,

$$Pr\Big(\sup_{\theta\in\Theta}\|\widehat{\beta}^h(\theta) - \beta(\theta)\| > \eta\Big) \leq \frac{1}{\eta}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|\widehat{\beta}^h(\theta) - \beta(\theta)\|\Big).$$

Combined with the compactness assumption, it suffices to show that

$$\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\leq\delta}\|\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)\|\Big) \leq C\delta, \tag{B.1}$$

where δ denotes a positive scalar that is arbitrarily small and C is a constant. The rest of the proof consists of three parts. Firstly, a representation of $\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)$ in terms of profiled likelihood is established. Then, the question is transformed to bounding terms related to components of the profiled log likelihood. Lastly, the different pieces are glued together to give an expression of C.

Step 1: Let $\widehat{Q}(\widehat{\beta}^h(\theta);\theta)$ denote the profiled log likelihood function using simulated data h,

$$\widehat{Q}(\beta;\theta) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}^h(\theta,\widehat{\alpha}_i)\ln\Big(\Phi(x_{it}'\beta+\widehat{\gamma}_i(\beta)\Big)+(1-y_{it}^h(\theta,\widehat{\alpha}_i))\ln\Big(1-\Phi(x_{it}'\beta+\widehat{\gamma}_i(\beta))\Big) \tag{B.2}$$

Then by definition, $\widehat{\beta}^h(\theta_1)$ and $\widehat{\beta}^h(\theta_1)$ satisfy the first-order conditions,

$$\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_1);\theta_1)}{\partial\beta} = 0, \qquad \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} = 0$$

A first-order Taylor expansion yields

$$\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_1);\theta_1)}{\partial\beta} = 0 = \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta} + \frac{\partial^2\widehat{Q}(\widetilde{\beta};\theta_1)}{\partial\beta\partial\beta'}(\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)),$$

where $\widetilde{\beta}$ is between $\widehat{\beta}^h(\theta_1)$ and $\widehat{\beta}^h(\theta_2)$. Therefore,

$$\frac{\partial^2 \widehat{Q}(\widetilde{\beta};\theta_1)}{\partial\beta\partial\beta'}(\widehat{\beta}^h(\theta_1)-\widehat{\beta}^h(\theta_2)) = \frac{\partial \widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} - \frac{\partial \widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta}.$$

Let λ_s denote the smallest eigenvalue of the Hessian of the profiled likelihood, then a quadratic inequality leads to

$$\lambda_s \|\widehat{\beta}^h(\theta_1)-\widehat{\beta}^h(\theta_2)\| \le \left|\frac{\partial \widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} - \frac{\partial \widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta}\right|,$$

where $\frac{\partial\widehat{\gamma}_i(\widehat{\beta}(\theta_2))}{\partial\beta} = 0$ by the envelope theorem. For binary response panel probit models, some algebra leads to the following expression of the right–hand–side term in the absolute sign,

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)\Big)\Big(\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big), \tag{B.3}$$

where $y_{it}^h(\theta) = \mathbf{1}\{x_{it}'\theta+\widehat{\alpha}_i \ge u_{it}^h\}$ and u_{it}^h is from the standard normal distribution. Therefore, to establish Condition (B.1), it suffices to focus on Eq. (B.3).

Step 2: By the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)\Big)\Big(\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big)\Big|\Big)\\
&\qquad\le \sqrt{\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)\Big|^2\Big)}\times\\
&\qquad\sqrt{\mathbb{E}\Big(\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big|^2\Big)}.
\end{aligned}$$

For each i and t, the following two L^2–smoothness conditions hold:

$$\sqrt{\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}|y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)|^2\Big)} \le \sqrt{\frac{\mathbb{E}\|x_{it}\|_2}{\sqrt{2\pi}}}\sqrt{\delta}, \tag{B.4}$$

$$\sqrt{\mathbb{E}\Big(\Big|\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big|^2\Big)} \le K_2, \tag{B.5}$$

where $\|x\|_2$ denotes the L_2-norm $|x'x|^{1/2}$. This corresponds to type IV class in Andrews (1994).

Proving condition (B.4): Denote $\Delta\theta := \theta_2 - \theta_1$ and note that

$$\sup_{\|\theta_1-\theta_2\|\le\delta} |y_{it}^h(\theta_1) - y_{it}^h(\theta_2)| = \sup_{\|\Delta\theta\|\le\delta} |\mathbf{1}\{x_{it}'\theta_1 + \widehat{\alpha}_i \ge u_{it}^h\} - \mathbf{1}\{x_{it}'(\theta_1 + \Delta\theta) + \widehat{\alpha}_i \ge u_{it}^h\}|.$$

The direction that obtains the supremum is given by

$$\Delta\theta = \pm\frac{\delta}{\|x_{it}\|_2}x_{it}.$$

Therefore

$$\mathbb{E}\Big[\sup_{\|\theta_1-\theta_2\|\le\delta} |y_{it}^h(\theta_1) - y_{it}^h(\theta_2)|\Big] \le \mathbb{E}\Big(\mathbf{1}\{x_{it}'\theta_1 + \widehat{\alpha}_i \ge u_{it}^h\} - \mathbf{1}\{x_{it}'\theta_1 - \|x_{it}\|_2\delta + \widehat{\alpha}_i \ge u_{it}^h\}\Big). \quad \text{(B.6)}$$

Because δ is a scalar, a proof strategy à la Chen et al. (2003) is employed to bound right–hand–side term in Equation (B.6). More specifically, note that

$$\mathbf{1}\{x_{it}'\theta_1 + \widehat{\alpha}_i \ge u_{it}^h\} - \mathbf{1}\{x_{it}'\theta_1 - \|x_{it}\|_2\delta + \widehat{\alpha}_i \ge u_{it}^h\}$$

takes value either 1 or 0, and the expectation is the probability that the following event occurs:

$$x_{it}'\theta_1 + \widehat{\alpha}_i \ge u_{it}^h \ge x_{it}'\theta_1 - \|x_{it}\|_2\delta + \widehat{\alpha}_i.$$

Applying law of iterated expectation on the right–hand–side term and first–order Taylor expansion around δ,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\Big(\mathbf{1}\{x_{it}'\theta_1 + \widehat{\alpha}_i \ge u_{it}^h\} - \mathbf{1}\{x_{it}'\theta_1 - \|x_{it}\|_2\delta + \widehat{\alpha}_i \ge u_{it}^h\} \mid x_{it}', \widehat{\alpha}_i\Big)\Big] \\
=&\mathbb{E}\Big[\Phi(x_{it}'\theta_1 + \widehat{\alpha}_i) - \Phi(x_{it}'\theta_1 - \|x_{it}\|_2\delta + \widehat{\alpha}_i)\Big] \\
=&\mathbb{E}\Big[\phi(x_{it}'\theta + \widehat{\alpha}_i)\|x_{it}\|_2\Big]\delta
\end{aligned}$$

Therefore,

$$\mathbb{E}\Big[\sup_{\|\theta_1-\theta_2\|\le\delta} |y_{it}^h(\theta_1) - y_{it}^h(\theta_2)|\Big] \le \mathbb{E}\Big[\phi(x_{it}'\theta + \widehat{\alpha}_i)\|x_{it}\|_2\Big]\delta \le \frac{\mathbb{E}\|x_{it}\|_2}{\sqrt{2\pi}}\delta,$$

where the last inequality uses the fact that $\phi(\cdot) \le \frac{1}{\sqrt{2\pi}}$.

Proving condition (B.5): Note that

$$\sqrt{\mathbb{E}\Big(\Big|\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big|^2\Big)}$$

is no greater than

$$\sqrt{\mathbb{E}\Big(\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma}\Big|\frac{\phi(x_{it}'\beta+\gamma)x_{it}}{\Phi(x_{it}'\beta+\gamma)[1-\Phi(x_{it}'\beta+\gamma)]}\Big|^2\Big)},$$

which is bounded based on Lipschitz condition.

Step 3: Because the supremum of sum is no greater than sum of the supremum,

$$\begin{aligned}\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\Big|\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)\Big|^2\Big) &\le \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}|y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)|^2\Big)\\ &\le \frac{\delta}{\sqrt{2\pi}}\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\|x_{it}\|_2,\end{aligned}$$

$$\mathbb{E}\Big(\Big|\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\frac{\phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))]}\Big|^2\Big)\le\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}K_{it}.$$

Therefore,

$$\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\|\widehat{\beta}(\theta_1)-\widehat{\beta}(\theta_2)\|\Big)\le\frac{\sqrt{\delta}}{(2\pi)^{1/4}}\sqrt{\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\|x_{it}\|_2K_{it}}.$$

This verifies condition (B.1) and hence establishes the stochastic equicontinuity condition.

Step 4: By Theorem 1 in Newey (1991), $\widehat{\beta}^h(\theta)$ converges to $\beta(\theta)$ uniformly over $\theta\in\Theta$. □

B.2 Proof of Theorem 1 (Consistency)

Proof. Following the argument as in Appendix 1 of Gouriéroux et al. (1993),, consistency of $\widetilde{\theta}^H$ requires the following three conditions to hold:

1. the function $\beta(\theta)$ is invertible;
2. $\widehat{\theta}$ converges to $\beta(\theta_0)$ in $\theta_0 \in \Theta$ pointwise;
3. $\widehat{\beta}^h(\theta)$ converges to $\beta(\theta)$ uniformly over $\theta \in \Theta$.

The first condition is satisfied because function is an identity. The second condition only involves fixed effect estimator using observed data, and is a standard result in large−T panel literature (e.g, Hahn and Kuersteiner, 2011, Theorem 3). The third condition is verified by Proposition 1.

Recall that $\widetilde{\theta}^H$ is the solution to the optimization problem:

$$\widetilde{\theta}^H = \arg\min_{\theta\in\Theta}[\widehat{\theta} - \widehat{\beta}^H(\theta)]'[\widehat{\theta} - \widehat{\beta}^H(\theta)],$$

where $\widehat{\beta}^H(\theta) := \frac{1}{H}\sum_{h=1}^{H}\widehat{\beta}^h(\theta)$. Therefore. the limit of the optimization problem becomes

$$\min_{\theta\in\Theta}[\theta_0 - \theta]'[\theta_0 - \theta],$$

which has a unique solution θ_0. Therefore,

$$\widetilde{\theta}^H \xrightarrow{p} \theta_0.$$

□

B.3 Proof of Theorem 2 (Asymptotic Normality)

Proof. By Assumption 6 and consistency of $\widetilde{\theta}^H$,

$$\widehat{\theta} = \widehat{\beta}^H(\widetilde{\theta}^H) = \widehat{\beta}^H(\theta_0) + \mathbb{E}(\widehat{\beta}^H(\widetilde{\theta}^H) - \widehat{\beta}^H(\theta_0)) + o_p\Big(\frac{1}{\sqrt{nT}}\Big).$$

By mean–value theorem,

$$\mathbb{E}(\widehat{\beta}^H(\widetilde{\theta}^H) - \widehat{\beta}^H(\theta_0)) = \frac{\partial \mathbb{E}\widehat{\beta}^H(\overline{\theta})}{\partial \theta}(\widetilde{\theta}^H - \theta_0),$$

where $\overline{\theta}$ is between θ_0 and $\widetilde{\theta}^H$. Therefore,

$$\begin{aligned}
\sqrt{nT}(\widetilde{\theta}^H - \theta_0) &= -\Big[\frac{\partial \mathbb{E}\widehat{\beta}^H(\overline{\theta})}{\partial \theta}\Big]^{-1}\sqrt{nT}\Big(\widehat{\beta}^H(\theta_0) - \widehat{\theta}\Big) \\
&= \sqrt{nT}\Big(\widehat{\theta} - \widehat{\beta}^H(\theta_0)\Big) + o_p(1),
\end{aligned}$$

where the last equality uses the property that $\beta(\theta) = \theta$. Therefore, it suffices to focus on $\sqrt{nT}(\widehat{\theta} - \widehat{\beta}^H(\theta_0))$. Hahn and Kuersteiner (2011) derive the representation of $\widehat{\theta} - \theta_0$ as follows:

$$\widehat{\theta} - \theta_0 = \frac{A(\theta_0, \alpha_0)}{\sqrt{nT}} + \frac{B(\theta_0, \alpha_0)}{T} + o\Big(\frac{1}{T}\Big),$$

Because the same regression is run on simulated data,

$$\widehat{\beta}^h(\theta_0) - \theta_0 = \frac{A^h(\theta_0, \widehat{\alpha})}{\sqrt{nT}} + \frac{B^h(\theta_0, \widehat{\alpha})}{T} + o\Big(\frac{1}{T}\Big),$$

where $\widehat{\alpha} := (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$. This implies

$$\widehat{\beta}^H(\theta_0) = \beta(\theta_0) + \frac{1}{H}\sum_{h=1}^{H}\frac{A^h(\theta_0, \widehat{\alpha}_i)}{\sqrt{nT}} + \frac{1}{H}\sum_{h=1}^{H}\frac{B^h(\theta_0, \widehat{\alpha}_i)}{T} + o\Big(\frac{1}{T}\Big).$$

Therefore,

$$\begin{aligned}
\sqrt{nT}\Big(\widehat{\theta} - \widehat{\beta}^H(\theta_0)\Big) &= \Big(A(\theta_0, \alpha_0) - \frac{1}{H}\sum_{h=1}^{H}A^h(\theta_0, \widehat{\alpha})\Big) \\
&\quad + \sqrt{\frac{n}{T}}\Big(B(\theta_0, \alpha_0) - \frac{1}{H}\sum_{h=1}^{H}B^h(\theta_0, \widehat{\alpha})\Big) + o\Big(\sqrt{\frac{n}{T^3}}\Big).
\end{aligned}$$

The rest of the proof shows that bias term cancels out and the asymptotic normality holds. To simplify notation, the rest of the proof proceeds by setting $H = 1$.

The CLT and bias terms are a combination of high–order derivatives of the log likelihood function and takes the following form (Hahn and Kuersteiner, 2011):

$$A(\theta_0, \alpha_0) = \Big(\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i\Big)^{-1} \frac{1}{\sqrt{nT}} \sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}, \tag{B.7}$$

$$B(\theta_0, \alpha_0) = -\Big(\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{V_{it}}{\mathbb{E}\Big[\frac{\partial V_i}{\partial \alpha_i}\Big]}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big(\frac{\partial U_i}{\partial \alpha_i} - \frac{\mathbb{E}\Big[\frac{\partial^2 U_{it}}{\partial \alpha_i^2}\Big]}{2\mathbb{E}\Big[\frac{\partial V_i}{\partial \alpha_i}\Big]} V_{it}\Big)\right], \tag{B.8}$$

where

$$\mathcal{I}_i := -\mathbb{E}\left[\frac{\partial U_{it}}{\partial \theta'}\right], \quad U_{it} := U_i(y_{it}, x_{it}, \theta_0, \alpha_{i0}), \quad V_{it} := V_i(y_{it}, x_{it}, \theta_0, \alpha_{i0}),$$

and

$$\begin{aligned}
U_i(y_{it}, x_{it}; \theta, \alpha_i) &= \frac{\partial \ln f(y_{it} \mid x_{it}; \theta, \alpha_i)}{\partial \theta} - \rho_{i0} \cdot \frac{\partial \ln f(y_{it} \mid x_{it}; \theta, \alpha_i)}{\partial \alpha_i}, \\
V_i(y_{it}, x_{it}; \theta, \alpha_i) &= \frac{\partial \ln f(y_{it} \mid x_{it}; \theta, \alpha_i)}{\partial \alpha_i}, \\
\rho_{i0} &= \mathbb{E}\left[\frac{\partial^2 \ln f(y_{it} \mid x_{it}; \theta_0, \alpha_{i0})}{\partial \theta \partial \alpha_i}\right] \Big/ \mathbb{E}\left[\frac{\partial^2 \ln f(y_{it} \mid x_{it}; \theta_0, \alpha_{i0})}{\partial \alpha_i^2}\right].
\end{aligned}$$

The simulation analog that uses θ and $\tau := (\tau_1, \ldots, \tau_n)$ for data generation is

$$\begin{aligned}
B^h(\theta, \tau) &= -\Big(\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i(\theta, \tau)\Big)^{-1} \times \\
&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{V_{it}^h(\theta, \tau_i)}{\int_X\int_U \Big[\frac{\partial V_i^h(\theta, \tau_i)}{\partial \gamma_i}\Big] dF_u dF_x}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big(\frac{\partial U_i^h(\theta, \tau_i)}{\partial \gamma_i} - \frac{\int_X\int_U \Big[\frac{\partial^2 U_{it}^h}{\partial \gamma_i^2}\Big] dF_u dF_x}{2\int_X\int_U \Big[\frac{\partial V_i^h(\theta, \tau_i)}{\partial \gamma_i}\Big] dF_u dF_x} V_{it}^h(\theta, \tau_i)\Big)\right],
\end{aligned} \tag{B.9}$$

where

$$\begin{aligned}
&\mathcal{I}_i(\theta, \tau_i) := -\int_X\int_U \frac{\partial U_{it}^h}{\partial \beta'} dF_u dF_x, \quad U_{it}^h(\theta, \tau_i) := U_i(y_{it}^h(\theta, \tau_i), x_{it}, \theta, \tau_i), \\
&V_{it}^h(\theta, \tau_i) := V_i(y_{it}^h(\theta, \tau_i), x_{it}, \theta, \tau_i),
\end{aligned}$$

and

$$U_i(y_{it}^h(\theta,\tau_i),x_{it};\beta,\gamma_i) = \frac{\partial \ln f(y_{it}^h(\theta,\tau_i) \mid x_{it};\beta,\gamma_i)}{\partial \beta} - \rho_i(\theta,\tau_i)\cdot \frac{\partial \ln f(y_{it}^h(\theta,\tau_i) \mid x_{it};\beta,\gamma_i)}{\partial \gamma_i},$$

$$V_i(y_{it}^h(\theta,\tau_i),x_{it};\beta,\gamma_i) = \frac{\partial \ln f(y_{it}^h(\theta,\tau_i) \mid x_{it};\beta,\gamma_i)}{\partial \gamma_i},$$

$$\rho_i(\theta,\tau_i) = \left[\int_X\int_U \frac{\partial^2 \ln f(y_{it}^h(\theta,\tau_i) \mid x_{it};\theta,\tau_i)}{\partial\beta\partial\gamma_i} dF_u dF_x\right] \Bigg/ \left[\int_X\int_U \frac{\partial^2 \ln f(y_{it}^h(\theta,\tau_i) \mid x_{it};\theta,\tau_i)}{\partial\gamma_i^2} dF_u dF_x\right].$$

In the simulation world, the integration is taken over the distribution of sampling of observed explanatory variables and the distribution of simulation draws. Define

$$B_i^h(\theta,\tau_i) := \left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{V_{it}^h(\theta,\tau_i)}{\int_X\int_U \left[\frac{\partial V_i^h(\theta,\tau_i)}{\partial\gamma_i}\right] dF_u dF_x}\right]\times \left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\frac{\partial U_i^h(\theta,\tau_i)}{\partial\gamma_i} - \frac{\int_X\int_U\left[\frac{\partial^2 U_{it}^h(\theta,\tau_i)}{\partial\gamma_i^2}\right] dF_u dF_x}{2\int_X\int_U\left[\frac{\partial V_i^h(\theta,\tau_i)}{\partial\gamma_i}\right] dF_u dF_x} V_{it}^h(\theta,\tau_i)\right)\right]. \tag{B.10}$$

Step 1: Bias correction is established in Appendix B.4.

Step 2: The proof sketch is provided here. The simulation analog of the CLT term is

$$A^h(\theta_0,\widehat{\alpha}) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}^h(\theta,\widehat{\alpha}_i)$$

First for iid data, find an envelope for $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}^h(\theta,\widehat{\alpha}_i)$, then by Corollary 19.35 in van der Vaart (2000). Then the expectation of each simulated CLT term can be bounded a finite uniform entropy integral. These individual bounds are then aggregated to form a measure of approximation error that is asymptotically negligible.

Then use Berbee's coupling lemma (Dedecker and Louhichi, 2002, Lemma 4.1) to extend the previous results to mixing data by constructing i.i.d sequences that approximate its simulated time series with a controlled remainder. The complication is to take care of the extra approximation errors. By triangular inequality,

$$|A^h(\theta_0,\widehat{\alpha}) - A^h(\theta_0,\alpha_0)| \leq \left|\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\right| \cdot \left|\frac{1}{\sqrt{nT}}\sum_{i=1}^n\sum_{t=1}^T \big[U_{it}^h(\theta_0,\widehat{\alpha}_i) - U_{it}^h(\theta_0,\alpha_{i0})\big]\right| \tag{B.11}$$

$$+ \left|\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1} - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\right| \cdot \left|\frac{1}{\sqrt{nT}}\sum_{i=1}^n\sum_{t=1}^T U_{it}^h(\theta_0,\alpha_{i0})\right|. \tag{B.12}$$

Note that

$$\begin{aligned}
&\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1} - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\\
=&\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \big[\mathcal{I}_i(\theta_0,\alpha_{i0}) - \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\big]\Big)\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}.
\end{aligned}$$

By continuous mapping theorem, $\mathcal{I}_i(\theta_0,\widehat{\alpha}_i) \xrightarrow{p} \mathcal{I}_i(\theta_0,\alpha_{i0})$ for each i, and thus

$$\left|\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1} - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\right| \xrightarrow{p} 0.$$

Therefore term (B.12) is $o_p(1)$, and it suffices to focus on term (B.11). □

B.4 Proof of Proposition 2

Proof. Consider an infeasible fixed effect estimator $\widehat{\beta}^h(\theta_0, \alpha_0)$ that is obtained from data simulated by (θ_0, α_0). Then the representation of $\widehat{\beta}^h(\theta_0, \alpha_0) - \theta_0$ takes the form

$$\widehat{\beta}^h(\theta_0, \alpha_0) - \theta_0 = \frac{A^h(\theta_0, \alpha_0)}{\sqrt{nT}} + \frac{B^h(\theta_0, \alpha_0)}{T} + o\Big(\frac{1}{T}\Big),$$

where the superscript h denotes the fact that the dependent variable in $B^h(\theta_0, \alpha_0)$ is $y^h_{it}(\theta_0, \alpha_{i0})$. Because $B(\theta_0, \alpha_0)$ and $B^h(\theta_0, \alpha_0)$ have the same probability limit, they converge to the same expectation, which is the asymptotic bias. Therefore, it suffices to show that $B^h(\theta_0, \widehat{\alpha})$ uniformly well approximates $B^h(\theta_0, \alpha_0)$.

Now prove bias correction of the following form:

$$|B^h(\theta, \widehat{\alpha}) - B(\theta, \alpha_0)| \xrightarrow{p} 0.$$

By Markov inequality, $\forall \eta > 0$,

$$Pr(|B^h(\theta, \widehat{\alpha}) - B(\theta, \alpha_0)| \geq \eta) \leq \frac{1}{\eta}\mathbb{E}(|B^h(\theta, \widehat{\alpha}) - B(\theta, \alpha_0)|).$$

Therefore it suffices to bound the RHS term. By triangular inequality,

$$\mathbb{E}(|B^h(\theta, \widehat{\alpha}) - B(\theta, \alpha_0)|) \leq \mathbb{E}(|B^h(\theta, \widehat{\alpha}) - B^h(\theta, \alpha_0)|) + \mathbb{E}(|B^h(\theta, \alpha_0) - B(\theta, \alpha_0)|). \tag{B.13}$$

The second RHS term in equation (B.13) is $o_p(1)$ because $B^h(\theta, \alpha_0)$ and $B(\theta, \alpha_0)$ have the same probability limit. Regarding the first RHS term, by triangular inequality,

$$\begin{aligned}
\mathbb{E}|B^h(\theta, \widehat{\alpha}) - B^h(\theta, \alpha_0)| \leq \mathbb{E}\Big|&\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^n B_i^h(\theta_0, \widehat{\alpha}_i) - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^n B_i^h(\theta_0, \alpha_{i0}) \\
&+ \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^n B_i^h(\theta_0, \alpha_{i0}) - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \alpha_{i0})\Big)^{-1}\frac{1}{n}\sum_{i=1}^n B_i^h(\theta_0, \alpha_{i0})\Big| \\
\leq\ & \mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\Big| \cdot \Big|\frac{1}{n}\sum_{i=1}^n \big[B_i^h(\theta_0, \widehat{\alpha}_i) - B_i^h(\theta_0, \alpha_{i0})\big]\Big| && \text{(B.14)} \\
&+ \mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1} - \Big(\frac{1}{n}\sum_{i=1}^n \mathcal{I}_i(\theta_0, \alpha_{i0})\Big)^{-1}\Big| \cdot \Big|\frac{1}{n}\sum_{i=1}^n B_i^h(\theta_0, \alpha_{i0})\Big|. && \text{(B.15)}
\end{aligned}$$

Therefore, it suffices to focus on bounding terms (B.14) and (B.15).

For term (B.15), note that

$$
\begin{aligned}
&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\\
=&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\big[\mathcal{I}_i(\theta_0,\alpha_{i0})-\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\big]\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}.
\end{aligned}
$$

By continuous mapping theorem, $\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\xrightarrow{p}\mathcal{I}_i(\theta_0,\alpha_{i0})$ for each i. Therefore,

$$
\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\Big|\cdot\Big|\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0,\alpha_{i0})\Big|\xrightarrow{p}0.
$$

For term (B.14), note that

$$
\begin{aligned}
\Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\big]\Big|&\le\frac{1}{n}\sum_{i=1}^{n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|\\
&\le\max_{1\le i\le n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|\cdot\Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\big]\Big|\\
\le&\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|\cdot\max_{1\le i\le n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|\\
\le&\sqrt{\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|^2}\cdot\sqrt{\mathbb{E}\max_{1\le i\le n}\Big|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\Big|^2},
\end{aligned}
$$

where the second inequality is due to Cauchy–Schwarz inequality. Note that $B_i^h(\theta,\tau_i)$ has a finite number of parameters. Therefore, it suffices to bound

$$
\mathbb{E}\max_{1\le i\le n}\Big|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\Big|^2.
$$

Recall from expression (B.9) that some elements in $B_i^h(\theta,\widehat{\alpha}_i)$ are not smooth, namely $\partial U_{it}^h(\theta,\tau_i)/\partial\gamma_i$ and $V_{it}^h(\theta,\tau_i)$ that contain $\mathbf{1}\{x_{it}'\theta+\widehat{\alpha}_i\ge u_{it}^h\}$, which is a function of a scalar $\widehat{\alpha}_i$. The terms $\int_X\int_U\Big[\frac{\partial V_i^h(\theta,\tau_i)}{\partial\gamma_i}\Big]dF_u dF_x$, $\int_X\int_U\Big[\frac{\partial^2 U_{it}^h(\theta,\tau_i)}{\partial\gamma_i^2}\Big]dF_u dF_x$ and $\int_X\int_U\Big[\frac{\partial V_i^h(\theta,\tau_i)}{\partial\gamma_i}\Big]dF_u dF_x$ are differentiable. For panel probit models, algebra derivation of $B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})$ can lead to bounding $\mathbb{E}\sup_{|\widehat{\alpha}_i-\alpha_{i0}|\le\delta}\mathbf{1}\{x_{it}'\theta+\widehat{\alpha}_i\ge u_{it}^h\}-\mathbf{1}\{x_{it}'\theta+\alpha_{i0}\ge u_{it}^h\}$, which is bounded using the argument in

Chen et al. (2003).

Hence, $B^h(\theta_0, \widehat{\alpha}_i)$ uniformly approximates $B^h(\theta, \alpha_0)$. Because $B^h(\theta_0, \alpha_0)$ and $B(\theta_0, \alpha_0)$ have the same probability limit $\mathbb{E}(B(\theta_0, \alpha_0))$, which is the asymptotic bias term, $B^h(\theta_0, \widehat{\alpha})$ also does, and bias correction is completed. $\square$

C Computation Appendix

C.1 Calibration Procedures

Simulation procedures for the labor force participation application.

1. Run the regression on the LFP data to obtain $\widetilde{\theta}$ and $\widetilde{\alpha}_i$'s. These are treated as true coefficients for the calibration exercise.

2. For each simulation $s = 1, \ldots, S$, create a synthetic panel data based on the equation

$$y_{it}^s = \mathbf{1}\{X_{it}\widetilde{\theta} + \widetilde{\alpha}_i > u_{it}^s\},$$

where $u_{it}^s \sim iid\mathcal{N}(0,1)$. The data $\{(y_{it}^s, X_{it})\}$ are considered as the observed data for simulation s.

3. **Implementing the estimation:**

   (a) Run Probit regression on $\{(y_{it}^s, X_{it})\}$ and obtain $\widehat{\theta}^s$ and $\widehat{\alpha}_i^s$. This denotes the fixed effect estimators using observed data.

   (b) **Data simulation**:

      i. Given a set of parameter θ, simulate dependent variable using

$$y_{it}^h(\theta) = \mathbf{1}\{X_{it}'\theta + \widehat{\alpha}_i^s > \varepsilon_{it}^h\}, \quad \varepsilon_{it}^h \sim iid\mathcal{N}(0,1)$$

      Run Probit regression on $\{y_{it}^h(\theta), X_{it}\}$to obtain $\widehat{\beta}^h(\theta)$.

      ii. Repeat step (i) for $H = 10$ times and compute

$$\widehat{\beta}^H(\theta) = \frac{1}{H}\sum_{h=1}^{H}\widehat{\beta}^h(\theta).$$

      iii. Compute the indirect inference estimator $\widetilde{\theta}^H$ by solving the following equation

$$\widehat{\theta}^s = \widehat{\beta}^H(\widetilde{\theta}^H).$$

4. Repeat steps 2 and 3 for $S = 500$ times.

C.2 Simulations for Dynamic Labor Force Participation

This subsection introduces dynamics into the specification and compare the performance of indirect fixed effect estimators with other estimators.

Positive serial correlation observed in employment outcomes motivates the question of identifying state dependence, i.e., the causal impact of past employment on future employment for married women. However, the positive correlation can also be driven by individual–specific unobserved heterogeneity such as willingness to work. Therefore, an important question of interest is to distinguish between state dependence and persistent unobserved heterogeneity.

Following the empirical specification in Fernández-Val (2009), this paper controls for time–invariant unobserved heterogeneity by adding individual fixed effects,

$$y_{it} = \mathbf{1}\{X_{it}'\theta + \alpha_i \geq u_{it}\}, \quad u_{it} \sim \mathcal{N}(0,1), \tag{C.1}$$

where the vector of pre–determined covariates $X_{it} := (x_{it}, y_{i,t-1})$ now contains an extra variable: $y_{i,t-1}$, which denotes the lagged dependent variable. The first year of the sample is excluded for use as the initial condition in the dynamic model. In the data simulation step, the dependent variable at time t has the following representation:

$$y_{it}^h(\theta, \widehat{\alpha}_i) = \mathbf{1}\{\theta_1 y_{i,t-1}^h(\theta, \widehat{\alpha}_i) + x_{it}'\theta_{-1} + \widehat{\alpha}_i \geq u_{it}^h\}, \quad u_{it}^h \sim \mathcal{N}(0,1). \tag{C.2}$$

where θ_{-1} denotes parameters other than the one for $y_{i,t-1}^h$.

Table (4) reports the coefficients estimates using different methods. The analytical bias correction (ABC) corresponds to the method proposed by Hahn and Kuersteiner (2011) and serves as a benchmark. The JBC method by Hahn and Newey (2004) is no longer applicable due to dynamics in the specification. The results are similar to the static case. When $H = 20$, the indirect inference estimator produces bias correction results close to the ABC. On the other hand, the SBC estimate of lagged LFP is larger. Regarding the standard errors, the indirect fixed effect estimator does not inflate the errors when $H = 20$, but SBC has larger standard errors across all variables.

Table (5) reports the results of the Monte Carlo simulations. Compared to the static case in Table (2), adding dynamics into the regression further deteriorates fixed effect estimators of strictly exogenous covariates, which are comparable with the standard deviations. On the other hand, indirect fixed effect estimators correct the bias significantly. Compared to SBC, the reduction of bias is comparable but the standard deviation is smaller, which is consistent with the theory: by construction SBC does not use the whole sample for bias correction and thus inflates the variance.

Table 4: Parameter Estimates for Dynamic LFP

	lfp_lagged	kids0_2	kids3_5	kids6_17	loghusinc	age	age2
FE	0.76	-0.55	-0.28	-0.07	-0.25	2.05	-0.25
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)
IFE−1	0.80	-0.41	-0.25	-0.06	-0.31	2.04	-0.24
	(0.06)	(0.08)	(0.08)	(0.06)	(0.08)	(0.54)	(0.07)
IFE−10	1.09	-0.39	-0.07	-0.04	-0.32	1.78	-0.19
	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)	(0.40)	(0.05)
IFE−20	1.11	-0.48	-0.22	-0.07	-0.28	1.75	-0.23
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.39)	(0.05)
SBC	1.35	-0.63	-0.34	-0.15	-0.31	1.79	-0.20
	(0.05)	(0.09)	(0.09)	(0.08)	(0.07)	(0.88)	(0.12)
ABC	0.99	-0.48	-0.21	-0.06	-0.23	1.84	-0.22
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)

Notes: Standard errors are stored in the parenthesis and are computed based on the Hessian matrix of profiled log likelihood. For details of the SBC estimates and standard errors computation, refer to page 1025 in Dhaene and Jochmans (2015).

Table 5: Simulation Results for Dynamic LFP

	FE			IFE−10			SBC		
	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE
lfp_lagged	-53.59	5.84	54.01	3.06	6.22	7.19	-6.43	7.32	9.74
kids0_2	33.45	13.64	36.12	-5.81	9.69	11.30	7.62	17.27	18.87
kids3_5	47.88	24.37	53.72	-8.53	18.65	20.50	24.14	31.89	39.98
kids6_17	53.38	73.44	90.76	-23.29	55.91	60.54	33.74	98.06	103.99
loghusinc	24.08	28.90	37.61	5.29	44.90	45.19	5.70	31.67	32.16
age	29.49	19.34	26.91	1.44	5.54	5.72	-1.46	33.73	33.74
age2	29.07	26.91	39.61	-1.67	20.75	20.81	-1.04	36.54	36.63

Notes: FE denotes fixed effects estimates. IFE−10 denotes indirect fixed effect estiamtes with $H = 10$. SBC denotes split−sample jackknife method. Simulations are conducted 1000 times, and all relative statistics are multiplied by 100.