Journal of Econometrics 175 (2013) 116–131

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Identification and estimation of nonlinear dynamic panel data models with unobserved covariates[☆]

Ji-Liang Shiu[a,*], Yingyao Hu[b]

[a] Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing 100872, PR China

[b] Department of Economics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, United States

ARTICLE INFO

*Article history:*
Received 8 July 2010
Received in revised form
29 January 2013
Accepted 5 March 2013
Available online 19 March 2013

*Keywords:*
Nonlinear dynamic panel data model
Dynamic discrete choice model
Dynamic censored model
Nonparametric identification
Initial condition
Correlated random effects
Unobserved heterogeneity
Unobserved covariate
Endogeneity

ABSTRACT

This paper considers nonparametric identification of nonlinear dynamic models for panel data with unobserved covariates. Including such unobserved covariates may control for both the individual-specific unobserved heterogeneity and the endogeneity of the explanatory variables. Without specifying the distribution of the initial condition with the unobserved variables, we show that the models are nonparametrically identified from two periods of the dependent variable $Y_{it}$ and three periods of the covariate $X_{it}$. The main identifying assumptions include high-level injectivity restrictions and require that the evolution of the observed covariates depends on the unobserved covariates but not on the lagged dependent variable. We also propose a sieve maximum likelihood estimator (MLE) and focus on two classes of nonlinear dynamic panel data models, i.e., dynamic discrete choice models and dynamic censored models. We present the asymptotic properties of the sieve MLE and investigate the finite sample properties of these sieve-based estimators through a Monte Carlo study. An intertemporal female labor force participation model is estimated as an empirical illustration using a sample from the Panel Study of Income Dynamics (PSID).



## 1. Introduction

There are very few papers that provide full nonparametric identification of panel data models in the existing literature. This paper provides sufficient conditions for nonparametric identification of nonlinear dynamic models for panel data with unobserved covariates. These models take into account the dynamic processes by allowing the lagged value of the dependent variable as one of the explanatory variables as well as containing observed and unobserved permanent (heterogeneous) or transitory (serially-correlated) individual differences. Let $Y_{it}$ be the dependent variable at period $t$ and $X_{it}$ be a vector of observed covariates for individual $i$. We consider nonlinear dynamic panel data models of the form:

$$Y_{it} = g\left(X_{it}, Y_{it-1}, U_{it}, \xi_{it}\right), \quad \forall i = 1, \ldots, N;\ t = 1, \ldots, T-1, \tag{1}$$

where $g$ is an unknown nonstochastic function, $U_{it}$ is an unobserved covariate correlated with other observed explanatory variables $(X_{it}, Y_{it-1})$, and $\xi_{it}$ stands for a random shock independent of all other explanatory variables $(X_{it}, Y_{it-1}, U_{it})$. The focuses of the above model are on the cases in which the time dimension, $T$, is fixed and the cross section dimension, $N$, grows without bound. The unobserved covariate $U_{it}$ may contain two components as follows:

$$U_{it} = V_i + \eta_{it},$$

where $V_i$ is the unobserved heterogeneity or the random effects correlated with the observed covariates $X_{it}$ and $\eta_{it}$ is an unobserved serially-correlated component.

If the unobserved heterogeneity $V_i$ is treated as a parameter for each $i$, then both $V_i$ and other unknown parameters need to be estimated for the model (1). When $T$ tends to infinity, the MLE provides a consistent estimator for $V_i$ and other unknown parameters. However, $T$ is fixed and usually small for the panel data model considered here, and therefore, there are not enough observations to estimate these parameters. The model suffers from an incidental parameters problem (Neyman and Scott, 1948). In this paper, the unobserved heterogeneity, $V_i$, is treated as an unobservable random variable which may be correlated with observed covariates from the same individual. This correlated random

☆ Ji-Liang Shiu acknowledges support from the National Science Council of Taiwan via Grant 98-2410-H-194-118. The authors would like to thank Ryan Bush and Chin-Wei Yang for proofreading the draft. Helpful comments by Cheng Hsiao and three anonymous referees are acknowledged. All errors remain our own.

* Corresponding author. Tel.: +886 5 2720411; fax: +886 5 2720816.
*E-mail addresses:* jishiu@ruc.edu.cn (J.-L. Shiu), yhu@jhu.edu (Y. Hu).

0304-4076/$ – see front matter Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.03.001

effect[1] approach (treating $V_i$ as a random variable correlated with the covariates) allows us to integrate out unobserved variables to construct sieve MLE. This reduces potential computational burden from the incidental parameters problem for sieve MLE estimators in the estimation.[2] The transitory component $\eta_{it}$ may be a function of all the time-varying RHS variables in the history, i.e., $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,\ldots,t-1}\right)$ for some function $\varphi$.[3] Both observed explanatory variables $X_{it}$ and $Y_{it-1}$ become endogenous if the unobserved covariate $U_{it}$ is ignored. In this paper, we provide assumptions, including high-level injectivity restrictions, under which the distribution of $Y_{it}$ conditional on $(X_{it}, Y_{it-1}, U_{it})$, i.e., $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, is nonparametrically identified. The nonparametric identification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ may lead to that of the general form of our model (1) under certain specifications of the distribution of the random shock $\xi_{it}$.

In this paper we adopt the correlated random effect approach for nonlinear dynamic panel data models without specifying the distribution of the initial condition. We treat the unobserved covariate in nonlinear dynamic panel data models as the latent true values in nonlinear measurement error models and the observed covariates as the measurement of the latent true values.[4] We then utilize the identification results in Hu and Schennach (2008a), where the measurement error is not assumed to be independent of the latent true values. Their results rely on a unique eigenvalue–eigenfunction decomposition of an integral operator associated with joint densities of observable variables and unobservable variables. Hu and Shum (2010) uses an identification technique described in Carroll et al. (2010). The two identification strategies are different although both use the spectral decomposition of linear operators. The discussion of the difference in the two techniques can be found in Carroll et al. (2010). The conditional independence assumptions in Hu and Shum (2010) are more general than those here but their results require five periods of data in the comparable setting. Our assumptions are more suitable for panel data models. Although some of our assumptions are stronger, our estimator requires only two periods of the dependent variable $Y_{it}$ and three periods of the covariate $X_{it}$. This advantage is important because semi-nonparametric estimators usually require the sample size to be large.

The strength of our approach is that we provide nonparametric identification of nonlinear dynamic panel data model using two periods of the dependent variable $Y_{it}$ and three periods of the covariate $X_{it}$ without specifying initial conditions. The model may be described by, $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, the conditional distribution of the dependent variable of interest for an individual $i$, $Y_{it}$, conditional on a lagged value of that variable $Y_{it-1}$, explanatory variables $X_{it}$, and an unobserved covariate $U_{it}$. We show that $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ can be nonparametrically identified from a sample of $\left\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\right\}$ without parametric assumptions on the distribution of the individuals' dependent variable conditional on the unobserved covariate in the initial period. The main identifying assumption requires that the dynamic process of the covariates $X_{it+1}$ depends on the unobserved covariate $U_{it}$ but is independent of the lagged dependent variables $Y_{it}$, $Y_{it-1}$, and $X_{it-1}$ conditional on $X_{it}$ and $U_{it}$.

The identification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ leads to the identification of the general form of our model in Eq. (1). We present below two motivating examples in the existing literature. The specifications in these two types of models can be used to distinguish between dynamic responses to lagged dependent variables, observed covariates, and unobserved covariates. While the state dependence $Y_{it-1}$ reflects that experiencing the event in one period should affect the probability of the event in the next period, the unobserved heterogeneity $V_i$ represents individual's inherent ability to resist the transitory shocks $\eta_{it}$.

**Example 1** (*Dynamic Discrete-choice Model with an Unobserved Covariate*)**.** A binary case of dynamic discrete choice models is as follows:

$$Y_{it} = 1\left(X_{it}'\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it} \ge 0\right) \quad \text{with } \forall i = 1, \ldots, n;\ t = 1, \ldots, T-1,$$

where $1(\cdot)$ is the 0–1 indicator function and the error $\varepsilon_{it}$ follows an AR(1) process $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$ for some constant $\rho$. The conditional distribution of the interest is then

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = \left(1 - F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]\right)^{Y_{it}} \times F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{1-Y_{it}},$$

where $F_{\xi_{it}}$ is the CDF of the random shock $\xi_{it}$, $U_{it} = V_i + \eta_{it}$, and $\eta_{it} = \rho\varepsilon_{it-1}$. Empirical applications of the dynamic discrete-choice model above have been studied in a variety of contexts, such as health status (Contoyannis et al., 2004; Halliday, 2002), brand loyalty (Chintagunta et al., 2001), welfare participation (Chay et al., 2001), and labor force participation (Heckman and Willis, 1977; Hyslop, 1999). Among these studies, the intertemporal labor participation behavior of married women is a natural illustration of the dynamic discrete choice model. In such a model, the dependent variable $Y_{it}$ denotes the $t$-th period participation decision and the covariates $X_{it}$ are the nonlabor income or other observable characteristics in that period. The heterogeneity $V_i$ is the unobserved individual skill level or motivation, while the idiosyncratic disturbance $\xi_{it}$ denotes unexpected change of child-care cost or fringe benefit for married women from working. Heckman (1978, 1981a,b) has termed the presence of $Y_{it-1}$ "true" state dependence and $V_i$ "spurious" state dependence.

**Example 2** (*Dynamic Censored Model with an Unobserved Covariate*)**.** In many applications, we may have

$$Y_{it} = \max\left\{X_{it}'\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it}, 0\right\} \quad \text{with } \forall i = 1, \ldots, n;\ t = 1, \ldots, T-1,$$

with $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$. It follows that

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{\mathbf{1}(Y_{it}=0)} \times f_{\xi_{it}}\left[Y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it}\right]^{\mathbf{1}(Y_{it}>0)} \quad (2)$$

where $F_{\xi_{it}}$ and $f_{\xi_{it}}$ are the CDF and the PDF of the random shock $\xi_{it}$ respectively. The dependent variable $Y_{it}$ may stand for the amount of insurance coverage chosen by an individual or a firm's expenditures on R&D. In each case, an economic agent solves an optimization problem and $Y_{it} = 0$ may be an optimal corner solution. For this reason, this type of censored regression models is also called a corner solution model or a censored model with lagged censored

[1] In several studies, random effect means $V_i$ is a random variable independent of the explanatory variables. The discussion here is based on definitions on p. 286 of Wooldridge (2010).

[2] The estimation of an individual parameter $V_i$ along with other model parameters leads to an incidental parameters problem. Our sieve MLE has a feature of random effect, treating $V_i$ as a random variable and integrating out a composite unobserved variable to construct a likelihood function. Thus, the proposed sieve MLE has a computational advantage over a fixed effect approach because the individual parameter $V_i$ does not appear in the likelihood function.

[3] By the definition of $\eta_{it}$, $U_{it}$ might not only contain the error terms in panels but also some unobserved covariates from the past. Hence, $U_{it}$ denotes an unobserved covariate in this paper.

[4] An ideal candidate for the "measurement" of the latent covariate would be the dependent variable because it is inherently correlated with the latent covariate. However, such a measurement is not informative enough when the dependent variable is discrete and the latent covariate is continuous.