

Identification and Estimation of Nonlinear Dynamic Panel Data Models with Unobserved Covariates*

Ji-Liang Shiu[†] and Yingyao Hu[‡]

June 2011

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 three periods of data. The main identifying assumption requires 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).

Keywords: dynamic nonlinear panel data model, dynamic discrete choice model, dynamic censored model, nonparametric identification, initial condition, correlated random effects, unobserved heterogeneity, unobserved covariate, endogeneity, intertemporal labor force participation,

*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 Cheng Hsiao for helpful comments and Chin-Wei Yang for proofreading the draft. All errors remain our own.

[†]Department of Economics, National Chung-Cheng University, 168 University Rd. Min-Hsiung Chia-Yi, Taiwan. Email: jishiu@ccu.edu.tw.

[‡]Department of Economics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218. Email: yhu@jhu.edu.

1. Introduction

This paper considers 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:

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

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 ξ_{it} stands for a random shock independent of all other explanatory variables $(X_{it}, Y_{it-1}, U_{it})$. The focuses of the above models 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 η_{it} is an unobserved serially-correlated component.¹ The transitory component η_{it} may be a function of all the time-varying RHS variables in the history, i.e., $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,\dots,t-1}\right)$ for some function φ .² 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 reasonable assumptions 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 iden-

¹If unobserved heterogeneity, V_i , are assumed to be fixed, then both V_i along with other unknown parameters are needed to be estimated for the model (1). When T tends to infinity, the MLE is consistent. However, T is fixed and usually small for the panel data models here. There are not enough observations to estimate these parameters. The models suffer from an incidental parameters problem (Neyman and Scott (1948)). In this paper, the unobserved heterogeneity, V_i , are treated as random and may be correlated with the covariates from the same individual. We therefore concentrate on estimation of the parameters of the lagged dependent variable and the vector of observed covariates. This correlated random effect approach (treating V_i as random variable correlated with the covariates) allows us to integrate out unobserved variables once to construct sieve MLE. This reduces potential computational burden from the curse of dimensionality for sieve estimators.

²By the definition of η_{it} , U_{it} might not only contain the error terms in panels but some unobserved covariates in the past. Hence, U_{it} denotes an unobserved covariate in this paper.

tification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ may lead to that of the general form of our model in equation (1) under certain specifications of the distribution of the random shock ξ_{it} .

In the econometric literature, there are two approaches to tackling the unobserved heterogeneity V_i : random effects and fixed effects. In the fixed effect approach, much attention has been devoted to linear models with an additive unobserved effect. The problem can be solved by first applying an appropriate transformation to eliminate the unobserved effect and then implementing instrument variables (IV) in a generalized method of moments (GMM) framework. Anderson and Hsiao (1982), Arellano and Bond (1991), Arellano and Bover (1995) and Ahn and Schmidt (1995) employ an IV estimator on a transformation equation through first-differencing. Eliminating the unobserved effects is notably more difficult in nonlinear models and some progress has been made in this area. Rasch (1960) and Chamberlain (1980, 1984) considers a conditional likelihood approach for logit models with strictly exogenous assumption. Honoré and Kyriazidou (2000) generalize the conditional probability approach to estimate the unknown parameters without formulating the distribution of the unobserved individual effects or the probability distribution of the initial observations for certain types of discrete choice logit models. However, their results have to rely on a very strong assumption to match the explanatory variables in different time-periods. Their estimator is consistent and asymptotically normal but the rate of convergence is not the inverse of the square root of the sample size. Honoré (1993), Hu (2002) and Honoré and Hu (2004) obtain moment conditions for estimating dynamic censored regression panel data models. Altonji and Matzkin (2005) develop two estimators for panel data models with nonseparable unobservable errors and endogenous explanatory variables.

On the other hand, it is often appealing to take a random effect specification by making assumptions on the distribution of the individual effects. The main difficulty of this approach is the so-called initial conditions problem.³ With a relatively short panel, the initial con-

³The random effect approach for dynamic models requires the specification on the initial conditions of the process. Specifically, consider a special case of our model (1), dynamic discrete choice models without observed covariates X_{it} , in the following form:

$$Y_{it} = 1(\gamma Y_{it-1} + V_i + \xi_{it} \geq 0).$$

Then the conditional distribution $f_{Y_{it}|Y_{it-1},V_i}$ can be specified and the corresponding likelihood function has the structure

$$\mathcal{L} = \int f_{Y_{i0}|V_i} \prod_{t=1}^{T-1} f_{Y_{it}|Y_{it-1},V_i} f_{V_i} dv_i,$$

where $f_{Y_{i0}|V_i}$ denotes the marginal probability of Y_{i0} given V_i . If the process is not observed from the start

ditions have a very strong impact on the entire path of the observations but they may not be observed in the sample. One remedy to this problem is to specify the distribution of the initial conditions given the unobserved heterogeneity. The drawbacks of this approach are that the corresponding likelihood functions typically involve high order integration and misspecification of the distributions generally results in inconsistent parameter estimates. The associated computational burden of high order integration has been reduced significantly by recent advances in simulation techniques.⁴ Hyslop (1999) analyzes the intertemporal labor force participation behavior of married women using maximum simulated likelihood (MSL) estimator to simulate the likelihood function of dynamic probit models with a nontrivial error structure. Wooldridge (2005) suggests a general method for handling the initial conditions problem by using a joint density conditional on the strictly exogenous variables and the initial condition. Honoré and Tamer (2006) relax the distributional assumption of the initial condition and calculate bounds on parameters of interest in panel dynamic discrete choice models. Evdokimov (2009) considers a nonparametric panel data model with nonadditive unobserved heterogeneity: $Y_{it} = m(X_{it}, V_i) + \varepsilon_{it}$ where individual-specific effects are allowed to be correlated with the covariates in an arbitrary manner. The model has a different focus since our model explicitly includes lags of the endogenous dependent variable Y_{it-1} and a nonadditive ε_{it} .

Although the proposed model (1) is concerned with nonlinear dynamic panel data models, there are several directions in panel data models that are closer in spirit to our work. Chernozhukov, Fernández-Val, Hahn, and Newey (2009) derive bounds for marginal effects in nonlinear panel models and show that they can tighten rapidly as the number of time series observations grows. They also provide two novel inference methods that produce uniformly valid confidence regions in large samples. Hoderlein and White (2009) consider identification of marginal effects in general nonseparable models with unrestricted correlated unobserved effects even if there are only two time periods though their approach explicitly rules out lagged dependent variables. Arellano and Bonhomme (2009) provide a characterization of the class

then the initial state for individual i , y_{i0} cannot be assumed fixed. However, it is not clear that how to derive the initial condition $f_{Y_{i0}|V_i}$ from $f_{Y_{it}|Y_{it-1}, V_i}$ so it could be internally inconsistent across different time periods if the evolution of these two process can not be connected. Heckman (1981b) suggested that using a flexible functional form to approximate the initial conditions.

⁴See Gourieroux and Monfort (1993), Hajivassiliou (1993), Hajivassiliou and Ruud (1994) and Keane (1993) for the reviews of the literature.

of weights for nonlinear panel data models that produce first-order unbiased estimators. The approach requires at least as many time periods as regressors to estimate, while we only need three periods of data. Although the focuses of the models in this paper are on the fixed time dimension, the setting can be changed to large T cases. The recent large- T literature for dynamic panel models can be found in Hahn and Kuersteiner (2004), Carro (2007), and Fernández-Val (2009).

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.⁵ 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 Schennach (2008a) uses a technique in Carroll, Chen, and Hu (2010) also for identification of measurement error models. 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, Chen, and Hu (2010). The conditional independence assumptions in Hu and Schennach (2008a) 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 three periods of data. 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 three periods of data without specifying initial conditions. The model may be described in, $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, 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 $\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ without

⁵An 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.

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 equation (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 η_{it} .

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

$$Y_{it} = 1 (X'_{it}\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it} \geq 0) \quad \text{with} \quad \forall i = 1, \dots, n; t = 1, \dots, T - 1,$$

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

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

where $F_{\xi_{it}}$ is the CDF of the random shock ξ_{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, Jones, and Rice (2004), Halliday (2002)), brand loyalty (Chintagunta, Kyriazidou, and Perktold (2001)), welfare participation (Chay, Hoynes, and Hyslop (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 covariate X_{it} is the wage or other observable characteristics in that period. The heterogeneity V_i is the unobserved individ-

ual skill level or motivation, while the idiosyncratic disturbance ξ_{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

$$Y_{it} = \max \{X'_{it}\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it}, 0\} \quad \text{with} \quad \forall i = 1, \dots, n; t = 1, \dots, T - 1,$$

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

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

where $F_{\xi_{it}}$ and $f_{\xi_{it}}$ are the CDF and the PDF of the random shock ξ_{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 dependent variables.⁶ Honoré (1993) and Honoré and Hu (2004) use a method of moments framework to estimate the model without making distributional assumptions about V_i .

Based on our nonparametric identification results, we propose a semi-parametric sieve maximum likelihood estimator (MLE) for the model. We show the consistency of our estimator and the asymptotic normality of its parametric components. The finite sample properties of the proposed sieve MLE are investigated through Monte Carlo simulations of dynamic discrete choice models and dynamic censored models. Our empirical application focuses on how the labor participation decisions of married women respond to their previous participation states, fertility decisions, and nonlabor incomes. We develop and test a variety of dynamic econometric models using a seven year longitudinal sample from the Panel Study of Income Dynamics (PSID) in order to compare the results with those in Hyslop (1999). In the empirical

⁶This setting rules out certain types of data censoring. For example, if the censoring is due to top-coding, then it makes sense to consider a lagged value of the latent variable, i.e., $Y_{it}^* = X'_{it}\beta + \gamma Y_{it-1}^* + v_i + \varepsilon_{it}$ and $Y_{it} = \max[Y_{it}^*, c_t]$. This top-coded dynamic censored model has been considered in Hu (2000, 2002).

application, we examine three different model specifications, i.e., a static probit model, a maximum simulation likelihood (MSL) model, and a semi-parametric dynamic probit model. Our results find a large significant state dependence of labor force participation, smaller significant negative effects on nonlabor income variables, and also negative effects of children age 0-2 in the current period and past period.

The paper is organized as follows. We present the nonparametric identification of nonlinear dynamic panel data models in Section 2. Section 3 discusses our proposed sieve MLE. Section 4 provides the Monte Carlo study. Section 5 presents an empirical application describing the intertemporal labor participation of married women. Section 6 concludes. Appendices include proofs of consistency and asymptotic normality of the proposed sieve MLE and discussions on how to impose restrictions on sieve coefficients in the sieve MLE.

2. Nonparametric Identification

In this section, we present the assumptions under which the distribution of the dependent variable Y_{it} conditional on Y_{it-1} , covariates X_{it} , and the unobserved covariate U_{it} , i.e., $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, is nonparametrically identified. We start with a panel data containing three periods, $\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ for $i = 1, 2, \dots, n$. The law of total probability leads to

$$f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}} = \int f_{X_{it+1}|Y_{it}, X_{it}, Y_{it-1}, X_{it-1}, U_{it}} f_{Y_{it}|X_{it}, Y_{it-1}, X_{it-1}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}} dU_{it},$$

where we omit the arguments in the density function to make the expressions concise.

We assume

Assumption 2.1. (*Exogenous shocks*) *The random shock ξ_{it} is independent of $\xi_{i\tau}$ for any $\tau \neq t$ and $\{X_{i\tau}, Y_{i\tau-1}, U_{i\tau}\}$ for any $\tau \leq t$.*

As shown in the two examples above, this assumption has been used in many existing studies in the literature. However, it is still stronger than necessary. For the nonparametric identification of $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$, we only need $f_{Y_{it}|X_{it}, Y_{it-1}, X_{it-1}, U_{it}} = f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$, which is implied by Assumption 2.1. Given Eq. (1), the condition $f_{Y_{it}|X_{it}, Y_{it-1}, X_{it-1}, U_{it}} = f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$ holds if

the random shock ξ_{it} is independent of the covariate X_{it-1} . Assumption 2.1 then implies

$$f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}} = \int f_{X_{it+1}|Y_{it}, X_{it}, Y_{it-1}, X_{it-1}, U_{it}} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}} dU_{it}.$$

Furthermore, we simplify the evolution of the observed covariates X_{it} as follows:

Assumption 2.2. (*Covariate evolution*) $f_{X_{it+1}|Y_{it}, X_{it}, Y_{it-1}, X_{it-1}, U_{it}} = f_{X_{it+1}|X_{it}, U_{it}}$, where $U_{it} = V_i + \eta_{it}$ and the transitory component η_{it} contains all the time-varying variables in the past, i.e., $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,\dots,t-1}\right)$ for some function φ .

Note that the assumption can be also written as $X_{it+1} \perp (Y_{it}, Y_{it-1}, X_{it-1}) | (X_{it}, U_{it})$ and the lagged effects of Y_{it} such as $Y_{it-1}, Y_{it-2}, \dots$ enter the evolution of X_{it+1} through the unobserved covariate U_{it} . Assumption 2.2 may be decomposed into three steps. The first step is a Markov-type assumption $f_{X_{it+1}|Y_{it}, X_{it}, Y_{it-1}, X_{it-1}, U_{it}} = f_{X_{it+1}|Y_{it}, X_{it}, U_{it}}$, which implies that the evolution of the observed covariate X_{it+1} only depends on all the explanatory variables in the previous period (Y_{it}, X_{it}, U_{it}) . The implication of the Markov assumption is that the unobserved covariate U_{it} captures all the latent serially-correlated variation in the process of X_{it} . For example, suppose that we have⁷

$$X_{it} = \rho X_{it-1} + W_i + V_i + v_{it},$$

where v_{it} are i.i.d, and a latent W_i is not perfectly correlated with $V_i (= U_{it})$. The Markov-type assumption may not hold. Since both W_i and V_i are unobserved, the Markov assumption may hold if we redefine $W_i + V_i$ as V_i . The definition $Y_{it} = g(X_{it}, Y_{it-1}, U_{it}, \xi_{it})$ then implies that $f_{X_{it+1}|Y_{it}, X_{it}, U_{it}} = f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}, \xi_{it}}$.

The second step is that conditional on X_{it} and U_{it} , X_{it+1} is independent of the exogenous shock ξ_{it} . Since U_{it} contains all past shocks $\{\xi_{i\tau}\}_{\tau < t}$, this step only excludes the immediate effect of the current shock ξ_{it} of Y_{it} on the future covariate X_{it+1} .⁸ This implies that $f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}, \xi_{it}} = f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}}$. The third step is a limited feedback assumption, i.e., $f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}} = f_{X_{it+1}|X_{it}, U_{it}}$ which rules out direct feedback from the lagged dependent

⁷We thank an anonymous referee for suggesting this example.

⁸The assumption imposes some restriction to regressors in panel data setting. For example, suppose that $U_{it} = V_i$. The assumption that X_{it+1} is independent of ξ_{it} given X_{it} and V_i implies that $E[X_{it+1}\xi_{it}] = 0$. Thus, X_{it} can be a predetermined regressor from period $t-1$, in the sense that $E[X_{it}\xi_{is}] \neq 0$ for $s < t-1$ and zero otherwise.

variable Y_{it-1} on the future value of the observed covariate X_{it+1} . The effect of Y_{it-1} on X_{it+1} is indirectly through X_{it} , and U_{it} . This assumption rules out the case where X_{it} is a predetermined regressor.

An alternative view for the third step is to carefully exam the definitions of U_{it} . Recall that $U_{it} = V_i + \eta_{it}$ and $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,\dots,t-1}\right)$ for some function φ . From the panel data model setting Eq. (1), we have $Y_{it-1} = g(Y_{it-2}, X_{it-1}, U_{it-1}, \xi_{it-1})$. Since Y_{it-2} , X_{it-1} , and ξ_{it-1} are in η_{it} and U_{it-1} is in U_{it} , the information of the lagged dependent variable Y_{it-1} is contained in U_{it} . Therefore, $f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}} = f_{X_{it+1}|X_{it}, U_{it}}$. If X_{it} is strictly exogenous, i.e., X_{it} and ξ_{is} are independent for all t, s then Assumption 2.2 holds under the Markov assumption. Overall, Assumption 2.2 implies that conditional on X_{it} and U_{it} , X_{it+1} is independent of the exogenous shock ξ_{it} . In other words, conditional on the past information, the future covariate X_{it+1} rules out the immediate effect of the current shock ξ_{it} of the dependent variable Y_{it} .

In nonlinear models, the dependent variable Y_{it} may either be discrete or truncated, while at least part of covariates X_{it} is continuous. In the case of the intertemporal labor force participation behavior of married women in Example 1, Assumption 2.2 holds if we closely verify it through those three steps. Hence, such a simplification may not lose too much generality. In fact, the third step, i.e., $f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}} = f_{X_{it+1}|X_{it}, U_{it}}$, is for the purpose of simplification. Our identification strategy still applies with $f_{X_{it+1}|Y_{it}, X_{it}, Y_{it-1}, X_{it-1}, U_{it}} = f_{X_{it+1}|X_{it}, Y_{it-1}, U_{it}}$ in Assumption 2.2. In other words, we may allow the dependent variable in period $t-1$ to affect the evolution of the covariate X_{it} .

Assumption 2.2 then implies that

$$(3) \quad f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}} = \int f_{X_{it+1}|X_{it}, U_{it}} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}} dU_{it}.$$

Based on this equation, we may apply the identification results in Hu and Schennach (2008) to show the all the unknown densities on the RHS are identified from the observed density on the LHS. Let $\mathcal{L}^p(\mathcal{X}), 1 \leq p < \infty$ stand for the space of function $h(\cdot)$ with $\int_{\mathcal{X}} |h(x)|^p dx < \infty$.

For any $1 \leq p \leq \infty$ and any given $(y_{it}, x_{it}, y_{it-1})$, we define operators as follows:

$$L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} : \mathcal{L}^p(\mathcal{X}_{t-1}) \rightarrow \mathcal{L}^p(\mathcal{X}_{t+1})$$

$$(L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} h)(u) = \int f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}}(u, y_{it}, x_{it}, y_{it-1}, x) h(x) dx,$$

and

$$D_{y_{it}|x_{it}, y_{it-1}, U_{it}} : \mathcal{L}^p(\mathcal{U}) \rightarrow \mathcal{L}^p(\mathcal{U})$$

$$(D_{y_{it}|x_{it}, y_{it-1}, U_{it}} h)(u) = f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, u) h(u).$$

Similarly, define

$$(L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} h)(u) = \int f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}(u, x_{it}, y_{it-1}, x) h(x) dx,$$

$$(L_{X_{it+1}|x_{it}, U_{it}} h)(x) = \int f_{X_{it+1}|X_{it}, U_{it}}(x|x_{it}, u) h(u) du,$$

$$(L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}} h)(u) = \int f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}(x_{it}, y_{it-1}, x, u) h(x) dx.$$

Eq. (3) is equivalent to the following operator relationship:

$$L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} = L_{X_{it+1}|x_{it}, U_{it}} D_{y_{it}|x_{it}, y_{it-1}, U_{it}} L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}.$$

Integrating out Y_{it} in Eq. (3) leads to $f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}} = \int f_{X_{it+1}|X_{it}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}} dU_{it}$, which is equivalent to

$$L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} = L_{X_{it+1}|x_{it}, U_{it}} L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}.$$

with $(L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} h)(u) = \int f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}(u, x_{it}, y_{it-1}, x) h(x) dx$. We may then apply the results in Hu and Schennach (2008) to identify $f_{X_{it+1}|X_{it}, U_{it}}$, $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$, and $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}$ from $f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}}$. We assume

Assumption 2.3. (*Invertibility*) For any $(x_{it}, y_{it-1}) \in \mathcal{X}_{it} \times \mathcal{Y}_{it-1}$, $L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}$ and $L_{X_{it+1}|x_{it}, U_{it}}$ are invertible.

Intuitively, this assumption guarantees that the observables contains enough information on

the unobserved covariate U_{it} and the covariates in period $t + 1$, X_{it+1} , depends on itself in period t , X_{it} . The invertibility of $L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}$ is imposed on observables, while we are aware of the difficulty of testing the completeness property empirically. However, the operator is constructed by the density of highly correlated variables, X_{it+1} , X_{it} , Y_{it-1} , and X_{it-1} .⁹ Thus, the invertibility may only require functional form restrictions on $f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}$. For example, if \mathcal{X}_{t+1} contains an open set then $f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}} = \phi(X_{it-1} - \alpha_1 X_{it+1} - \alpha_2 X_{it} - \alpha_3 Y_{it-1})$ satisfies Assumption 2.3 where ϕ is the standard normal pdf and $\alpha_i \neq 0$.¹⁰ On the other hand, the invertibility of $L_{X_{it+1}|x_{it}, U_{it}}$ requires the covariates in period $t + 1$, X_{it+1} , contains enough information on the unobserved covariate U_{it} conditional on X_{it} . For example, we may have $X_{it+1} = X_{it} + U_{it} + h(X_{it})\epsilon_{it}$, where ϵ_{it} is independent of X_{it} and U_{it} and has a nonvanishing characteristic function on the real line. We use X_{it+1} instead of Y_{it+1} for the information on U_{it} because the dependent variable Y_{it+1} is discrete and U_{it} is continuous in many interesting applications. In that case, the operator mapping from functions of U_{it} to those of Y_{it+1} can't be invertible. On the other hand, when Y_{it+1} is continuous, it would be more reasonable to impose invertibility on the operator mapping from functions of U_{it} to those of Y_{it+1} , while U_{it} or V_i is allowed to be independent of the observed covariates X_{it} .¹¹ Necessary conditions for Assumption 2.3 include that $f_{X_{it+1}, Y_{it-1}, X_{it}|X_{it-1}} \neq f_{X_{it+1}, Y_{it-1}, X_{it}}$ and $f_{X_{it+1}|X_{it}, U_{it}} \neq f_{X_{it+1}|X_{it}}$. These necessary conditions rule out the case where X_{it+1} and X_{it-1} are independent or X_{it+1} and U_{it} are independent. In other words, Assumption 2.3 permits the existence of serial correlation among X_{it} and correlation between X_{it+1} and U_{it} .

There are more detailed discussions and general conditions for an invertible integral operator or complete conditional distributions in $\mathcal{L}^2(\mathcal{X})$ in Hu and Shiu (2011). In particular, the invertibility may require certain functional restrictions on the kernel function $f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}$. By Lemma 4 in Hu and Shiu (2011), general sufficient conditions for the injectivity of $L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}$ are as follows: for every $x_{it+1}, x_{it}, y_{it-1}$, let $f_{x_{it+1}, x_{it}, y_{it-1}, X_{it-1}}(\cdot)$

⁹The invertibility of $L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}$ can be justified by the fact that most variables in economics are correlated across time which reveal a pattern of serial correlation or autocorrelation.

¹⁰The result is from Theorem 2.2 in Newey and Powell (2003). Suppose that the distribution of x conditional on z is $N(a + bz, \sigma^2)$ for $\sigma^2 > 0$ and the support of z contains an open set, then the integral operator corresponding to $\phi_{\sigma^2}(x - a - bz)$ is invertible from $\mathcal{L}^p(\mathcal{X})$ to $\mathcal{L}^p(\mathcal{Z})$. There are more detailed discussions and general conditions for an invertible integral operator or complete conditional distributions in $\mathcal{L}^2(\mathcal{X})$ in Hu and Shiu (2011).

¹¹Assumption 2.3 requires $L_{X_{it+1}|x_{it}, U_{it}}$ is invertible and it demands the unobservable U_{it} to be correlated with the observed X_{it+1} . This case is complementary to the existing models where U_{it} is independent of X_{it+1} . Honoré and Kyriazidou (2000) and Honoré and Tamer (2006) identify the parameters under certain assumptions on the strictly exogenous covariates.

be in $\mathcal{L}^2(\mathbb{R})$ with

$$f_{x_{it+1}, x_{it}, y_{it-1}, X_{it-1}}(\cdot) \equiv f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}(x_{it+1}, x_{it}, y_{it-1}, \cdot).$$

Suppose that for every x_{it}, y_{it-1} , there exists a point x_{it+1}^0 with its open neighborhood $\mathcal{N}(x_{it+1}^0) \subset \mathcal{X}_{t+1}$ such that i) the Fourier transform $\phi_{x_{it+1}^0}(t)$ of $f_{x_{it+1}^0, x_{it}, y_{it-1}, X_{it-1}}(\cdot)$ satisfies $0 < |\phi_{x_{it+1}^0}(t)| < C e^{-\delta|t|}$ for all $t \in \mathbb{R}$ and some constants $C, \delta > 0$; ii) $\frac{\partial}{\partial x_{it+1}} f_{x_{it+1}, x_{it}, y_{it-1}, X_{it-1}}(\cdot)$ for $x_{it+1} \in \mathcal{N}(x_{it+1}^0)$ and $\frac{\partial}{\partial x_{it-1}} f_{x_{it+1}, x_{it}, y_{it-1}, X_{it-1}}(\cdot)$ are in $\mathcal{L}^2(\mathbb{R})$; iii) there exists a sequence $\{x_{it+1}^k : k = 1, 2, \dots\}$ of distinct $x_{it+1}^k \in \mathcal{N}(x_{it+1}^0)$ converging to x_{it+1}^0 such that the sequence $\{f_{x_{it+1}^k, x_{it}, y_{it-1}, X_{it-1}}(\cdot) : k = 1, 2, \dots\}$ is linearly independent, i.e.,

$$\sum_{j=1}^J c_j f_{x_{it+1}^k, x_{it}, y_{it-1}, X_{it-1}}(\cdot) = 0 \text{ for all } x_{it-1} \in \mathcal{X}_{t-1} \text{ implies } c_j = 0 \text{ for all } j = 1, 2, \dots, J.$$

Primitive conditions for the linear independence are summarized in Lemma 3 in Hu and Shiu (2011). For example, one of the sufficient conditions for the linear independence is

$$\lim_{x_{it-1} \rightarrow -\infty} \frac{f_{x_{it+1}^{k+1}, x_{it}, y_{it-1}, X_{it-1}}(x_{it-1})}{f_{x_{it+1}^k, x_{it}, y_{it-1}, X_{it-1}}(x_{it-1})} = 0.$$

Sufficient conditions for the invertibility of operator $L_{X_{it+1}|x_{it}, U_{it}}$ may also be derived similarly.

In addition, the invertibility of the operator $L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} = L_{X_{it+1}|x_{it}, U_{it}} L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}$ does imply restrictions on the initial condition through the operator $L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}$. For example, in a case where X_{it} and U_{it} are discrete and the linear operators are matrices, the invertibility of these operators are equivalent to the invertibility of corresponding matrices. However, the operators or matrices may still have a flexible form so that there is no need to specify the initial condition.

Note that when the unobserved component U_{it} is continuous-valued, the invertibility of $L_{X_{it+1}|x_{it}, U_{it}}$ implies that the set of the explanatory variables X_{it} contains a continuous element Z_{it} . The existence of the continuous component, Z_{it} is essential. It is impossible to nonparametrically identify a distribution of a continuous unobservable variable only by observed discrete variables. The restriction imposed on the continuous Z_{it+1} guarantees that the explanatory variables X_{it+1} contains enough information to identify unobserved compo-

ment U_{it} . A sufficient condition for identification with continuous-valued U_{it} can be obtained from the well-known completeness property of exponential families.¹² Thus, if \mathcal{X}_{it+1} is an open set then \mathcal{U}_{it} must be an open set and vice versa. In the case of the intertemporal labor force participation behavior of married women, since the covariate X_{it} contains wage and U_{it} includes the unobserved individual skill level or motivation the example can hold the invertibility property.

This assumption enables us to have

$$L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}^{-1} = L_{X_{it+1} | x_{it}, U_{it}} D_{y_{it} | x_{it}, y_{it-1}, U_{it}} L_{X_{it+1} | x_{it}, U_{it}}^{-1},$$

which implies a spectral decomposition of the observed operators on the LHS. The eigenvalues are the kernel function of the diagonal operator $D_{y_{it} | x_{it}, y_{it-1}, U_{it}}$ and the eigenfunctions are the kernel function $f_{X_{it+1} | X_{it}, U_{it}}$ of the operator $L_{X_{it+1} | x_{it}, U_{it}}$. In order to make the eigenvalues distinctive, we assume

Assumption 2.4. (*Distinctive eigenvalues*) *There exists a known function $\omega(\cdot)$ such that $E[\omega(Y_{it}) | x_{it}, y_{it-1}, u_{it}]$ is monotonic in u_{it} for any given (x_{it}, y_{it-1}) .*

The function $\omega(\cdot)$ may be specified by users, such as $\omega(y) = y$, $\omega(y) = I(y > 0)$, or $\omega(y) = y^2$. For example, we may have $\omega(y) = I(y = 0)$ in the two examples above. In both cases, $E[I(Y_{it} = 0) | x_{it}, y_{it-1}, u_{it}] = F_{\xi_{it}}[-(x'_{it}\beta + \gamma y_{it-1} + u_{it})]$, which is monotonic in u_{it} . Assumption 2.4 implies that for all $\hat{U}_{it}, \tilde{U}_{it} \in \mathcal{U}$, the set $\{y : f_{Y_{it} | X_{it}, Y_{it-1}, \hat{U}_{it}} \neq f_{Y_{it} | X_{it}, Y_{it-1}, \tilde{U}_{it}}\}$ for any given (x_{it}, y_{it-1}) has a positive probability whenever $\hat{U}_{it} \neq \tilde{U}_{it}$. Since the identification from the spectral decomposition is only identified up to u_{it} and its monotone transformation, we make a normalization assumption to pins down the unobserved covariate u_{it} .

Assumption 2.5. (*Normalization*) *For any given $x_{it} \in \mathcal{X}_{it}$, there exists a known functional G such that $G[f_{X_{it+1} | X_{it}, U_{it}}(\cdot | x_{it}, u_{it})] = u_{it}$.*

The functional G may be the mean, the mode, median, or a quantile. For example, we may have $X_{it+1} = X_{it} + U_{it} + h(X_{it})\epsilon_{it}$ with an unknown function $h(\cdot)$ and a zero median independent error ϵ_{it} . Then U_{it} is the median of the density function $f_{(X_{it+1} - X_{it}) | X_{it}, U_{it}}(\cdot | x_{it}, u_{it})$. In the case of the intertemporal labor force participation behavior of married women, this

¹²See Newey and Powell (2003) for details.

normalization imposes restrictions on the conditional density of wage. See Hu and Schennach (2008) for more discussion of this assumption.

Notice that Theorem 1 in Hu and Schennach (2008) implies that all three densities $f_{X_{it+1}|X_{it},U_{it}}$, $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, and $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$ are identified under the assumptions introduced above. The model of interest is described in the density $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$. While the initial condition at period $t - 1$ is contained in the joint distribution $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$, the evolution of the covariates X_{it} is described in $f_{X_{it+1}|X_{it},U_{it}}$. We summarize our identification results as follows:

Theorem 2.1. *Under Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, the joint distribution $f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}}$ uniquely determines the model of interest $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, together with the evolution density of observed covariates $f_{X_{it+1}|X_{it},U_{it}}$ and the initial joint distribution $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$.¹³*

Since the unobserved covariate U_{it} appearing in $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ does not have natural units of measurement or it is unclear which values are appropriate for U_{it} , in some economic problems the partial effects averaged across the distribution of U_{it} are more appealing. The average partial effects are based on the effect on a mean response after averaging the unobserved heterogeneity across the population. Suppose that we are interested in the conditional mean of $\omega(y_t)$, which is a scalar function of y_t . The average partial effect can be defined as:

$$(4) \quad \mu(X_t, Y_{t-1}) = E_{U_t} [E_{Y_t} [\omega(y_t) | X_t, Y_{t-1}, U_t] | X_t, Y_{t-1}].$$

Theorem 2.1 also allows us to go beyond estimation of the distribution of interest and to obtain estimated partial effects,

$$\mu(X_t, Y_{t-1}) = \int_{U_t} \left[\int_{Y_t} \omega(y_t) f_{Y_t|X_t, Y_{t-1}, U_t} dY_t \right] f_{U_t|X_t, Y_{t-1}} dU_t,$$

where $f_{Y_t|X_t, Y_{t-1}, U_t}$ and $f_{U_t|X_t, Y_{t-1}} = \frac{\int_{X_{t-1}} f_{X_t, Y_{t-1}, X_{t-1}, U_t} dX_{t-1}}{\int_{U_t} \int_{X_{t-1}} f_{X_t, Y_{t-1}, X_{t-1}, U_t} dX_{t-1} dU_t}$ are both identified from Theorem 2.1. These discussions lead to the following result:

Corollary 2.1. *Under Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, the average partial effect $\mu(X_t, Y_{t-1})$ defined in Eq. (4) can be identified and estimated by a panel data containing three periods,*

¹³The identification techniques is illustrated in Appendix A using a finite dimensional discrete example where the linear operators are matrices.

$\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ for $i = 1, 2, \dots, n$.

3. Estimation

The dynamic panel data model (1) specifies the relationship between the dependent variable of interest for an individual i , Y_{it} , and the explanatory variables including a lagged dependent variable Y_{it-1} , a set of possibly time-varying explanatory variables X_{it} , an unobserved covariate U_{it} . If we are willing to make a normality assumption on ξ_{it} , then the model in example 1 becomes a probit model and the model in example 2 becomes a tobit model. The general specification here covers a number of other dynamic nonlinear panel data model in one framework.

Given that the random shocks $\{\xi_{it}\}_{t=0}^T$ is exogenous, the conditional distribution $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$ is a combination of the function g and the distribution of ξ_{it} . In most applications, the function g and the distribution of ξ_{it} have a parametric form. That means the model may be parameterized in the following form,

$$f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta),$$

where θ includes the unknown parameters in both the function g and the distribution of ξ_{it} . Under the rank condition in the regular identification of parametric models, the nonparametric identification of $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$ implies that of the parameter θ , and therefore, the identification of the function g and the distribution of ξ_{it} . In general, we may allow $\theta = (b, \lambda)^T$, where b is a finite-dimensional parameter vector of interest and λ is a potentially infinite-dimensional nuisance parameter or nonparametric component.¹⁴ What is not specified in the model is the evolution of the covariate X_{it} , together with the unobserved component U_{it} , i.e., $f_{X_{it+1}|X_{it}, U_{it}}$, and the initial joint distribution of all the variables $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}$. We consider the nonparametric elements $(f_{X_{it+1}|X_{it}, U_{it}}, \lambda, f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}})^T$ as infinite-dimensional nuisance parameters in our semi-parametric estimator.

Our semi-parametric sieve maximum likelihood estimator (sieve MLE) does not require the initial condition assumption for the widely used panel data models, such as dynamic discrete-

¹⁴A partition of α_0 into finite-dimensional parameters and infinite-dimensional parameters does not affect our sieve MLE. More examples of a partition can be found in Shen (1997).

response models and dynamic censored models. In Section 2, we have shown equation (3) uniquely determines $(f_{X_{it+1}|X_{it},U_{it}}, f_{Y_{it}|X_{it},Y_{it-1},U_{it}}, f_{X_{it},Y_{it-1},X_{it-1},U_{it}})^T$. While the dynamic panel data model component $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ will be parameterized, the other components are treated as nonparametric nuisance functions. Eq. (3) implies

$$\begin{aligned}\alpha_0 &\equiv (f_{X_{it+1}|X_{it},U_{it}}, \theta, f_{X_{it},Y_{it-1},X_{it-1},U_{it}})^T \\ &= \arg \max_{(f_1, \theta, f_2)^T \in \mathcal{A}} E \ln \int f_1(x_{it+1}|x_{it}, u_{it}) f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta) \\ &\quad \times f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}) du_{it},\end{aligned}$$

which suggests a corresponding semi-parametric sieve MLE using an i.i.d. sample $\left\{ x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1} \right\}_{i=1}^n$,

$$\begin{aligned}(5) \quad \hat{\alpha}_n &\equiv \left(\hat{f}_1, \hat{\theta}, \hat{f}_2 \right)^T \\ &= \arg \max_{(f_1, \theta, f_2)^T \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n \ln \int f_1(x_{it+1}|x_{it}, u_{it}) f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta) \\ &\quad \times f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}) du_{it}.\end{aligned}$$

The function space \mathcal{A} contains the corresponding true densities and \mathcal{A}_n is a sequence of approximating sieve spaces.

Our estimator is a direct application of the general semi-parametric sieve MLE in Shen (1997), Chen and Shen (1998), and Ai and Chen (2003). In the appendix, we provide sufficient conditions for the consistency of our semi-parametric estimator $\hat{\alpha}_n$ and those for the \sqrt{n} asymptotic normality of the parametric component $\hat{\theta}$. The asymptotic theory of the proposed sieve MLE and the detailed development of sieve approximations of the nonparametric components are also provided in Appendix B.

With the consistency of the semi-parametric estimator $\hat{\alpha}_n$, a consistent estimator of the average partial effect can be obtained by

$$(6) \quad \hat{\mu}(X_t, Y_{t-1}) = \int_{U_t} \left[\int_{Y_t} \omega(y_t) f_{Y_t|X_t, Y_{t-1}, U_t}(y_t|x_t, y_{t-1}, u_t; \hat{\theta}) dY_t \right] \hat{f}_2(U_t|X_t, Y_{t-1}) dU_t,$$

where $\hat{f}_2(U_t|X_t, Y_{t-1}) = \frac{\int_{X_{t-1}} \hat{f}_2(x_t, Y_{t-1}, X_{t-1}, U_t) dX_{t-1}}{\int_{U_t} \int_{X_{t-1}} \hat{f}_2(x_t, Y_{t-1}, X_{t-1}, u_t) dX_{t-1} dU_t}$. Thus, the average partial effects of

the state dependence at interesting values of the explanatory variables can be computed by changes or derivatives of Eq. (6) with respect to Y_{t-1} .

Note that the proposed sieve MLE is ran on only 3 periods. This means that when a DGP is generated through the dynamic process (1), a three-periods data is enough to recovery the parameter of the interest θ . When there are more periods of data, the approach is still tractable. For example, if $T = 4$ and we assume the dynamic panel data specification (1) then estimation results form periods 1, 2, 3 should be the same as ones from 2, 3, 4. If the estimated results are significantly different, we would suspect model misspecification. Under the assumptions of stationary and ergodicity, an alternative way with data more than 3 periods is to transform the data into 3 periods of data by rearranging them as 3 periods of data and stacking them into a larger cross-sectional data. For example, suppose that there are 5 periods of data $\{D_t, D_{t+1}, D_{t+2}, D_{t+3}, D_{t+4}\}$. It can be transformed into three observations of three periods of data, i.e., $\{D_t, D_{t+1}, D_{t+2}\}$, $\{D_{t+1}, D_{t+2}, D_{t+3}\}$, and $\{D_{t+2}, D_{t+3}, D_{t+4}\}$. For model with a larger number of observed covariates, we can consider a single-index response model with $X'_{it}\beta$. That is: X_{it} is a d -dimensional vector of explanatory variables, $X'_{it}\beta$ is the index, the scalar product of X_{it} with β , a vector of parameters whose values are unknown. Since our assumptions do not exclude time dependence in covariates, time dummies are allowed to be in X_{it} . Many widely used parametric models have this form. In our empirical application, we adopt this approach to deal with a case of many observed covariates.

3.1. Implementation

As we discussed above, we propose a semi-parametric sieve MLE using an i.i.d. sample $\{x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1}\}$ for $i = 1, 2, \dots, n$. The unknown densities are associated with the observed distribution as follows:

$$f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}} = \int f_{X_{it+1}|X_{it}, U_{it}} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}} dU_{it}.$$

The parametric part is the model of interest $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta)$. The two non-parametric nuisance functions include $f_{X_{it+1}|X_{it}, U_{it}}$ and $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}$. The sieve MLE transforms a semi-parametric MLE to a parametric MLE by replacing the nonparametric nuisance functions with their Fourier approximations. For example, the sieve estimator for

the covariate evolution may be constructed by the Fourier series as follows:

$$f_1(x_{t+1}|x_t, u_t; \delta_1) = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \delta_{1,ijk} \varphi_{1i}(x_{t+1} - u_t) \varphi_{2j}(x_t) \varphi_{3k}(u_t),$$

where i_n, j_n, k_n are smoothing parameters and $\varphi_{1i}, \varphi_{2j}, \varphi_{3k}$ are known basis functions. Similarly, we may have a sieve approximation of the initial joint density, $f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2)$, where δ_2 is a vector of all the sieve coefficients. The fact that the parametric functions $f_1(x_{t+1}|x_t, u_t; \delta_1)$ and $f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2)$ are approximations of probability density functions implies certain restrictions on the sieve coefficients (δ_1, δ_2) , which is discussed in the Appendix C. In the sieve MLE, we may estimate $(\theta, \delta_1, \delta_2)$ as a parametric MLE with a density function as follows:

$$f(x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1}; \theta, \delta_1, \delta_2) = \int f_1(x_{it+1}|x_{it}, u_{it}; \delta_1) f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta) \times f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2) du_{it}.$$

In the Appendix B, we show the consistency and asymptotic normality as sample size goes to infinity.

4. Monte Carlo Evidence

In this section we present a Monte Carlo study that investigates the finite sample properties of the proposed sieve MLE estimators in the two different settings, dynamic discrete-choice models and dynamic censored models. We start with the specification of the models as follows.

Semi-parametric Dynamic Probit Models

First, we adopt a parametric assumption for ε_{it} . Suppose that ε_{it} has a stationary AR(1) with an independent Gaussian white noise process, $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$, $\xi_{it} \sim N(0, 1/2)$. We have

$$f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} = \Phi_{\xi_{it}}(X'_{it}\beta + \gamma Y_{it-1} + U_{it})^{Y_{it}} [1 - \Phi_{\xi_{it}}(X'_{it}\beta + \gamma Y_{it-1} + U_{it})]^{1-Y_{it}},$$

with $U_{it} = V_i + \rho\varepsilon_{it-1}$.

The density $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$ is fully parameterized and θ only contain the parametric com-

ponent $b = (\gamma, \beta)^T$. We approximate $f_{X_{it+1}|X_{it}, U_{it}}$, and $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}$ by truncated series in the estimation. The average partial effect in the dynamic probit model is

$$(7) \quad \hat{\mu}(X_t, Y_{t-1}) = \int_{U_t} \Phi_{\xi_t} \left(X_t' \hat{\beta} + \hat{\gamma} Y_{t-1} + U_t \right) \hat{f}_2(U_t | X_t, Y_{t-1}) dU_t,$$

which represents the conditional mean of $\omega(y_t) = y_t$.

Semi-parametric Dynamic Tobit Models:

We also assume that ε_{it} has a stationary AR(1) with an independent Gaussian white noise process, $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$. This gives

$$(8) \quad \begin{aligned} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} &= [1 - \Phi_{\varepsilon_{it}}(X_{it}'\beta + \gamma Y_{it-1} + U_{it})]^{\mathbf{1}(Y_{it}=0)} \phi_{\varepsilon_{it}}(y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it})^{\mathbf{1}(Y_{it}>0)} \\ &= \left[1 - \Phi \left(\frac{X_{it}'\beta + \gamma Y_{it-1} + U_{it}}{\sigma_{\xi}} \right) \right]^{\mathbf{1}(Y_{it}=0)} \times \\ &\quad \left[\frac{1}{\sigma_{\xi}} \phi \left(\frac{y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it}}{\sigma_{\xi}} \right) \right]^{\mathbf{1}(Y_{it}>0)}, \end{aligned}$$

and the parameter is $\theta = b = (\gamma, \beta, \sigma_{\xi}^2)^T$. Since $\xi_{it} \sim N(0, \sigma_{\xi})$, $E_{Y_i} [y_t | X_t, Y_{t-1}, U_t] = \Phi \left(\frac{X_t'\beta + \gamma Y_{t-1} + U_t}{\sigma_{\xi}} \right) (X_t'\beta + \gamma Y_{t-1} + U_t) + \sigma_{\xi} \phi \left(\frac{X_t'\beta + \gamma Y_{t-1} + U_t}{\sigma_{\xi}} \right)$. The average partial effect in the dynamic tobit model is

$$(9) \quad \begin{aligned} \hat{\mu}(X_t, Y_{t-1}) &= \int_{U_t} \left[\Phi \left(\frac{X_t' \hat{\beta} + \hat{\gamma} Y_{t-1} + U_t}{\hat{\sigma}_{\xi}} \right) (X_t' \hat{\beta} + \hat{\gamma} Y_{t-1} + U_t) \right. \\ &\quad \left. + \hat{\sigma}_{\xi} \phi \left(\frac{X_t' \hat{\beta} + \hat{\gamma} Y_{t-1} + U_t}{\hat{\sigma}_{\xi}} \right) \right] \hat{f}_2(U_t | X_t, Y_{t-1}) dU_t. \end{aligned}$$

The data generating process (DGP) for dynamic discrete choice models and dynamic censored models in the Monte Carlo experiments are according to the following processes respectively:

$$(10) \quad \begin{aligned} Y_{it} &= 1 (\beta_0 + \beta_1 X_{it} + \gamma Y_{it-1} + U_{it} + \xi_{it} \geq 0) \quad \text{with} \\ U_{it} &= V_i + \rho\varepsilon_{it-1} \quad \forall \quad i = 1, \dots, N; t = 1, \dots, T-1. \end{aligned}$$

and

$$(11) \quad \begin{aligned} Y_{it} &= \max \{ \beta_0 + \beta_1 X_{it} + \gamma Y_{it-1} + U_{it} + \xi_{it}, 0 \} \quad \text{with} \\ U_{it} &= V_i + \rho \varepsilon_{it-1} \quad \forall \quad i = 1, \dots, N; t = 1, \dots, T - 1. \end{aligned}$$

where $V_i \sim N(1, 1/2)$. To construct the sieve MLE, it is necessary to integrate out the unobserved covariate U_{it} . Here U_{it} has an unbounded domain $(-\infty, \infty)$ and we adopted Gauss-Hermite quadrature for approximating the value of the integral. Our generating processes of covariate evolution have the following form $X_{it+1} = X_{it} + h(X_{it})\varepsilon_{it} + U_{it}$ or

$$f_{X_{it+1}|X_{it}, U_{it}}(x_{t+1}|x_t, u) = \frac{1}{h(x_t)} f_\epsilon \left(\frac{x_{t+1} - x_t - u}{h(x_t)} \right),$$

where f_ϵ is a density function that can be specified under different identification conditions of Assumption 2.5.¹⁵ We consider the mode condition in this paper, and $f_\epsilon(x) = \exp(x - e^x)$ in all simulated data. In addition, we set $h(x) = 0.3 \exp(-x)$ to allow heterogeneity and assume the initial observation (y_0, x_0) and the initial component $\xi_0 (= \varepsilon_{i0})$ equal to zero.

We consider five different values of $(\gamma, \sigma_\xi^2, \rho)$ in the experiments: $(\gamma, \sigma_\xi^2, \rho) = (0, 0.5, 0)$, $(0, 0.5, 0.5)$, $(1, 0.5, 0)$, $(1, 0.5, 0.5)$, $(1, 0.5, -0.5)$ and the parameters in the intercept and the exogenous variable are held fixed: $\beta_0 = 0$ and $\beta_1 = -1$. In summary, the data generating processes are as follows:

$$\begin{aligned} \text{DGP I:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 0, 0.5, 0) \\ \text{DGP II:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 0, 0.5, 0.5) \\ \text{DGP III:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, 0) \\ \text{DGP IV:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, 0.5) \\ \text{DGP V:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, -0.5). \end{aligned}$$

The first two DGPs are not state dependent ($\gamma = 0$) while the rest are state dependent with $\gamma = 1$. Three different sample sizes N are considered: 250, 500, 1000. To secure a more stationary sample, the sampling data are drawn over $T = 7$ periods but only last three periods

¹⁵This generating process is also adopted in Hu and Schennach (2008a) and it can be adjusted to a variety of identification conditions, the mean, the mode, median, or a quantile.

are utilized. 100 simulation replications are conducted at each estimation.

Tables 1, 2, and 3 present simulation results under the semi-parametric probit model. The simulation results of DGP I (only allows for unobserved heterogeneity) show downward bias exists in the structural model coefficients (β_1, γ) for sample sizes $N=250$ and 1000 . For DGP II, the results have downward bias in the structural model coefficient β_1 . In addition, with nontrivial transitory component ($\rho \neq 0$) in DGP II, the standard errors of $(\beta_0, \beta_1, \gamma)$ are not much different from DGP I. As for DGPs with nontrivial state dependence, bias for $(\beta_0, \beta_1, \gamma)$ for these DGPs is around 0.01 or less and their standard errors are around 0.1. The coefficient estimators of γ in these DGPs have very small bias for all sample sizes, which means that our estimation for state dependence is very precise among processes with serial correlation ($\rho \neq 0$). In general, the means and medians of (β_1, γ) are very close to each other, reflecting little skewness in their respective distributions. Table 4 shows the simulation of the average partial effects in dynamic probit models in these DGPs. When there is no state dependence (DGP I & II), the estimates for average partial effects do not vary much with the lagged value Y_{t-1} . However, when DGPs contain state dependence, the different in the average response are up to 0.3.

Tables 5, 6, and 7 report the results of estimates for the semi-parametric tobit model. In the tobit model, there is negative bias in β_1 for all DGPs with trivial state dependence except for DGP I in $N=1000$. In tobit case, we have additional parameters to estimate, σ_ξ^2 . There is upward bias of the parameter in all DGPs and their standard errors are a little bit higher in DGPs with nontrivial state dependence. For these DGPs in positive state dependence, estimation results of γ show that there are small bias and precision is within 0.05. Also, the means and medians of all model parameters are not much different, reflecting low degree of skewness in distributions. Table 8 shows the results of the average partial effects in dynamic tobit models. There are higher standard errors of average partial effects conditional on \bar{y} in DGPs with positive state dependence. Our sieve MLE estimators tend to perform well.

There are two nuisance parameters, $f_{X_{t+1}|X_t, U_t}$ and $f_{X_t, Y_{t-1}, X_{t-1}, U_t}$ in our Monte Carlo simulation and we use Fourier series to approximate the evolution density and the square root of the initial joint distribution. Since a higher dimensional sieve space is constructed by tensor product of univariate sieve series, approximation series can be formed from several univariate Fourier series. In the semi-parametric probit model, while in the approximation

of the evolution densities we use three univariate Fourier series with the number of term, $i_n = 5$, $j_n = 2$, and $k_n = 2$, in the approximation of the initial joint distribution we have $i_n = 5$, $j_n = 2$, $k_n = 2$, and $l_n = 2$.¹⁶ While a formal selection rule for these smoothing parameters would be desirable, it is difficult to provide a general guideline. In general, one should pick the smoothing parameters to minimize the approximate mean squared errors of the estimator. In the Monte Carlo study, this is relatively easy to do because the true values are known. But in empirical applications where the true values of the parameters are unknown, it is still a difficult task. A rule of thumb is to pick the smoothing parameters such that the estimates are not much sensitive to small variations in the smoothing parameter. While the Fourier approximations to the evolution density $f_{X_{t+1}|X_t,U_t}$ have the density restriction and the identification restriction, there exists only density restriction for the approximations to the square root of the initial joint distribution $f_{X_t,Y_{t-1},X_{t-1},U_t}^{1/2}$ using Fourier basis. The semi-parametric sieve MLE using this construction does not encounter any negative integral inside the logarithm on Eq. (5) in our Monte Carlo study. As for the semi-parametric tobit model, we have similar choices of approximation series. The detailed sieve expression of these nuisance parameters can be found in Appendix C.

In summary, the Monte Carlo study shows that our semi-parametric sieve MLE performs well with a finite sample since mean and median estimates are close to the true values with reasonable standard errors.

5. Empirical Example

In this section, we apply our estimator to a dynamic discrete choice model, which describes the labor-force participation decisions of married women given their past participation state and other covariates. The advantage of our estimator is that our model may include (i) arbitrary and unspecified correlated random effects between unobserved time invariant factors such as skill level or motivation and time-varying X'_{it} s, and (ii) no initial conditions assumption.¹⁷ We will compare our estimates with those in Hyslop (1999), which studied a similar empirical

¹⁶The numbers of term, i_n , j_n , and k_n represent the length of three univariate Fourier series. See Appendix C for details.

¹⁷In Hyslop (1999), a correlated random-effects (CRE) specification for v_i is:

$$v_i = \sum_{s=0}^T (\delta_{1s} \cdot (\#Kids0-2)_{is} + \delta_{2s} \cdot (\#Kids3-5)_{is} + \delta_{3s} \cdot (\#Kids6-17)_{is}) + \sum_{s=0}^{T-1} \delta_{4s} \cdot y_{mtis} + \eta_i,$$

model with less general assumptions. The study in Hyslop (1999) also specified parametric forms of the unobserved heterogeneity V_i and AR(1) time dependence ρ of the transitory error component ε_{it} . Since these two terms are not separately identified from our main result Theorem 2.1, the empirical study here will emphasize on the parameters of exogenous explanatory variables and lagged dependent variable not the distributions of the error terms. On the other hand, these estimations might not comparable across specifications, because of the estimator-specific normalizations in binary choice models. Since the average partial effect is identified in Corollary 2.1, the empirical study also focuses on comparable average partial effects.

5.1. Data Descriptive

In order to provide comparison of the models developed in this paper and by Hyslop (1999), we also used the data related to waves 12-19 of the Michigan Panel Survey of Income Dynamics from the calendar years 1979-85 to study married women's employment decisions. The seven-year sample consists of women aged 18-60 in 1980, continuously married, and the husband is a labor force participant in each of the sample years. A woman is defined to be a labor market participant if she works for money any time in the sample year.¹⁸ The sample contains 1752 married women and also includes both the random Census subsample of families and the nonrandom Survey of Economic Opportunities (SEO) subsample of families.¹⁹

The number of possible binary participation sequences over a 7-period panel is $2^7 = 128$ and the sequences can be expressed as sequences of zeros and ones of the length 7.²⁰ If we partition the full sample based on all the observed annual participation outcomes of women during the seven-year period, the number of subsamples is up to 128. To provide a useful analysis of the differences of women's work propensity due to the number of years worked

where y_{mtis} is i 's transitory nonlabor income in year s . An alternative CRE specification can be:

$$v_i = \delta_1 \cdot (\overline{\#Kids0-2})_i + \delta_2 \cdot (\overline{\#Kids3-5})_i + \delta_3 \cdot (\overline{\#Kids6-17})_i + \delta_4 \cdot \bar{y}_{mti} + \eta_i,$$

where $\bar{x}_i = \sum_{t=0}^T x_{it}$.

¹⁸A standard definition of a participant is that an individual reports both positive annual hours worked and annual earnings. Hyslop (1995) provided a description of the extent of aggregation bias which results from ignoring intra-year labor force transition.

¹⁹Hyslop (1995) obtains a sample consisted of 1812 observations. The PSID contains an over-sample of low-income families called the Survey of Economic Opportunity (SEO).

²⁰An '1' in the t -th position of the sequence denotes participation in year t , while a '0' denotes nonparticipation.

and the associated participation sequences, we choose a small group of dividing criteria. The mutually exclusive sub-sample partition is as follows: we have in column (2) women who work in each year corresponding to a sequence '1111111'; in column (3), women who never work during the sample period corresponding to a sequence '0000000'; in column (4), women who experience a single transition from employment to nonemployment-that is, six participation sequences '1000000',..., '1111110'; in column (5), women who experience a single transition from nonemployment to employment corresponding to another six participation sequences '1000000',..., '1111110'; and in column (6), women who experience more than a single transition in their participation status corresponding to the rest of participation sequences.

Table 9 reports the descriptive statistics from the resulting subsamples. The selection of variables of interest in the table is close to the sample characteristics in Hyslop (1999) and the variables show similar trends and features. Column (1) presents the characteristics of the variables for the whole sample. Comparison of the observed annual participation outcomes with individual's independent participation decision form a binomial distribution with fixed probability of 0.7 (the average participation rate) indicates there is strong persistence in the married women's annual participation decisions. If there does not exist any persistence, then about 8 percent of the sample would be expected to work each year, and only 0.02 percent would not work at all, which are quite different from the sample relative frequencies, 47 percent and 9 percent respectively. In addition, the rest of the columns demonstrate the difference in the observable variables across the subsamples. For 825 women in the sample whom we observe employment in each of the seven years (column (2)), they are more likely to be better educated than average, Black, have fewer dependent children (especially children's age under 6 years), and their husbands' labor incomes are lower than average. Women who are never employed (column (3)) are older, less educated, and their husbands' labor income are higher than average. The women in this group have slightly fewer young children, reflecting the older age of the group. In column (4), women who make a single transition from employment to nonemployment have fewer dependent children but are more likely to have infant children (aged 0-2 years), and their husbands have above average earnings. Women who experience a single transition from nonemployment to employment (column (5)) are less likely to be black and have significantly more children (aged 0-17 years). The last column (6) indicates women who experience multiple employment transitions are younger, have more dependent children

of all ages, and their husbands have below-average labor income.

The description of the sample characteristics according to various subsamples suggests that there are several patterns between observable characteristics of individuals and their participation behavior. First, there is a negative income effect from husband's labor income on women's willingness for labor market participation (column (2) vs column (3)). Secondly, the presence of younger children tends to reduce the participation of women more except for women who never work in the sample (column (3)). The numbers of very young and older children between the single-transition subsamples in columns (4) and (5) are 0.34 and 0.24 (aged 0-2 years) and 0.67 and 1.21 (aged 6-17 years) respectively. The differences suggest that women leave employment to have children and re-enter employment as their children reach school age. The life-cycle interpretation is plausible by slight age difference between women in these groups (35.66 and 35.81). However, the age differences between these groups in Hyslop (1999) suggests that the composition of these samples is determined by more than simply fertility considerations. Finally, column (6) indicates that the presence of children in all age group (aged 0-17 years), together with low husband's labor income, increases the number of employment transitions of women.

5.2. Specifications and Estimation Results

According to a theoretic model in Hyslop (1999), the labor-force participation decisions of married women depend on whether or not their market wage offer exceeds their reservation wage, which in turn may depend on their past participation state, namely, suppose Y_t is the t -th period participation decision, W_t is the wage, and W_{0t}^* is a reservation wage then period t participation decision can be formulated by

$$(12) \quad Y_t = 1(W_t > W_{0t}^* - \gamma Y_{t-1})$$

where $1(\cdot)$ denotes an indicator function that is equal to 1 if the expression is true and 0 otherwise. An empirical reduced form specification for Eq. (12) is the following

$$Y_{it} = 1(X'_{it}\beta + \gamma Y_{t-1} + U_{it} + \xi_{it} > 0) \quad \forall i = 1, \dots, N; t = 1, \dots, T - 1$$

where X_{it} is a vector of observed demographic and family structure variables U_{it} captures the effects of unobserved factors, and β and γ are parameters. There are two latent sources for the unobserved term U_{it} :

$$U_{it} = V_i + \rho\varepsilon_{it-1}$$

where V_i is an individual-specific component, which captures unobserved time invariant factors possibly correlated with the time-varying X'_{it} s such as skill level or motivation; and ε_{it} is a serially correlated error term, which captures factors such as transitory wage movements.

The estimation results for the various models of labor force participation are presented in Table 10 which includes estimates from static probit models with random effect (column 1), maximum simulation likelihood (MSL) models with random effect²¹ (column 2), semi-parametric dynamic probit models (column 3). All specifications include unrestricted time effects, a quadratic in age, race, years of education, permanent and transitory nonlabor income y_{mp} & y_{mt} , current realizations of the number of children aged 0-2, 3-5, and 6-17, lagged realizations of the number of children aged 0-2.²² While the first two models are estimated using full seven years of data, the last one is estimated over three periods of data. In addition, the last model is the dynamic model without an initial conditions specification. The static probit model is estimated by MSL with 200 replications. It allows for individual-specific random effects but ignores possible dynamic effects of the past employment and potential correlation between the unobserved heterogeneity and the regressors. The results are that permanent nonlabor income has a significantly negative effect, transitory income reduces the contemporaneous participation, and preschool children have substantially negative effect. In addition, the variance of unobserved heterogeneity is 0.786. We now turn to dynamic specifications. The specifications in MSL model contain random effects, a stationary AR(1) error component, and first-order state dependence (SD(1)). The results show a large and significant first-order state dependence effect (1.117). The addition of SD(1) and AR(1) error component reduced the effects of nonlabor income variables largely (-0.007 & -0.004) and the contemporaneous fertility variables like $\#Kid3-5_t$ and $\#Kid6-17_t$ by approximately 50 percent.

²¹A detailed discussion of MSL models can be found in Hyslop (1999). There are more specifications in the paper. Here we only compare the models allowing the three sources of persistence.

²²The labor earnings of the husband are used as a proxy for nonlabor income. Permanent nonlabor income y_{mp} is estimated by the sample average, and transitory income y_{mt} is measured as deviations from the sample average

But the estimated effects of younger kids in the past and current periods $\#Kid0-2_{t-1}$ and $\#Kid0-2_t$ have stronger negative effects on women's participation decisions (-0.117 & -0.380). Including state dependence and serial correlation error component reduce the error variance (0.313) due to unobserved heterogeneity. The estimated AR(1) coefficient ρ is -0.146.²³

As the identification of the models hinges on assumptions in Section 2, a careful discussion of them in this labor force application is necessary. Assumption 2.1 is a model specification and it implies whatever is in X_{it}, Y_{it-1}, U_{it} , enough information have been included so that further lags of participation decision and the explanatory variables including nonlabor income, fertility status, etc, do not matter for explaining the current participation decision Y_{it} directly. Assumption 2.3 imposes functional form restrictions on the covariate evolution and the initial joint distribution. Assumption 2.4 in the empirical application may be $E[I(Y_{it} = 0) | x_{it}, y_{it-1}, u_{it}] = F_{\xi_{it}}[-(x'_{it}\beta + \gamma y_{it-1} + u_{it})]$, which is decreasing in u_{it} . Since u_{it} can represent or contain unobserved heterogeneity such as individual ability or motivation, the assumption suggests that the conditional expectation of absence from labor force decreases with ability or motivation. Our choice of G in Assumption 2.5 is the mode since the covariate X_{it} contain income variables. In Current Population Survey (CPS), it was found that the mode of misreported income conditional on true income is equal to the true income (see Bound and Krueger (1991) and Chen, Hong, and Tarozzi (2008)). Using the mode condition may relief concerns on measurement errors. Obviously, this is not the only choice of the functional G . As discussed before, we may use mean or median also.

We then focus on Assumption 2.2. The discussion of the assumption in Section 2 suggests that it imposes the key restriction that conditional on X_{it} and U_{it} , X_{it+1} is independent of the exogenous shock ξ_{it} and the lagged effects of Y_{it} such as $Y_{it-1}, Y_{it-2}, \dots$ enter the evolution of X_{it+1} through U_{it} . The regressors of interest in this empirical application are the nonlabor income variables and the fertility variables. There are several scenarios for the exogenous participation shock ξ_{it} . First, if ξ_{it} denotes measurement error, then the conditional independence between ξ_{it} and the future nonlabor income and fertility variables is plausible. Second, if ξ_{it} represents luck in labor markets such as unexpected change of child-care cost or fringe benefit for married women from working, the assumption rules out the immediate effect of the cur-

²³A correlated random-effects (CRE) is adopted in Hyslop (1999) to test the exogeneity of fertility with respect to participation decisions. His results show that there is no evidence against the exogeneity of fertility decision in dynamic model specifications.

rent shock ξ_{it} on the future nonlabor income and fertility variables. This implies that married women do not adjust their nonlabor income and fertility variables to the latest participation shock ξ_{it} but consider all other past period information. If there was a negative shock on participation, married women’s nonlabor income and fertility decisions would wait one period to response it. Therefore, Assumption 2.2 may be plausible in our model of the intertemporal labor force participation behavior of married women. Nevertheless, Assumption 2.2 does rule out the possible correlation between the fertility decisions in X_{it+1} and a negative shock on labor force participation ξ_{it} even conditioning on the fertility decisions in the previous period in X_{it} . While the lagged effects of Y_{it} enter the evolution of X_{it+1} indirectly here, our identification strategy still applies with $f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{X_{it+1}|X_{it},Y_{it-1},U_{it}}$ in Assumption 2.2 if Y_{it-1} has direct influence on X_{it+1} . This alternative specification implies that the labor force participation in period $t - 1$ to affect married women’s future nonlabor income and fertility decisions.

We then apply the sieve MLE method introduced in Section 3 & 4 and maintain a single-index form and a mode condition. The results also show a large first-order state dependence effect in the semi-parametric model (1.089). There exists a strong dependence between married women’s current labor force participation and past labor force participation and relaxing the initial conditions assumption increase the negative effects of nonlabor income variables and their significance in the dynamic models. Permanent income and transitory income both reduce the probability of participation but the effect of permanent nonlabor income has substantially greater magnitude.

Under the assumptions made in Section 2, the average partial effects are identified and we can obtain estimated partial effects at interesting values of the explanatory variables by Eq. (7). Table 10 reports the results of the average probabilities of being in the labor force in these specifications.²⁴ There are significant differences on the estimated probabilities. The static model does not have state dependence effect and it shows the estimated probability is 0.220. For a married woman not in the labor force at period $t - 1$ and with one kid aged 0-2, the estimated probabilities of being in the labor force at period t are 0.104 in the MSL model and 0.067 in the semi-parametric probit model, respectively. Comparing the results with a married woman in the labor force at period $t - 1$, the differences are estimates of state

²⁴The average partial effects are calculated using $X_t=(y_{mp}=\bar{y}_{mp}, y_{mt}=\bar{y}_{mt}, \#Kid0-2_{t-1}=0, \#Kid0-2_t=1, \#Kid3-5_t=0, \#Kid6-17_t=0, age_t=\text{mean of age}, education_t=\text{mean of education}, race_t=\text{Non-black})$.

dependence of being in the labor force. The magnitudes are 0.219 and 0.208 for the MSL model and the semi-parametric probit model respectively.²⁵

The fertility variables in the model are generally similar to those in column (1) and (2) but with less magnitude. That is: each of them has a significantly negative effect on married women's current labor force participation status, and younger children have stronger effect than older. In our semi-parametric probit model, the unobserved heterogeneity and the AR(1) component have been mixed into the unobserved covariate U_{it} . They are not identified so there are no any estimation results.

Although the model allows more flexible approach, its ability to predict the observed participation outcomes does not increase much. We compare frequencies of the participation outcomes predicted by the models in Table 11 to assess their fitting ability. Table 11 presents the frequencies of sample distribution and these predicted outcomes by the various estimated models over seven years period. Column (2) presents the predicted frequencies from the static probit model with random effect. The fraction of the predicted outcomes greatly over-predicts the frequencies of zero, one and six years worked and greatly under-predicts the frequencies of seven years worked. The results from the model MSL greatly under-predicts the frequencies of zero and seven year worked. As a result, the model substantially under-predicts the frequencies of the outcomes with no change in participation status over periods.

The final column in Table 11 contains the predicted frequencies from the semi-parametric probit model. The model predicts the frequencies in each participation outcome adequately for never work, and always work. It over-predicts the frequency of one and six years worked and under-predict the frequency of two, three, four, and five years worked. This is expected if there are larger lagged effect of participation decisions. Without initial conditions assumption, the model predicts the distribution of the number of years worked reasonably well. However, the predictive power from the model (column 3) without initial conditions assumption relative to the dynamic model with initial conditions assumption (column 2) is relatively small. One possible explanation of this is that the source of the serial persistence in participation outcomes over time is not well identified by those regressors. We might need other important regressors like child-care cost or welfare benefit from working.

In comparison to the results in the dynamic probit models allowing for CRE, AR(1),

²⁵The effect of state dependence is the marginal effect of the binary lagged dependent variable which is defined as the difference of the average responses, $\mu(X_t, Y_{t-1} = 1) - \mu(X_t, Y_{t-1} = 0)$.

and SD(1) in Hyslop (1999), adding unspecified CRE and avoiding initial conditions have significant effect on the model. Our results find a smaller significant negative effects on nonlabor income variables (-0.221 and -0.106 v.s. -0.285 and -0.140, respectively) and negative effects of children age 0-2 in the current period and past period which increases by 30 % (from -0.252 to -0.316) and decline by 50% (from -0.115 to -0.055) respectively. The changes of estimated parameters due to relaxing assumptions here are similar to the changes in Hyslop (1999).

6. Conclusion

This paper presents the nonparametric identification of nonlinear dynamic panel data models with unobserved covariates. We show the models are identified using only three periods of data without initial conditions assumptions, and we propose a sieve MLE estimator, which is applied to two examples, a dynamic discrete-choice model and a dynamic censored model. Both of them allow for three sources of persistence, "true" state dependence, unobserved individual heterogeneity ("spurious" state dependence), and possible serially correlated transitory error. Monte Carlo experiments have shown that how to deal with specific implementation issues and the sieve MLE estimators perform well for these models. Our sieve MLE is shown to be root n consistent and asymptotically normal. Finally, we apply our estimator to an intertemporal female labor force participation model using a sample from the Panel Study of Income Dynamics (PSID).

Appendix

A. Identification in the Discrete Case

We will show how to utilize the identification techniques in Section 2 for the discrete case. The discrete case refers to that the variables X_{it} and U_{it} is discrete:

$$X_{it} \in \mathcal{X}_t \equiv \{1, 2, \dots, J_1\} \text{ and } U_{it} \in \mathcal{U} \equiv \{1, 2, \dots, J_2\}.$$

In this finite dimensional discrete example, linear integral operators are matrices, which might be useful to give some intuition about how the identification is achieved. For simplicity, assume that $J_1 = J_2 = J$. Based on Eq. (3) which is the consequence of Assumption 2.1 and 2.2, the key equation of the discrete case is:

$$(13) \quad f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}} = \sum_{u_{it}=1}^J f_{X_{it+1}|X_{it}, U_{it}} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}.$$

Given $(y_{it}, x_{it}, y_{it-1})$, define J -by- J matrices

$$\begin{aligned} L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} &= [f_{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}}(u, y_{it}, x_{it}, y_{it-1}, x)]_{u, x} \\ L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} &= [f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}}(u, x_{it}, y_{it-1}, x)]_{u, x} \\ L_{X_{it+1}|x_{it}, U_{it}} &= [f_{X_{it+1}|X_{it}, U_{it}}(x|x_{it}, u)]_{x, u} \\ L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}} &= [f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}(x_{it}, y_{it-1}, x, u)]_{u, x} \end{aligned}$$

and a J -by- J diagonal matrix

$$D_{y_{it}|x_{it}, y_{it-1}, U_{it}} = \begin{bmatrix} f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, 1) & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}(y_{it}|x_{it}, y_{it-1}, J) \end{bmatrix}.$$

Using these matrixes, Eq. (13) can be expressed into a matrix notation as

$$(14) \quad L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} = L_{X_{it+1}|x_{it}, U_{it}} D_{y_{it}|x_{it}, y_{it-1}, U_{it}} L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}.$$

Integrating out Y_{it} in Eq. (13) leads to

$$(15) \quad f_{X_{it+1}, X_{it}, Y_{it-1}, X_{it-1}} = \sum_{u_{it}=1}^J f_{X_{it+1}|X_{it}, U_{it}} f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}.$$

which is equivalent to

$$(16) \quad L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}} = L_{X_{it+1}|x_{it}, U_{it}} L_{x_{it}, y_{it-1}, X_{it-1}, U_{it}}.$$

Assumption 2.3 guarantees that the above matrix $L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}}$ is invertible. It follows that

$$L_{X_{it+1}, y_{it}, x_{it}, y_{it-1}, X_{it-1}} L_{X_{it+1}, x_{it}, y_{it-1}, X_{it-1}}^{-1} = L_{X_{it+1}|x_{it}, U_{it}} D_{y_{it}|x_{it}, y_{it-1}, U_{it}} L_{X_{it+1}|x_{it}, U_{it}}^{-1}.$$

The observed matrix on the LHS has a matrix factorization, the product of a diagonal matrix with a matrix of eigenvectors. Uniqueness of the factorization requires the distinct eigenvalues and normalization of the unobserved covariate U_{it} . Assumption 2.4 and 2.5 are imposed to make these conditions hold. Since the eigenvalues and eigenvectors in the matrix factorization are $f_{Y_{it}|X_{it}, Y_{it-1}, U_{it}}$ and $f_{X_{it+1}|X_{it}, U_{it}}$ respectively, the identification of the model is reached. By Eq. (14), the initial joint distribution $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}}$ is also identified.

B. Asymptotic Properties of the Sieve Maximum Likelihood Estimator

This appendix presents the consistency of our estimator and the asymptotic normality of the parametric component of our estimator. Furthermore, we provide further details on the implementation of the semi-parametric sieve estimator, i.e., how to impose restrictions on the sieve coefficients. We basically applied the estimators in Chen, Hong, and Tamer (2005) and Hu and Schennach (2008a) to our model. Therefore, we refer to those papers for technical details.

Our asymptotic analysis relies on regularity restrictions on function containing the parameters of interest α . First, we introduce a typical space of smooth functions, Hölder space. Given a $d \times 1$ vector of nonnegative integers, $a = (a_1, \dots, a_d)'$ and denote $[a] = a_1 + \dots + a_d$ and let D^a denote the differential operator defined by $D^a = \frac{\partial^{[a]}}{\partial \xi_1^{a_1} \dots \partial \xi_d^{a_d}}$. Let m denote the largest integer satisfying $\gamma > \underline{\gamma}$ and set $\gamma = \underline{\gamma} + p$. The Hölder space $\Lambda^\gamma(\nu)$ of order $\gamma > 0$ is a collection of functions which are m times continuously differentiable on ν and the $\underline{\gamma}$ -th derivative are Hölder continuous with the exponent p . The Hölder space becomes a Banach space with the Hölder norm, i.e., $\forall g \in \Lambda^\gamma(\nu)$

$$(17) \quad \|g\|_{\Lambda^\gamma} = \sup_{\xi \in \nu} |g(\xi)| + \max_{a_1 + \dots + a_d = \underline{\gamma}} \sup_{\xi \neq \xi' \in \nu} \frac{|D^a g(\xi) - D^a g(\xi')|}{\|\xi - \xi'\|_E^p}.$$

The weighted Hölder norm is defined as $\|g\|_{\Lambda^{\gamma,\omega}} \equiv \|\tilde{g}\|_{\Lambda^\gamma}$ for $\tilde{g}(\xi) \equiv g(\xi)\omega(\xi)$ and the corresponding weighted Hölder space is $\Lambda^{\gamma,\omega}(\nu)$. Define a weighted Hölder ball as $\Lambda_c^{\gamma,\omega}(\nu) \equiv \{g \in \Lambda^{\gamma,\omega}(\nu) : \|g\|_{\Lambda^{\gamma,\omega}} \leq c < \infty\}$. Let $\varepsilon \in \mathbb{R}$, and $W \in \mathcal{W}$ with \mathcal{W} a compact convex subset in \mathbb{R}^{d_w} . Define the following spaces:

$$\begin{aligned}\mathcal{F}_1 &= \{f_1(\cdot|\cdot, \cdot) \in \Lambda_c^{\gamma_1,\omega}(\mathcal{X} \times \mathcal{X} \times \mathcal{U}) : f_1(\cdot) \geq 0 \text{ and } \int_{\mathcal{X}} f_1(\cdot|x, u)dx = 1, \forall(x, u) \in \mathcal{X} \times \mathcal{U}\}, \\ \mathcal{M} &= \{\lambda(\cdot) \in \Lambda_c^{\gamma_m,\omega}(\mathbb{R}) : 0 \leq \lambda(\cdot) \leq 1, \lambda(-\infty) = 0 \text{ and } \lambda(\infty) = 1\}, \\ \mathcal{F}_2 &= \{f_2(\cdot)^{1/2} \in \Lambda_c^{\gamma_3,\omega}(\mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{U}) : f_2(\cdot) \geq 0 \text{ and } \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{U}} f_2(\cdot, \dots, \cdot) dx dy dx du = 1\},\end{aligned}$$

where $\gamma_i > 1 \forall i = 1, 2$, and $\gamma_m > 1$. Recall that the parameter of the dynamic panel data model is $\theta = (b, \lambda)$. Suppose that \mathcal{B} is a compact set such that its interior containing the true parametric component of the dynamic panel data model b_0 . If the dynamic panel data model component $f_{Y_t|X_t, Y_{t-1}, U_t}$ is fully parameterized, then we do not need the infinite-dimensional function space \mathcal{M} . In the case, the parameter θ of the dynamic panel data model only contain a finite-dimensional parameter vector b . Without loss of generality, we assume that θ contains an unknown function λ . We assume that the parameters of interest $f_{X_{t+1}|X_t, U_t}$, θ , and $f_{X_t, Y_{t-1}, X_{t-1}, U_t}$ belong to the spaces, \mathcal{F}_1 , $\Theta \equiv \mathcal{B} \times \mathcal{M}$, and \mathcal{F}_2 respectively. The following smoothness and boundedness restrictions to limit the size of the parameter spaces.

Assumption B.1. *With $\gamma_i > 1 \forall i = 1, 2$, and $\gamma_m > 1$, we have (i) $f_1(\cdot|\cdot, \cdot) \in \mathcal{F}_1$, (ii) $\lambda(\cdot) \in \mathcal{M}$, (iii) $f_2(\cdot)^{1/2} \in \mathcal{F}_2$.*

Set $\mathcal{A} = \mathcal{F}_1 \times \Theta \times \mathcal{F}_2$ and $\alpha \equiv (f_1, \theta, f_2)'$. Then the true parameter α_0 maximizes:

$$\sup_{\alpha \in \mathcal{A}} E \left[\ln \int f_1(x_{t+1}|x_t, u_t) f_{Y_t|X_t, Y_{t-1}, U_t}(y_t|x_t, y_{t-1}, u_t; \theta) f_2(x_t, y_{t-1}, x_{t-1}, u_t) du_t \right].$$

An estimator could then be obtained by maximizing the sample analog of the above equation. Define

$$\begin{aligned}(18) \quad \hat{Q}_n(z_t; \alpha) &= \frac{1}{n} \sum_{i=1}^n \ln f_{Z_t}(z_{it}; \alpha) \text{ with} \\ \ln f_{Z_t}(z_{it}; \alpha) &\equiv \ln \int f_1(x_{it+1}|x_{it}, u_t) f_{Y_{it}|X_{it}, Y_{it-1}, U_t}(y_{it}|x_{it}, y_{it-1}, u_t; \theta) \\ &\quad \times f_2(x_{it}, y_{it-1}, x_{it-1}, u_t) du_t,\end{aligned}$$

where z_t is a realization of a random variable $Z_t \equiv (X_{t+1}, Y_t, X_t, Y_{t-1}, X_{t-1})$. However, when the function spaces \mathcal{A} is large, the estimation method could yield an inconsistent estimator or a consistent estimator which converges very slowly. Denote $\Theta^n \equiv \mathcal{B} \times \mathcal{M}^n$. The sieve spaces $\mathcal{A}_n \equiv \mathcal{F}_1^n \times \Theta^n \times \mathcal{F}_2^n$ will be introduced to replace the function spaces \mathcal{A} to overcome the problem, namely, maximizing $\widehat{Q}_n(z_t; \alpha)$ over \mathcal{A}_n , a sequence of approximation spaces to \mathcal{A} . In the sieve approximation, we consider a finite-dimensional sieve \mathcal{A}_n as follows. Let $p^k(\cdot) = (p_1(\cdot), \dots, p_k(\cdot))'$ be a vector of some known univariate basis function and $p^k(\cdot, \dots, \cdot) = (p_1(\cdot, \dots, \cdot), \dots, p_k(\cdot, \dots, \cdot))'$ be multivariate basis function generated by tensor product construction. The sieve spaces are

$$\begin{aligned}\mathcal{F}_1^n &= \{f_1(x_{t+1}|x_t, u_t; \delta_1) = p^{k_{n1}}(x_{t+1}, x_t, u_t)' \delta_1 \in \mathcal{F}_1\}, \\ \mathcal{M}^n &= \{\lambda(\varepsilon) = p^{k_{n\lambda}}(\varepsilon)' \beta_\lambda \in \mathcal{M}\}, \\ \mathcal{F}_2^n &= \{f_2(x_t, y_{t-1}, x_{t-1}, u_t; \delta_2)^{1/2} = p^{k_{n2}}(x_t, y_{t-1}, x_{t-1}, u_t)' \delta_2 \in \mathcal{F}_2\}.\end{aligned}$$

A consistent sieve MLE $\widehat{\alpha}_n$ is given by

$$\widehat{\alpha}_n = \arg \max_{\alpha \in \mathcal{A}_n} \widehat{Q}_n(z_t; \alpha).$$

The rest of this appendix show the consistency of $\widehat{\alpha}_n$ and its convergence rate under different metrics and the \sqrt{n} asymptotic normality of the parametric component b .

B.1. Consistency and Convergence Rates

In this section, we first introduce a strong norm $\|\cdot\|_s$ in Newey and Powell (2003) which would be used to show the consistency of the sieve estimator and then the Fisher norm, $\|\cdot\|$, in which the sieve estimator is consistent with a rate faster than $n^{-1/4}$.

For $\alpha \equiv (f_1, \theta, f_2)^T$,

$$(19) \quad \|\alpha\|_s = \|b\|_E + \|\lambda\|_{s,\omega} + \sum_{i=1}^2 \|f_i\|_{s,\omega}$$

where $\|b\|_E$ is the Euclidean norm and $\|f_i\|_{s,\omega} \equiv \sup_{\xi} |f_i(\xi)\omega(\xi)|$ with $\omega(\xi) = (1 + \|\xi\|_E^2)^{-\varsigma/2}$, $\varsigma > 0$. Since the supports of the unobserved variables v and ε could be unbounded, the

weighting function w is introduced to deal with unbounded support and has been used in Chen, Hansen, and Scheinkman (1997), Chen, Hong, and Tamer (2005) and Hu and Schennach (2008a). We make the following assumptions:

Assumption B.2. (i) The data $\{(Z_{it})_{i=1}^n\}$ are i.i.d.; (ii) The density function of Z_t , f_{Z_t} , satisfies $\int \omega(\xi)^{-2} f_{Z_t}(\xi) d\xi < \infty$.

Assumption B.3. (i) $b_0 \in \mathcal{B}$, a compact subset of \mathbb{R}^b ; (ii) Assumption B.1 holds under the norm $\|\alpha\|_s$.

Assumption B.4. (i) For any $\alpha \in \mathcal{A}$, there exists $\Pi_n \alpha \in \mathcal{A}_n$ such that $\|\Pi_n \alpha - \alpha\|_s = o(1)$; (ii) $k_{ni} \rightarrow +\infty$ and $k_{ni}/n \rightarrow 0$ for $i = 1, \lambda, 2$.

Definition B.1. $\ln f_{Z_t}(z_t; \alpha)$ is Hölder continuous with respect to $\alpha \in \mathcal{A}$ if there exists a measurable function $c_h(Z_t)$ with $E\{c_h(Z_t)^2\} < \infty$ such that, for all $\alpha_1, \alpha_2 \in \mathcal{A}$, and Z_t , we have

$$(20) \quad |\ln f_{Z_t}(z_t; \alpha_1) - \ln f_{Z_t}(z_t; \alpha_2)| \leq c_h(Z_t) \|\alpha_1 - \alpha_2\|_s.$$

The next assumption ensures $\ln f_{Z_t}(z_t; \alpha)$ is Hölder continuous with respect to $\alpha \in \mathcal{A}$.

Assumption B.5. (i) $E\{|\ln f_{Z_t}(z_t; \alpha)|^2\}$ is bounded; (ii) There exists a measurable function $\tilde{h}(Z_t)$ with $E\{\tilde{h}(Z_t)^2\} < \infty$ such that, for any $\bar{\alpha}_{12} = (\bar{f}_1, \bar{\theta}, \bar{f}_2, \bar{f}_3)$ and $\bar{\omega}(z_t, \varepsilon) = [1, \omega^{-1}(x_{t+1}, x_t, u_t), \omega^{-1}(\varepsilon), \omega^{-1}(x_t, y_{t-1}, x_{t-1}, u_t)]^T$, we have $|h_1(z_t, \bar{\alpha}_{12}, \bar{\omega})| < \tilde{h}(Z_t)$. (The function $h_1(z_t, \bar{\alpha}_{12}, \bar{\omega})$ is related to the derivatives of $f_{Z_t}(z_t; \alpha)$ which can be constructed by the similar derivation in Hu and Schennach (2008a).).

Applying Theorem 4.1 in Newey and Powell (2003) or Theorem 3.1 of Chen (2007) to the sieve estimator $\hat{\alpha}_n$ with these assumptions, we obtain the following lemma.

Lemma B.1. Let $\hat{\alpha}_n$ be the sieve MLE for α_0 identified in Section 2 and Assumptions B.1-B.5 holds, then $\|\hat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

Proof: The proof is similar to that of Lemma 2 in Hu and Schennach (2008a), and therefore, is omitted.

Lemma B.1 provides a consistency result under the metric $\|\cdot\|_s$ but the convergence rate under the metric is not fast enough to establish our semi-parametric asymptotic normality

and \sqrt{n} consistency result. In order to achieve this we consider the Fisher norm $\|\cdot\|$ in which $\hat{\alpha}_n$ converges at a rate faster than $n^{-1/4}$. In addition, Lemma B.1 allows us to restrict the sieve estimator $\hat{\alpha}_n$ to local $\|\cdot\|_s$ -neighborhood around the true parameter α_0 . For simplicity, we can assume the function space \mathcal{A} is convex. For any $v \in \bar{V}$, define the pathwise derivative as:

$$\frac{d \ln f_{Z_t}(z_t; \alpha)}{d\alpha}[v] \equiv \left. \frac{d \ln f_{Z_t}(z_t; \alpha + \tau v)}{d\tau} \right|_{\tau=0} \quad \text{a.s. } Z_t.$$

In particular, the pathwise derivative at the direction $[\alpha_1 - \alpha_2]$ evaluated at α_0 is:

$$\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \equiv \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha}[\alpha_2 - \alpha_0] \quad \text{a.s. } Z_t.$$

Expanding the pathwise derivative of $\ln f_{Z_t}(z_t; \alpha_0)$ gives:

$$\begin{aligned} & \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha}[\alpha - \alpha_0] \\ &= \frac{1}{f_{Z_t}(z_t; \alpha_0)} \left(\int (f_1 - f_{X_{t+1}|X_t, U_t}) f_{Y_t|X_t, Y_{t-1}, U_t} f_{X_t, Y_{t-1}, X_{t-1}, U_t} du_t \right. \\ & \quad + \int f_{X_{t+1}|X_t, U_t} \frac{d}{d\theta} f_{Y_{it}|X_{it}, Y_{it-1}, U_t} (\theta - \theta_0) f_{X_t, Y_{t-1}, X_{t-1}, U_t} du_t \\ & \quad \left. + \int f_{X_{t+1}|X_t, U_t} f_{Y_{it}|X_{it}, Y_{it-1}, U_t} (f_2 - f_{X_t, Y_{t-1}, X_{t-1}, U_t}) du_t \right). \end{aligned}$$

Following the notation, for any $\alpha_1, \alpha_2 \in \mathcal{A}$ we define the Fisher norm:

$$\|\alpha_1 - \alpha_2\|^2 \equiv \mathbb{E} \left\{ \left(\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \right)^2 \right\}.$$

We make the following assumptions to obtain a rate faster than $n^{-1/4}$.

Assumption B.6. Let k_n be the total number of sieve coefficients in the sieve estimator $\hat{\alpha}_n$, i.e., $k_n = k_{n1} + d_b + k_{n\lambda} + k_{n2}$. $(k_n n^{-1/2} \ln n) \times \sup_{\xi \in (\mathcal{X} \times \mathcal{X} \times \mathcal{U} \cup \mathbb{R} \cup \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{U})} \|p^{k_n}(\xi)\|_E^2 = o(1)$.

Assumption B.7. (i) There exist a measurable function $c(Z_t)$ with $E\{c(Z_t)^4\} < \infty$ such that $|\ln f_{Z_t}(z_t; \alpha)| \leq c(Z_t)$ for all Z_t and $\alpha \in \mathcal{A}_n$; (ii) $\ln f_{Z_t}(z_t; \alpha) \in \Lambda_c^{\tau, \omega}(\mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{X})$ with $\tau > d_z/2$, for all $\alpha \in \mathcal{A}_n$, where d_z is the dimension of Z_t .

Assumption B.8. \mathcal{A} is convex in α_0 , and $f_{Y_t|X_t, Y_{t-1}, U_t}(Y_t|X_t, Y_{t-1}, u_t; \theta)$ is pathwise differentiable at θ_0 .

Assumption B.9. $\ln N(\delta, \mathcal{A}_n) = O(k_n \ln(k_n/\delta))$ where $N(\delta, \mathcal{A}_n)$ is the minimum number of balls with radius δ under the $\|\cdot\|_s$ norm covering \mathcal{A}_n .

Assumption B.10. There exists $c_1, c_2 > 0$,

$$c_1 E \left(\ln \frac{f_{Z_t}(z_t; \alpha_0)}{f_{Z_t}(z_t; \alpha)} \right) \leq \|\alpha - \alpha_0\|^2 \leq c_2 E \left(\ln \frac{f_{Z_t}(z_t; \alpha_0)}{f_{Z_t}(z_t; \alpha)} \right)$$

holds for all $\alpha \in \mathcal{A}_n$ with $\|\alpha - \alpha_0\|_s = o(1)$.

Assumption B.11. For any $\alpha \in \mathcal{A}$, there exists $\Pi_n \alpha \in \mathcal{A}_n$ such that $\|\Pi_n \alpha - \alpha\| = o(k_n^{-\mu_1})$ and $k_n^{-\mu_1} = o(n^{-1/4})$.

The following lemma is a direct application of Theorem 3.1 of Ai and Chen (2003) and a similar proof can also be found in that of Theorem 2 in Hu and Schennach (2008b); we omit its proof.

Theorem B.1. Suppose that α_0 is identified and Assumptions B.6-B.11 hold, then $\|\hat{\alpha}_n - \alpha_0\| = o_p(n^{-1/4})$.

B.2. Asymptotic Normality

In this section, we follow the semi-parametric MLE framework of Hu and Schennach (2008b) to show the asymptotic normality of the parametric component b which represents the parameter of interest in dynamic panel data models. Let V be the space spanned by $\mathcal{A} - \alpha_0$ and \bar{V} be completion of V under the Fisher norm $\|\cdot\|$. It follows that $(\bar{V}, \|\cdot\|)$ is a Hilbert space with the inner product

$$\langle v_1, v_2 \rangle \equiv E \left\{ \left(\frac{d}{d\alpha} \ln f_{Z_t}(z_t; \alpha_0)[v_1] \right) \left(\frac{d}{d\alpha} \ln f_{Z_t}(z_t; \alpha_0)[v_2] \right) \right\},$$

and $\langle v, v \rangle = \|v\|^2$. For any fixed and nonzero $\kappa \in \mathbb{R}^{d_b}$, $f_\kappa(\alpha - \alpha_0) \equiv \kappa^T (b - b_0)$ is linear in $\alpha - \alpha_0$ and $f_\kappa(\alpha - \alpha_0)$ is a linear functional on $(\bar{V}, \|\cdot\|)$. Shen (1997) and Van Der Vaart (1991) show that $f(\alpha) \equiv \kappa^T b$ is a bounded linear functional on \bar{V} under the operator norm. That is:

$$(21) \quad \|f_\kappa\| \equiv \sup_{\{\alpha \in \mathcal{A}: \|\alpha - \alpha_0\| > 0\}} \frac{|f_\kappa(\alpha - \alpha_0)|}{\|\alpha - \alpha_0\|} < \infty.$$

By the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that for any $\alpha \in \mathcal{A}$, I have $f_\kappa(\alpha - \alpha_0) = \langle \alpha - \alpha_0, v^* \rangle$. and $\|f_\kappa\| = \|v^*\|$.²⁶ Denote $\bar{V} = \mathbb{R}^{d_b} \times \bar{W}$ and $\bar{W} \equiv \overline{\mathcal{F}_1^n \times \mathcal{M}^n \times \mathcal{F}_2^n} - (f_{X_{t+1}|X_t, U_t}, \lambda, f_{X_t, Y_{t-1}, X_{t-1}, U_t})^T$.²⁷ We can expand the first pathwise derivative out as follows:

$$\begin{aligned} \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha} [\alpha - \alpha_0] &= \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df_1} [f_1 - f_{X_{t+1}|X_t, U_t}] + \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{db} [b - b_0] \\ &+ \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\lambda} [\lambda - \lambda_0] + \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df_2} [f_2 - f_{X_t, Y_{t-1}, X_{t-1}, U_t}]. \end{aligned}$$

For each component b_j of b , $j = 1, 2, \dots, d_b$, we define $w_j^* \in \bar{W}$ to be the solution to the following minimization problem associated with the denominator of the operator norm,

$$\begin{aligned} w_j^* = \arg \min_{w_j = (f_1, \lambda, f_2, f_3)' \in \bar{W}} E \left[\left(\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{db_j} - \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df_1} [f_1] \right. \right. \\ \left. \left. - \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\lambda} [\lambda] - \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df_2} [f_2] \right) \right]. \end{aligned}$$

Define $w^* = (w_1^*, \dots, w_{d_b}^*)$,

$$\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df} [w^*] = \left(\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{dh} [w_1], \dots, \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df} [w_{d_b}^*] \right),$$

and

$$(22) \quad D_{w^*}(z_t) \equiv \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d'_b} - \frac{d \ln f_{Z_t}(z_t; \alpha_0)}{df} [w^*].$$

With these notation,

$$\|f_\kappa\|^2 = \sup_{\{\alpha \in \mathcal{A}: \|\alpha - \alpha_0\| > 0\}} \frac{|f_\kappa(\alpha - \alpha_0)|^2}{\|\alpha - \alpha_0\|^2} = \kappa^T (E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\})^{-1} \kappa,$$

$v^* \equiv (v_b^*, v_h^*) \in \bar{V}$ with $v_b^* = (E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\})^{-1} \kappa$ and $v_h^* = -w^* \times v_b^*$. In addition, $f_\kappa(\alpha - \alpha_0) = \kappa^T (b - b_0) = \langle \alpha - \alpha_0, v^* \rangle$ and $\frac{d \ln f_{Z_t}(z_t; \alpha_0)}{d\alpha} [v^*] = D_{w^*}(z_t) v_b^*$. See Chen (2007) for detailed discussion about this linear functional approach. Therefore, the asymptotic distribution of parametric component \hat{b}_n reduces to when the linear functional f_κ is bounded and

²⁶Stein (1956) pointed out that v^* yields the most difficult one-dimensional sub-problem. Begun, Hall, Huang, and Wellner (1983) mentioned that v^* represents a worst possible direction to nonparametric component for estimating parametric component.

²⁷ \bar{W} is a function space of nonparametric components.

what is the asymptotic distribution of $\langle \widehat{\alpha}_n - \alpha_0, v^* \rangle$. That is:

$$\begin{aligned} \kappa^T (\widehat{b}_n - b_0) &= \langle \widehat{\alpha}_n - \alpha_0, v^* \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \frac{d \ln f_{Z_t}(z_{it}; \alpha_0)}{d\alpha} [v^*] + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \kappa^T (E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\})^{-1} D_{w^*}(z_{it})^T + o_p(n^{-1/2}), \end{aligned}$$

and $\sqrt{n}(\widehat{b}_n - b_0) \rightarrow N(0, (E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\})^{-1})$.

We make the following sufficient conditions for the \sqrt{n} -normality of \widehat{b}_n which are also conditions in Ai and Chen (2003) and Hu and Schennach (2008b):

Assumption B.12. (i) $E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\}$ is positive-definite and bounded; (ii) $b_0 \in \text{int}(\mathcal{B})$.

Assumption B.13. There is a $v_n^* = (v_b^*, -\Pi_n w^* \times v_b^*) \in \mathcal{A}_n - \alpha_0$ such that $\|v_n^* - v^*\| = o_p(n^{-1/4})$.

We use the \sqrt{n} consistency results in the previous section to focus on a smaller neighbor of α_0 , Define $\mathcal{N}_{on} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s = o(1), \|\alpha - \alpha_0\| = o(n^{-1/4})\}$ and $\mathcal{N}_o \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_s = o(1), \|\alpha - \alpha_0\| = o(n^{-1/4})\}$.

Assumption B.14. There exists a measurable function $\widehat{h}(Z_t)$ with $E\{\widehat{h}(Z_t)^2\} < \infty$ such that, for any $\bar{\alpha} = (\bar{f}_1, \bar{\theta}, \bar{f}_2)$, we have

$$(23) \quad \left| h_2(z_t, \bar{\alpha}, \bar{\omega}) \right| + \left| h_1(z_t, \bar{\alpha}, \bar{\omega}) \right|^2 < \widehat{h}(Z_t).$$

(The similar definition of $h_2(z_t, \bar{\alpha}, \bar{\omega})$ can be found in Hu and Schennach (2008a).)

For $\tilde{f} = f_1, \lambda, f_2$, denote

$$\begin{aligned} \frac{df_{Z_t}(z_t; \alpha_0)}{d\tilde{f}}[p^{k_{ni}}] &= \left(\frac{df_{Z_t}(z_t; \alpha_0)}{d\tilde{f}}[p_1^{k_{ni}}], \dots, \frac{df_{Z_t}(z_t; \alpha_0)}{d\tilde{f}}[p_{k_{ni}}^{k_{ni}}] \right)^T \quad \forall i = 1, \lambda, 2, \\ \frac{df_{Z_t}(z_t; \alpha_0)}{db} &= \left(\frac{df_{Z_t}(z_t; \alpha_0)}{db_1}, \dots, \frac{df_{Z_t}(z_t; \alpha_0)}{db_{d_b}} \right)^T, \\ \frac{df_{Z_t}(z_t; \alpha_0)}{d\alpha}[p^{k_n}] &= \left(\left(\frac{df_{Z_t}(z_t; \alpha_0)}{db} \right)^T, \left(\frac{df_{Z_t}(z_t; \alpha_0)}{df_1}[p^{k_{n1}}] \right)^T, \left(\frac{df_{Z_t}(z_t; \alpha_0)}{d\lambda}[p^{k_{n\lambda}}] \right)^T, \right. \\ &\quad \left. \left(\frac{df_{Z_t}(z_t; \alpha_0)}{df_2}[p^{k_{n2}}] \right)^T \right)^T, \end{aligned}$$

and

$$\Omega_{k_n} = E \left\{ \left(\frac{df_{Z_t}(z_t; \alpha_0)}{d\alpha}[p^{k_n}] \right) \left(\frac{df_{Z_t}(z_t; \alpha_0)}{d\alpha}[p^{k_n}] \right)^T \right\}.$$

Assumption B.15. *The smallest eigenvalue of the matrix Ω_{k_n} is bounded away from zero, and $\|p_j^{k_{ni}}\|_{s,\omega} < \infty$ for $j = 1, 2, \dots, k_{ni}$ uniformly in k_{ni} .*

Assumption B.16. *For all $\alpha \in \mathcal{N}_{on}$, there exists a measurable function $h(Z_t)$ with $E|h(Z_t)| < \infty$ such that*

$$(24) \quad \left| \frac{d^4 \ln f_{Z_t}(z_t; \bar{\alpha} + t(\alpha - \alpha_0))}{dt^4} \right|_{t=0} \leq h(Z_t) \|\alpha - \alpha_0\|_s^4.$$

Theorem B.2. *Suppose that α_0 is identified and Assumptions B.6-B.11 and B.12-B.16 hold, then $\sqrt{n}(\hat{b}_n - b_0) \Rightarrow N(0, V^{-1})$ where $V = E\{D_{w^*}(Z_t)^T D_{w^*}(Z_t)\}$.*

Proof: The likelihood function $f_{Z_t}(z_t; \alpha)$ has a similar expression as the likelihood function in Hu and Schennach (2008a). The proof there can directly apply to our case, and therefore, is omitted.

C. Restrictions on the Sieve Coefficients

This appendix describes the sieve MLE method used to estimate nonlinear dynamic panel data models. We provide detailed derivation of the method based on the likelihood function in Eq. (3). According to Eq. (3), there are several essential parts in the likelihood function, $f_{X_{t+1}|X_t, U_t}$, $f_{Y_t|X_t, Y_{t-1}, U_t}$, and $f_{X_t, Y_{t-1}, X_{t-1}, U_t}$. While the specifications of $f_{Y_t|X_t, Y_{t-1}, U_t}$ have

been provided in Section 3, $f_{X_{t+1}|X_t, U_t}$, and $f_{X_t, Y_{t-1}, X_{t-1}, U_t}$ will be treated here. We will show sieve approximations and their constraints of those nonparametric components in the two examples. First, we introduce the sieve estimators for the covariate evolution, $f_{X_{t+1}|X_t, U_t}$, since we can use the same sieve approximates for them in the examples. Suppose that $x_t, u_t \in [0, l_1]$ and $(x_{t+1} - u_t) \in [-l_2, l_2]$.²⁸ The sieve estimators for the covariate evolution are constructed by Fourier series as follows:

$$f_1(x_{t+1}|x_t, u_t; \delta_1) = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \delta_{1,ijk} p_{1i}(x_{t+1} - u_t) p_{2j}(x_t) p_{3k}(u_t),$$

where

$$p_{1i}(x_{t+1} - u_t) = \cos \frac{i\pi}{l_2}(x_{t+1} - u_t) \text{ or } \sin \frac{i\pi}{l_2}(x_{t+1} - u_t),$$

$$p_{2j}(x_t) = \cos \frac{j\pi}{l_1}x_t, \text{ and, } p_{3k}(u_t) = \cos \frac{k\pi}{l_1}u_t.$$

The conditional density restrictions $\int f_1(x_{t+1}|x_t, u_t) dx_{t+1} = 1 \forall x_t, u_t$ implies that a constant term in the sieve expression $f_1(x_{t+1}|x_t, u_t)$ equals $\frac{1}{2l_2}$.

Next, since $f_{X_{t+1}|X_t, U_t}$ is identified through Assumption 2.5, one thing remained to show is how to implement the normalization assumption in estimation. Consider the zero mode case, we have $\frac{\partial}{\partial x_{t+1}} f_1(x_{t+1}|x_t, u_t)|_{x_{t+1}=u_t} = 0$ for all x_t, u_t . By properties of the trigonometric functions, sieve coefficients related to terms like $\sin \frac{i\pi}{l_2}(x_{t+1} - u_t)$ survive. The identification restrictions impose constraints on those coefficients.

²⁸While the range of x_t can be obtained from data set, the domain of u_t depends on the modeling of unobserved heterogeneity. In our simulation design, the domain of u_t is $(-\infty, \infty)$.

We consider the following simple case:

$$\begin{aligned}
& f_1(x_{t+1}|x_t, u_t; \delta_1) \\
&= \left(c_{00} + c_{01} \cos \frac{\pi}{l_1} x_t + c_{02} \cos \frac{2\pi}{l_1} x_t \right) \left(a_{00} + a_{01} \cos \frac{\pi}{l_1} u_t + a_{02} \cos \frac{2\pi}{l_1} u_t \right) \\
&\quad + \sum_{i=1}^5 \left(c_{i0} + c_{i1} \cos \frac{\pi}{l_1} x_t + c_{i2} \cos \frac{2\pi}{l_1} x_t \right) \left(a_{00} + a_{01} \cos \frac{\pi}{l_1} u_t + a_{02} \cos \frac{2\pi}{l_1} u_t \right) \\
&\quad \quad \times \cos \frac{i\pi}{l_2} (x_{t+1} - u_t) \\
&\quad + \sum_{i=1}^5 \left(d_{i0} + d_{i1} \cos \frac{\pi}{l_1} x_t + d_{i2} \cos \frac{2\pi}{l_1} x_t \right) \left(a_{00} + a_{01} \cos \frac{\pi}{l_1} u_t + a_{02} \cos \frac{2\pi}{l_1} u_t \right) \\
&\quad \quad \times \sin \frac{i\pi}{l_2} (x_{t+1} - u_t).
\end{aligned}$$

Then the density restriction gives $c_{00}a_{00} = \frac{1}{2l_2}$ and the identification restriction on the coefficients are

$$\sum_{i=1}^5 id_{i0} = \sum_{i=1}^5 id_{i1} = \sum_{i=1}^5 id_{i2} = 0.$$

As for nonparametric series estimator of $f_{X_t, Y_{t-1}, X_{t-1}, U_t; \delta_2}$, we have to separate it into two cases to fit into our examples. First, we handle with dynamic discrete choice models and a sieve estimator of $f_{X_t, Y_{t-1}, X_{t-1}, U_t; \delta_2}$ is given by the following:

$$f_{X_t, Y_{t-1}, X_{t-1}, U_t; \delta_2} = (f_{X_t, X_{t-1}, U_t | Y_{t-1}=0} f_{Y_{t-1}=0})^{1-Y_{t-1}} (f_{X_t, X_{t-1}, U_t | Y_{t-1}=1} (1 - f_{Y_{t-1}=0}))^{Y_{t-1}},$$

where

$$(25) \quad (f_{X_t, X_{t-1}, U_t | Y_{t-1}=0})^{1/2} = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \hat{a}_{ijk} q_i(x_t - x_{t-1} - u_t) q_j(x_{t-1}) q_k(u_t),$$

and

$$(26) \quad (f_{X_t, X_{t-1}, U_t | Y_{t-1}=1})^{1/2} = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \tilde{a}_{ijk} q_i(x_t - x_{t-1} - u_t) q_j(x_{t-1}) q_k(u_t).$$

Our choice of q'_i 's and q'_j 's are the orthonormal Fourier series:

$$\begin{aligned} q_0(u_t) &= \frac{1}{\sqrt{l_1}}, q_k(u_t) = \sqrt{\frac{2}{l_1}} \cos\left(\frac{k\pi}{l_1} u_t\right), q_0(x_{t-1}) = \frac{1}{\sqrt{l_1}} \text{ and } q_j(x_{t-1}) = \sqrt{\frac{2}{l_1}} \cos\left(\frac{j\pi}{l_1} x_{t-1}\right) \\ q_0(x_t - x_{t-1} - u_t) &= \frac{1}{\sqrt{l_2}} \text{ and } q_i(x_t - x_{t-1} - u_t) = \frac{1}{\sqrt{l_2}} \sin\left(\frac{i\pi}{l_2} (x_t - x_{t-1} - u_t)\right) \text{ or} \\ q_i(x_t - x_{t-1} - u_t) &= \frac{1}{\sqrt{l_2}} \cos\left(\frac{i\pi}{l_2} (x_t - x_{t-1} - u_t)\right). \end{aligned}$$

The density restrictions $\int f_{X_t, X_{t-1}, U_t | Y_{t-1}=0} dx_t dx_{t-1} du_t = 1$ and $\int f_{X_t, X_{t-1}, U_t | Y_{t-1}=1} dx_t dx_{t-1} du_t = 1$ amount to

$$\sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} (\hat{a}_{ijk})^2 = 1, \text{ and } \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} (\tilde{a}_{ijk})^2 = 1.$$

We consider the case where $i_n = 5$, $j_n = 2$, and $k_n = 2$:

$$\begin{aligned} & (f_{X_t, X_{t-1}, U_t | Y_{t-1}=0})^{1/2} \\ &= \left(\hat{c}_{00} + \hat{c}_{01} \cos \frac{\pi}{l_1} x_{t-1} + \hat{c}_{02} \cos \frac{2\pi}{l_1} x_{t-1} \right) \left(\hat{a}_{00} + \hat{a}_{01} \cos \frac{\pi}{l_1} u_t + \hat{a}_{02} \cos \frac{2\pi}{l_1} u_t \right) \\ &+ \sum_{i=1}^5 \left(\hat{c}_{i0} + \hat{c}_{i1} \cos \frac{\pi}{l_1} x_{t-1} + \hat{c}_{i2} \cos \frac{2\pi}{l_1} x_{t-1} \right) \left(\hat{a}_{i0} + \hat{a}_{i1} \cos \frac{\pi}{l_1} u_t + \hat{a}_{i2} \cos \frac{2\pi}{l_1} u_t \right) \\ &\quad \times \cos \frac{i\pi}{l_2} (x_t - x_{t-1} - u_t) \\ &+ \sum_{i=1}^5 \left(\hat{c}_{i0} + \hat{c}_{i1} \cos \frac{\pi}{l_1} x_{t-1} + \hat{c}_{i2} \cos \frac{2\pi}{l_1} x_{t-1} \right) \left(\hat{b}_{i0} + \hat{b}_{i1} \cos \frac{\pi}{l_1} u_t + \hat{b}_{i2} \cos \frac{2\pi}{l_1} u_t \right) \\ &\quad \times \sin \frac{i\pi}{l_2} (x_t - x_{t-1} - u_t). \end{aligned}$$

The sieve expression of $f_{X_{it}, Y_{it-1}, X_{it-1}, U_{it}; \delta_2}$ in dynamic censored models can be constructed similarly. Eq. (25) is still applicable for $Y_{t-1} = 0$ part,

$$(f_{X_t, X_{t-1}, U_t | Y_{t-1}=0})^{1/2} = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \hat{a}_{ijk} q_i(x_t - x_{t-1} - u_t) q_j(x_{t-1}) q_k(u_t).$$

Suppose that $y_{t-1} \in (0, l_3]$. Consider

$$(f_{X_t, Y_{t-1}>0, X_{t-1}, U_t})^{1/2} = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \sum_{l=0}^{l_n} \tilde{a}_{ijkl} \tilde{q}_i(x_t - x_{t-1} - u_t) \tilde{q}_j(x_{t-1}) \tilde{q}_k(u_t) \tilde{q}_l(y_{t-1}).$$

The density restriction $\int f_{X_t, X_{t-1}, U_t | Y_{t-1}=0} + \left(\int f_{X_t, Y_{t-1}>0, X_{t-1}, U_t} dy_{t-1} \right) dx_t dx_{t-1} du_t = 1$ is

$$(27) \quad \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} (\hat{a}_{ijk})^2 + \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \sum_{l=0}^{l_n} (\tilde{a}_{ijkl})^2 = 1.$$

In the simulation of the censored tobit model in Section 4, our choice of $Y_{t-1} > 0$ part is $i_n = 5$, $j_n = 2$, $k_n = 2$, and $l_n = 2$.

References

- AHN, S., AND P. SCHMIDT (1995): “Efficient Estimation of Models for Dynamic Panel Data,” *Journal of Econometrics*, 68(1), 5–28.
- AI, C., AND X. CHEN (2003): “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71(6), 1795–1843.
- ALTONJI, J., AND R. MATZKIN (2005): “Cross Section and Panel Data Estimators for Non-separable Models with Endogenous Regressors,” *Econometrica*, 73(4), 1053–1102.
- ANDERSON, T., AND C. HSIAO (1982): “Formulation and Estimation of Dynamic Models Using Panel Data,” *Journal of Econometrics*, 18(1), 47–82.
- ARELLANO, M., AND S. BOND (1991): “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *The Review of Economic Studies*, 58(2), 277–297.
- ARELLANO, M., AND S. BONHOMME (2009): “Robust Priors in Nonlinear Panel Data Models,” *Econometrica*, 77(2), 489–536.
- ARELLANO, M., AND O. BOVER (1995): “Another Look at the Instrumental Variable Estimation of Error Component Models,” *Journal of Econometrics*, 68(1), 29–51.
- BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER (1983): “Information and Asymptotic Efficiency in Parametric-nonparametric Models,” *The Annals of Statistics*, 11(2), 432–452.
- CARRO, J. (2007): “Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects,” *Journal of Econometrics*, 140(2), 503–528.

- CARROLL, R., X. CHEN, AND Y. HU (2010): “Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors,” *Journal of Nonparametric Statistics*, 22(4), 379–399.
- CHAMBERLAIN, G. (1980): “Analysis of Covariance with Qualitative Data,” *The Review of Economic Studies*, 47(1), 225–238.
- (1984): “Panel Data,” in *Handbook of Econometrics, Vol. 2, ed. by Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland, Amsterdam.*, pp. 1247–1318.
- CHAY, K., H. HOYNES, AND D. HYSLOP (2001): “A Non-experimental Analysis of True State Dependence in Monthly Welfare Participation Sequences,” University of California, Berkeley.
- CHEN, X. (2007): “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in *Handbook of Econometrics, Vol. 6, ed. by J. Heckman and E. Leamer. Amsterdam: North-Holland, Amsterdam.*, pp. 5549–5632.
- CHEN, X., L. HANSEN, AND J. SCHEINKMAN (1997): “Shape-preserving Estimation of Diffusions,” *Unpublished Manuscript, University of Chicago, Dept. of Economics.*
- CHEN, X., H. HONG, AND E. TAMER (2005): “Measurement Error Models with Auxiliary Data,” *Review of Economic Studies*, 72(2), 343–366.
- CHEN, X., AND X. SHEN (1998): “Sieve Extremum Estimates for Weakly Dependent Data,” *Econometrica*, 66(2), 289–314.
- CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2009): “Identification and Estimation of Marginal Effects in Nonlinear Panel Models,” *CeMMAP Working Paper.*
- CHINTAGUNTA, P., E. KYRIAZIDOU, AND J. PERKTOLD (2001): “Panel Data Analysis of Household Brand Choices,” *Journal of Econometrics*, 103(1-2), 111–153.
- CONTOYANNIS, P., A. JONES, AND N. RICE (2004): “Simulation-based Inference in Dynamic Panel Probit Models: An Application to Health,” *Empirical Economics*, 29(1), 49–77.
- EVDOKIMOV, K. (2009): “Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity,” *Working Paper.*

- FERNÁNDEZ-VAL, I. (2009): “Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models,” *Journal of Econometrics*, 150(1), 71–85.
- GOURIEROUX, C., AND A. MONFORT (1993): “Simulation-based Inference: A Survey with Special Reference to Panel Data Models,” *Journal of Econometrics*, 59(1-2), 5–33.
- HAHN, J., AND G. KUERSTEINER (2004): “Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects,” *Unpublished Manuscript*.
- HAIJIVASSILIOU, V. (1993): “Simulation Estimation Methods for Limited Dependent Variable Models,” *Handbook of Statistics*, 59.
- HAIJIVASSILIOU, V., AND P. RUUD (1994): “Classical Estimation Methods for LDV Models Using Simulation,” in *Handbook of Econometrics, Vol. 4, ed. by R. Engle and D. McFadden. Amsterdam: North-Holland, Amsterdam.*, pp. 2383–2441.
- HALLIDAY, T. (2002): “Heterogeneity, State Dependence and Health,” Princeton University.
- HECKMAN, J. (1978): “Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence,” pp. 227–269.
- (1981a): “Statistical Models for Discrete Panel Data,” in *Structural Analysis of Discrete Panel Data with Econometric Applications, ed. by C. Manski and D. McFadden. Cambridge: MIT Press*, pp. 114–178.
- (1981b): “The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process,” in *Structural Analysis of Discrete Panel Data with Econometric Applications, ed. by C. Manski and D. McFadden. Cambridge: MIT Press*, pp. 179–195.
- HECKMAN, J., AND R. WILLIS (1977): “A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women,” *The Journal of Political Economy*, 85(1), 27–58.
- HODERLEIN, S., AND H. WHITE (2009): “Nonparametric Identification in Nonseparable Panel Data Models With Generalized Fixed Effects,” *CeMMAP Working Paper*.

- HONORÉ, B. (1993): “Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables,” *Journal of Econometrics*, 59(1-2), 35–61.
- HONORÉ, B., AND L. HU (2004): “Estimation of Cross Sectional and Panel Data Censored Regression Models with Endogeneity,” *Journal of Econometrics*, 122(2), 293–316.
- HONORÉ, B., AND E. KYRIAZIDOU (2000): “Panel Data Discrete Choice Models with Lagged Dependent Variables,” *Econometrica*, 68(4), 839–874.
- HONORÉ, B., AND E. TAMER (2006): “Bounds on Parameters in Panel Dynamic Discrete Choice Models,” *Econometrica*, 74(3), 611–629.
- HU, L. (2000): “Estimating a Censored Dynamic Panel Data Model with an Application to Earnings Dynamics,” *Unpublished Manuscript, Northwestern University*.
- (2002): “Estimation of a Censored Dynamic Panel Data Model,” *Econometrica*, 70(6), 2499–2517.
- HU, Y., AND S. SCHENNACH (2008a): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76(1), 195–216.
- (2008b): “Supplement to Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76(1), 195–216.
- HU, Y., AND J. SHIU (2011): “Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness,” *Jonhs Hopkins University, Department of Economics Working Paper*, 581.
- HYSLOP, D. (1995): “State Dependence, Serial Correlation and Heterogeneity in Intertemporal Participation Behavior: Monte Carlo Evidence and Empirical Results for Married Women,” *Industrial Relations Section Working Paper #347, Princeton University*.
- (1999): “State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women,” *Econometrica*, 67(6), 1255–1294.
- KEANE, M. (1993): “Simulation Estimation for Panel Data Models with Limited Dependent Variables,” *Handbook of Statistics*, 11, 545–571.

- NEWKEY, W., AND J. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71(5), 1565–1578.
- NEYMAN, J., AND E. SCOTT (1948): “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16, 1–32.
- RASCH, G. (1960): “Probabilistic Models for Some Intelligence and Attainment Tests,” (*Copenhagen: Danmarks Paedagogiske Institut*).
- SHEN, X. (1997): “On Methods of Sieves and Penalization,” *The Annals of Statistics*, 25, 2555–2591.
- STEIN, C. (1956): “Efficient Nonparametric Testing and Estimation,” in *Proc. Third Berkeley Symp. Math. Statist. Probab*, vol. 1, pp. 187–195.
- VAN DER VAART, A. (1991): “On Differentiable Functionals,” *The Annals of Statistics*, 19, 178–204.
- WOOLDRIDGE, J. (2005): “Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity,” *Journal of Applied Econometrics*, 20(1), 39–54.

Table 1: Monte Carlo Simulation of Semi-parametric Probit model (n=250)

		Parameters		
DGP		β_0	β_1	γ
DGP I:	true value	0	-1	0
	mean estimate	-0.009	-1.000	-0.004
	median estimate	0.003	-1.000	-0.009
	standard error	0.103	0.099	0.100
DGP II:	true value	0	-1	0
	mean estimate	0.006	-1.015	-0.003
	median estimate	0.011	-1.014	-0.003
	standard error	0.113	0.109	0.087
DGP III:	true value	0	-1	1
	mean estimate	-0.013	-1.006	1.011
	median estimate	-0.008	-1.020	1.006
	standard error	0.095	0.105	0.106
DGP IV:	true value	0	-1	1
	mean estimate	0.008	-0.997	1.006
	median estimate	0.009	-0.991	1.014
	standard error	0.106	0.096	0.095
DGP V:	true value	0	-1	1
	mean estimate	-0.013	-1.006	1.011
	median estimate	-0.008	-1.021	1.006
	standard error	0.095	0.104	0.106

Note: The simulated date has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric probit model. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 2: Monte Carlo Simulation of Semi-parametric Probit model (n=500)

		Parameters		
DGP		β_0	β_1	γ
DGP I:	true value	0	-1	0
	mean estimate	0.008	-0.994	-0.013
	median estimate	0.010	-1.006	-0.002
	standard error	0.086	0.103	0.108
DGP II:	true value	0	-1	0
	mean estimate	-0.003	-1.010	0.007
	median estimate	-0.012	-1.004	0.011
	standard error	0.087	0.095	0.110
DGP III:	true value	0	-1	1
	mean estimate	0.008	-0.991	0.997
	median estimate	0.016	-0.994	1.000
	standard error	0.093	0.105	0.106
DGP IV:	true value	0	-1	1
	mean estimate	-0.005	-1.005	1.008
	median estimate	0.003	-1.024	1.010
	standard error	0.092	0.104	0.121
DGP V:	true value	0	-1	1
	mean estimate	-0.001	-0.996	0.996
	median estimate	0.012	-1.002	0.982
	standard error	0.112	0.095	0.093

Note: The simulated date has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric probit model. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 3: Monte Carlo Simulation of Semi-parametric Probit model (n=1000)

		Parameters		
DGP		β_0	β_1	γ
DGP I:	true value	0	-1	0
	mean estimate	-0.009	-1.002	-0.014
	median estimate	-0.014	-1.004	-0.002
	standard error	0.100	0.097	0.092
DGP II:	true value	0	-1	0
	mean estimate	0.007	-1.003	-0.001
	median estimate	0.015	-1.013	0.008
	standard error	0.100	0.107	0.098
DGP III:	true value	0	-1	1
	mean estimate	0.012	-1.002	1.000
	median estimate	0.020	-1.013	1.009
	standard error	0.100	0.106	0.098
DGP IV:	true value	0	-1	1
	mean estimate	-0.006	-0.996	0.995
	median estimate	-0.011	-0.995	1.004
	standard error	0.087	0.112	0.096
DGP V:	true value	0	-1	1
	mean estimate	0.004	-0.990	1.013
	median estimate	-0.004	-0.988	1.016
	standard error	0.101	0.095	0.094

Note: The simulated date has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric probit model. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 4: Simulation of Average Partial Effects in Probit model (n=500)

	State Dependence	Average Partial Effects
DGP I:	$Y_{t-1} = 0$	0.413 (0.056)
	$Y_{t-1} = 1$	0.409 (0.056)
DGP II:	$Y_{t-1} = 0$	0.411 (0.061)
	$Y_{t-1} = 1$	0.413 (0.061)
DGP III:	$Y_{t-1} = 0$	0.415 (0.057)
	$Y_{t-1} = 1$	0.753 (0.083)
DGP IV:	$Y_{t-1} = 0$	0.411 (0.055)
	$Y_{t-1} = 1$	0.740 (0.082)
DGP V:	$Y_{t-1} = 0$	0.422 (0.060)
	$Y_{t-1} = 1$	0.726 (0.083)

Note: The average partial effects are reported at the mean value of the explanatory variable and two different outcomes of Y_{t-1} , 0 and 1. Standard errors of these average partial effects are computed by the standard deviation of the estimates across 100 simulations.

Table 5: Monte Carlo Simulation of Semi-parametric Tobit model (n=250)

		Parameters			
DGP		β_0	β_1	γ	σ_ξ^2
DGP I:	true value	0	-1	0	0.5
	mean estimate	-0.007	-1.001	-0.005	0.523
	median estimate	0.010	-1.008	-0.010	0.522
	standard error	0.104	0.098	0.100	0.028
DGP II:	true value	0	-1	0	0.5
	mean estimate	-0.014	-1.037	0.010	0.524
	median estimate	-0.008	-1.030	0.006	0.525
	standard error	0.095	0.097	0.106	0.028
DGP III:	true value	0	-1	1	0.5
	mean estimate	-0.015	-1.019	1.003	0.523
	median estimate	-0.015	-1.029	0.998	0.525
	standard error	0.094	0.114	0.130	0.039
DGP IV:	true value	0	-1	1	0.5
	mean estimate	-0.012	-1.035	0.971	0.523
	median estimate	0.003	-1.038	0.967	0.523
	standard error	0.104	0.112	0.107	0.046
DGP V:	true value	0	-1	1	0.5
	mean estimate	-0.016	-1.050	1.000	0.523
	median estimate	-0.005	-1.053	0.991	0.525
	standard error	0.092	0.125	0.140	0.040

Note: The simulated data has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric Tobit models. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 6: Monte Carlo Simulation of Semi-parametric Tobit model (n=500)

		Parameters			
	DGP	β_0	β_1	γ	σ_ξ^2
DGP I:	true value	0	-1	0	0.5
	mean estimate	0.007	-1.006	0.002	0.525
	median estimate	0.006	-0.992	0.009	0.523
	standard error	0.092	0.111	0.103	0.031
DGP II:	true value	0	-1	0	0.5
	mean estimate	0.001	-1.009	0.017	0.526
	median estimate	-0.014	-1.009	0.019	0.524
	standard error	0.112	0.096	0.098	0.030
DGP III:	true value	0	-1	1	0.5
	mean estimate	0.015	-1.011	0.989	0.528
	median estimate	0.014	-1.003	0.994	0.526
	standard error	0.100	0.112	0.114	0.035
DGP IV:	true value	0	-1	1	0.5
	mean estimate	0.007	-1.015	0.988	0.501
	median estimate	0.017	-1.023	0.986	0.523
	standard error	0.093	0.103	0.101	0.036
DGP V:	true value	0	-1	1	0.5
	mean estimate	-0.002	-1.030	0.997	0.528
	median estimate	0.008	-1.026	0.996	0.527
	standard error	0.108	0.099	0.120	0.035

Note: The simulated data has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric Tobit models. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 7: Monte Carlo Simulation of Semi-parametric Tobit model (n=1000)

		Parameters			
	DGP	β_0	β_1	γ	σ_ξ^2
DGP I:	true value	0	-1	0	0.5
	mean estimate	-0.006	-0.991	0.016	0.524
	median estimate	-0.001	-0.991	-0.005	0.526
	standard error	0.106	0.106	0.098	0.031
DGP II:	true value	0	-1	0	0.5
	mean estimate	0.008	-1.019	-0.010	0.524
	median estimate	-0.005	-1.020	-0.001	0.526
	standard error	0.084	0.102	0.089	0.032
DGP III:	true value	0	-1	1	0.5
	mean estimate	0.002	-1.020	0.986	0.525
	median estimate	-0.008	-1.020	0.985	0.527
	standard error	0.087	0.106	0.098	0.036
DGP IV:	true value	0	-1	1	0.5
	mean estimate	-0.005	-1.004	0.990	0.525
	median estimate	-0.010	-1.011	1.002	0.526
	standard error	0.086	0.109	0.094	0.033
DGP V:	true value	0	-1	1	0.5
	mean estimate	-0.007	-1.010	0.991	0.525
	median estimate	-0.010	-1.004	1.003	0.522
	standard error	0.087	0.111	0.104	0.036

Note: The simulated data has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric Tobit models. Standard errors of the parameters are computed by the standard deviation of the estimates across 100 simulations since standard errors from the limiting distribution of the estimators require much computational burden.

Table 8: Simulation of Average Partial Effects in Tobit model (n=500)

	State Dependence	Average Partial Effects
DGP I:	$Y_{t-1} = 0$	0.174 (0.053)
	$Y_{t-1} = \bar{y}$	0.174 (0.053)
DGP II:	$Y_{t-1} = 0$	0.206 (0.072)
	$Y_{t-1} = \bar{y}$	0.207 (0.072)
DGP III:	$Y_{t-1} = 0$	0.183 (0.059)
	$Y_{t-1} = \bar{y}$	0.461 (0.522)
DGP IV:	$Y_{t-1} = 0$	0.194 (0.067)
	$Y_{t-1} = \bar{y}$	0.431 (0.357)
DGP V:	$Y_{t-1} = 0$	0.246 (0.096)
	$Y_{t-1} = \bar{y}$	0.442 (0.407)

Note: The average partial effects are reported at the mean value of the explanatory variable and two different outcomes of Y_{t-1} , 0 and \bar{y} . Standard errors of these average partial effects are computed by the standard deviation of the estimates across 100 simulations.

Table 9: Sample Characteristics

Variables	Full Sample (1)	Employed 7 years (2)	Employed 0 years (3)	Single Transition from Work (4)	Single Transition to Work (5)	Multiple Transitions (6)
Age	35.38 (0.22)	35.22 (0.31)	39.90 (0.80)	35.66 (0.84)	35.81 (0.65)	33.65 (0.44)
Education ²⁹	12.99 (0.05)	13.34 (0.08)	11.88 (0.17)	12.85 (0.18)	13.04 (0.14)	12.74 (0.10)
Race (1=Black)	0.21 (0.01)	0.24 (0.01)	0.23 (0.03)	0.17 (0.03)	0.16 (0.02)	0.19 (0.02)
No. Children ³⁰	0.25 (0.01)	0.20 (0.01)	0.24 (0.03)	0.34 (0.02)	0.24 (0.02)	0.33 (0.02)
aged 0-2 years						
No. Children aged 3-5 years	0.27 (0.01)	0.22 (0.01)	0.26 (0.03)	0.26 (0.02)	0.33 (0.02)	0.36 (0.02)
No. Children aged 6-17 years	0.96 (0.02)	0.92 (0.03)	0.96 (0.07)	0.67 (0.06)	1.21 (0.06)	1.01 (0.04)
Husband's Labor ³¹ Income (\$1000)	27.30 (0.38)	25.85 (0.49)	32.59 (1.73)	28.49 (1.26)	29.93 (1.48)	26.40 (0.64)
Participation	0.71 (0.01)	1 -	0 -	0.51 (0.02)	0.54 (0.02)	0.57 (0.01)
No. years worked ³²						
zero	9.34	-	100	-	-	-
one	5.90	-	-	20.79	14.54	10.40
two	5.51	-	-	15.73	15.86	9.98
three	6.29	-	-	12.92	14.54	14.97
four	7.18	-	-	11.24	12.33	20.37
five	9.39	-	-	15.73	20.26	24.32
six	9.29	-	-	23.60	22.47	19.96
seven	47.10	100	-	-	-	-
Sample size	1752	825	164	153	196	414

Note: Standard error of means σ/\sqrt{n} in parentheses. Sample selection criteria: continuously married couples, aged 18-60 in 1980, with positive husband's annual earnings and hours worked each year.

²⁹ Years of Education are imputed from the following categorical scheme: 1 = '0-5 grades' (2.5 years); 2 = '6-8' (7 years); 3 = '9-11' (10 years); 4 = '12' (12 years); 5 = '12 plus non-academic training' (13 years); 6 = 'some college' (14 years); 7 = 'college degree, not advanced' (16 years); 8 = 'college advanced degree' (18 years). Education is measured as the highest level reported in the 1980-86 surveys.

³⁰ Sample averages: child variables based on 8 observations from waves 12-19 of the PSID; participation and male earnings based on 7 observations from 1979 to 1985.

³¹ The amounts are computed in constant (1987) dollars deflated by the consumer price index (CPI).

³² Column percentages.

Table 10: Estimates of Married Women's Participation Outcomes

	Static Probit+RE (1)	MSL, RE AR(1)+SD(1) (2)	Semi-parametric Probit (3)
y_{t-1}	–	1.117	1.089
	–	(0.528)	(0.077)
y_{mp}	-0.312	-0.007	-0.221
	(0.045)	(0.017)	(0.012)
y_{mt}	-0.106	-0.004	-0.106
	(0.026)	(0.028)	(0.056)
#Kid0-2 $_{t-1}$	-0.022	-0.117	-0.055
	(0.010)	(0.013)	(0.048)
#Kid0-2 $_t$	-0.330	-0.380	-0.316
	(0.021)	(0.145)	(0.061)
#Kid3-5 $_t$	-0.400	-0.206	-0.137
	(0.015)	(0.027)	(0.028)
#Kid6-17 $_t$	-0.120	-0.056	-0.062
	(0.011)	(0.037)	(0.011)
Cov. Parameters			
σ_v^2	0.786	0.313	–
	(0.071)	(0.323)	–
ρ	–	-0.146	–
	–	(0.140)	–
Average Partial Effects			
No y_{t-1}	0.220	–	–
$y_{t-1} = 0$	–	0.104	0.067
$y_{t-1} = 1$	–	0.323	0.275

Note: Bootstrap standard errors are reported in parentheses, using 100 bootstrap replications. The models in the first two columns are estimated using full seven years of data but the last two columns are estimated over three-period data. The average partial effects are reported at $X_t=(y_{mp}=\bar{y}_{mp}, y_{mt}=\bar{y}_{mt}, \#Kid0-2_{t-1}=0, \#Kid0-2_t=1, \#Kid3-5_t=0, \#Kid6-17_t=0, age_t=\text{mean of age}, education_t=\text{mean of education}, race_t=\text{Non-black})$.

Table 11: Predicted Frequencies of Married Women's Participation Outcomes

	Sample Distribution (1)	Static Probit+RE (2)	MSL, RE AR(1)+SD(1) (3)	Semi-parametric Probit (4)
No. years worked				
zero	9.34 —	12.32 (0.003)	6.26 (0.005)	12.25 (0.003)
one	5.90 —	15.15 (0.005)	5.20 (0.005)	10.07 (0.004)
two	5.51 —	7.09 (0.005)	5.70 (0.005)	2.96 (0.004)
three	6.29 —	6.14 (0.005)	7.19 (0.006)	2.69 (0.003)
four	7.18 —	6.57 (0.005)	9.11 (0.006)	2.74 (0.003)
five	9.39 —	8.42 (0.006)	12.91 (0.007)	3.72 (0.004)
six	9.29 —	21.93 (0.006)	19.72 (0.009)	26.42 (0.004)
seven	47.10 —	22.38 (0.004)	33.92 (0.008)	39.15 (0.004)
Total	100	100	100	100

Note: Frequencies are computed as average values of 1000 predicted outcomes of 7 periods. They are reported in percentages and their standard deviations are reported in parentheses. The unobserved covariate U_{it} in the Semi-parametric Probit model is generated using the estimated parameters (σ_v^2, ρ) in column (2).