Nonparametric Identification of Discrete Choice Models with Lagged Dependent Variables

Benjamin Williams*

George Washington University

April 23, 2019

Abstract

Variation in covariates can be used to nonparametrically identify a discrete choice model with a lagged dependent variable and discrete unobserved heterogeneity (Kasahara and Shimotsu, 2009; Browning and Carro, 2014). In some cases the number of support points of the unobserved heterogeneity distribution is restricted only by the number of points of support in the distribution of the covariates. This paper provides conditions under which these models can be identified with continuous heterogeneity using continuous variation in the covariates. The identification argument is related to that of Honore and Lewbel (2002) in that it requires a "special regressor" (Lewbel, 1998), but it does not assume an additively separable latent index. Identification requires only 3 time periods, neither stationarity nor time homogeneity is imposed, and the distribution of the initial condition is not restricted apart from conditions required for the special regressor. I demonstrate the result through a Monte Carlo simulation.

JEL codes: C14, C23, C25

Keywords: discrete choice, binary choice, nonparametric panel data, lagged dependent variable, special regressor

*Department of Economics, George Washington University, 2115 G St. NW, Washington DC 20052, phone: (202)994-6685, email: bdwilliams@gwu.edu. This paper has benefited from comments from Yingyao Hu, Shakeeb Khan, Arthur Lewbel, Bob Phillips, Nicholas Papageorge, Joris Pinkse, Yuya Sasaki, Susanne Schennach, Suyong Song, and conference and seminar participants at the 2014 Greater NY Metropolitan Area Econometric Colloquium, the 2018 Southern Economic Association meetings, the 2018 Midwest Econometric Group meeting, Boston College, Johns Hopkins University, and the University of Iowa, as well as the editor and three anonymous referees. This paper was previously titled "Nonparametric Identification of Binary Choice Models with Lagged Dependent Variables".

1 Introduction

This paper provides conditions under which a discrete choice panel model with a lagged dependent variable and continuously distributed individual effects is nonparametrically identified using continuous variation in the covariates. I use a spectral decomposition due to Hu and Schennach (2008) to show nonparametric identification with as few as three periods of data. The argument does not require specification of the distribution of the initial condition. It also allows for general time-dependence of the model parameters. The main identifying assumption is that there is a continuously distributed time-varying covariate, V_{it}, with large support, that satisfies a conditional independence condition. This assumption is similar to the assumption of the existence of a special regressor in Honore and Lewbel (2002).[1] The argument also relies on a version of the completeness condition, which has received increased attention recently (Canay et al., 2013; Andrews, 2017; D'Haultfoeuille, 2011; Hu and Shiu, 2018).

Consider the following dynamic binary choice model. For each time period $t = 1, \ldots, T$ and each observational unit $i = 1, \ldots, n$,

$$Y_{it} = \mathbf{1}(\delta_t + \gamma_t Y_{i(t-1)} + \beta_t' W_{it} + \alpha_t F_i \geq U_{it}), \tag{1.1}$$

where U_{it} are independent errors, F_i is the unobserved individual effect, and W_{it} is a vector of observed covariates. This model incorporates both state dependence and unobserved heterogeneity as sources of serial dependence in outcomes. It is well known that maximum likelihood estimation of this model, with the distribution of U_{it} specified but the individual effects treated as parameters to be estimated, suffers from inconsistency due to the incidental parameters problem unless $T \to \infty$ (Neyman and Scott, 1948).[2] In some cases, the individual effects can be removed by conditioning on a sufficient statistic if the coefficients in (1.1) are time-invariant and the errors are all logistically distributed (Rasch, 1961; Andersen, 1970; Honore and Kyriazidou, 2000; Aguirregabiria et al., 2018). This approach has been extended to a semiparametric model, where neither the

[1] Chen et al. (2018b), who label this an "excluded" covariate, show that the conditional independence assumption in Honore and Lewbel (2002) implicitly requires serial independence of the special regressor. They also provide a different identification argument under an alternative assumption regarding the special regressor that allows for serial dependence. The assumption used in this paper is neither stronger nor weaker than either of these but it does allow for serial dependence.

[2] Chen et al. (2018a), Boneva and Linton (2017) and Ando and Bai (2018) study estimation of this model with large T.

distribution of F_i nor the distribution of U_{it} is specified, though stationarity of the unobservables, time-invariance of the coefficients, and linearity of the latent index are maintained (Manski, 1987; Honore and Kyriazidou, 2000; Honore and Lewbel, 2002). An alternative is a random effects specification where the distribution of F_i conditional on covariates, in addition to the distribution of U_{it}, is specified (up to a finite-dimensional parameter). As shown by Heckman (1981a), specification of the random effects distribution is complicated by the presence of the lagged dependent variable.

Models based on equation (1.1) have been employed in a vast empirical literature. Among many other applications, such models have been used to study labor force participation (Heckman, 1981b,a; Hyslop, 1999), brand switching behavior (Chintagunta et al., 2001), health (Contoyannis et al., 2004), educational attainment (Cameron and Heckman, 1998, 2001), stock market participation (Alessie et al., 2004), and firm behavior (Roberts and Tybout, 1997; Kerr et al., 2014). The main result of this paper shows that the conditional choice probabilities, $Pr(Y_{it} = y_t \mid Y_{i(t-1)} = y_{t-1}, W_{it} = w_t, F_i = f)$, the initial condition distribution, $Pr(Y_{i1} = y_1 \mid W_{i1} = w_1, \ldots, W_{iT} = w_T, F_i = f)$, and the density of the individual effect, $f_{F_i|W_{i1}=w_1,\ldots,W_{iT}=w_T}$, are identified. While the result applies to the model of equation (1.1) as a special case, it does not require a linear latent index and also allows for general discrete-valued outcomes.

Let $W_{it} = (V_{it}, X_{it}')'$, where V_{it} is a scalar. The main identifying assumption is that V_{it} is a time-varying, continuously distributed covariate satisfying two exclusion restrictions. The restrictions are that neither the initial condition distribution nor the density of the individual effect depends on $V_{i2}, \ldots, V_{iT}$. That is, (Y_{i1}, F_i) is conditionally independent of $(V_{i2}, \ldots, V_{iT})$ given $(V_{i1}, X_{i1}, \ldots, X_{iT})$. The first exclusion restriction is implicitly imposed in most of the empirical applications cited above, nearly all of which assume that the initial conditions distribution depends only on initial period covariates. The second exclusion restriction is also imposed in many of the empirical models cited above. For example, Cameron and Heckman (2001), Alessie et al. (2004), Contoyannis et al. (2004) and Roberts and Tybout (1997) assume that the individual effect is independent of all covariates. Kerr et al. (2014) use the Mundlak (1978) projection, assuming that $F_i \mid W_i$ is normally distributed with mean $\gamma' \left(T^{-1} \sum_{t=1}^{T} W_{it} \right)$. Hyslop (1999) allows the individual effect to be correlated with non-labor income and the presence of children in the household in estimating a model of female labor supply but finds that these correlations are statistically insignificant.

The second exclusion restriction is similar to the special regressor assumption of Honore and Lewbel (2002). As they argue, this assumption is natural in many cases where $\gamma_t Y_{i(t-1)} + \beta_{1t}' X_{it} + \alpha_t F_i$ represents a measure of "benefits" and $\beta_{2t} V_{it}$ represents an observed "cost" of a decision. In an economic model where Y_{it} represents an individual choice, F_i often allows for variation in individual preferences, abilities, or character traits. In such models, while X_{it} may include many observed individual characteristics that are naturally correlated with F_i, V_{it} could be related to an institutional factor that leads to exogenous variation in costs of participation across individuals. Another potential application is an extension of Lewbel et al. (2011)'s model for contingent valuation, or willingness to pay, to allow for state dependence and unobserved heterogeneity, where the special regressor would be assigned as part of the experimental design. Unlike in Honore and Lewbel (2002), the special regressor, V_{it}, has to be time-varying for the identification result in this paper. More specifically, the conditional support of $V_{i2} \mid V_{i1}$ must be the same as the unconditional support of V_{i2}. Honore and Lewbel (2002) are able to avoid this support restriction because of their stationarity and time-invariant coefficients assumptions. However, this rules out some of the potential special regressors suggested by Honore and Lewbel (2002), such as age or date of birth, for which V_{i2} is a deterministic function of V_{i1}.

There are few results in the literature regarding nonparametric identification of models like the model of equation (1.1). Kasahara and Shimotsu (2009) provide several nonparametric identification results for finite mixture models of dynamic discrete choice. The only result in that paper that allows for lagged dependent variables requires $T \geq 6$ and imposes stationarity, that is, time-invariance of the conditional choice probabilities. Moreover, they assume throughout that the individual effect has a finite support. Hu and Shum (2012) do not assume finite support of F_i but require $T = 5$ and do not prove identification of the initial conditions distribution or the distribution of the individual effects. Shiu and Hu (2013) show that by strengthening some of the assumptions in Hu and Shum (2012), identification is possible with only 2 periods of data on the outcome and 3 periods of data on covariates. Both Shiu and Hu (2013) and Hu and Shum (2012) use an argument that treats the observed covariates as a sort of proxy for F_i. Browning and Carro (2014) provide several interesting results for the binary choice case. However, they do not provide any results for the nonstationary case with observed covariates and they assume the support of F_i is finite.

My identification argument suggests nonparametric and semiparametric sieve maximum likelihood estimators (MLE) that are a natural generalization of the standard random effects MLE. I provide simulations that demonstrate the practical importance of the new identification result through a Monte Carlo study of a particular semiparametric sieve MLE in the binary choice model of equation (1.1). Through these simulations I show that a random effects MLE of the coefficients in the linear latent index model can be biased when the initial conditions distribution is misspecified. A semiparametric sieve MLE that treats both the initial conditions distribution and the distribution of individual effects nonparametrically effectively eliminates the bias and in some cases leads to a reduction in the mean squared error.

The remainder of this paper is organized as follows. Section 2 provides some additional discussion of the related literature. In Section 3, I discuss the model and assumptions. Section 4 provides the main identification result. Section 5 discusses lower level conditions for the assumptions stated in Section 3. Section 6 reports the results of the Monte Carlo study and Section 7 concludes.

2 Related literature

One common solution to the incidental parameters problem (Neyman and Scott, 1948) is the random effects approach in which the (conditional) distribution of the unobserved individual effects is modeled parametrically. Heckman (1981a) noted that in a dynamic model this approach requires specifying the distribution of initial conditions as well.[3] The random effects approach has been extended subsequently to minimize the restrictions that must be imposed (see, inter alia, Newey, 1994; Arellano and Carrasco, 2003; Gayle and Namoro, 2013).

A separate approach is to avoid the specification of random effects and initial conditions distributions entirely. In a linear model, individual effects can be differenced out, leading to identification of a broad class of models where the dependence between individual effects and covariates is unrestricted (Ahn et al., 2001, 2013; Anderson and Hsiao, 1982; Arellano and Bond, 1991; Arellano and Bover, 1995; Holtz-Eakin et al., 1988). In some cases the differencing approach can be extended to nonlinear models (see Bonhomme, 2012, for a general treatment). For example, Rasch (1961) and Andersen (1970) showed that the differencing approach can be extended to the static binary choice

[3] Alternatively, Wooldridge (2005) recommended conditioning on the initial condition.

panel model if the errors are logistically distributed, and Honore and Kyriazidou (2000) showed the same for the dynamic binary choice model.

The differencing approach has been extended to a semiparametric model for binary outcomes as well. Manski (1987) obtained identification of the static binary choice model without any distributional assumptions and Honore and Kyriazidou (2000) extended this idea to a model with lagged dependent variables. Honore and Lewbel (2002) provide an alternative argument for the same linear latent index model that uses the special regressor method of Lewbel (1998). Cameron and Heckman (1998) prove a similar result for a dynamic model of educational attainment, though they assume all covariates are independent of the unobserved individual effects. Recently, Khan et al. (2016), Pakes and Porter (2016), and Shi et al. (2018) have extended this idea to a semiparametric model for multinomial outcomes. Honore and Tamer (2006) demonstrate that if the support conditions on the regressors required by Honore and Kyriazidou (2000) do not hold then the model is not identified if the distribution of the errors is specified as something other than logistic, though the identified set for some parameters can be very small. Chernozhukov et al. (2013) apply similar ideas to a nonseparable model, deriving nonparametric and semiparametric bounds for average and quantile effects.

Browning and Carro (2007) demonstrate how the semiparametric model restricts the nature of unobserved heterogeneity because of the reliance on a linear latent index. Kasahara and Shimotsu (2009) and Browning and Carro (2014) provide identification results for dynamic binary choice models by imposing a finite support for the individual effects. They prove several important results showing how identification depends on the number of support points for the distribution of individual effects relative to the length of the panel. Another key insight of these papers is that variation in the histories of the covariates in the presence of restrictions on the dependence between the covariates and the individual effects, increases the number of allowable support points of the individual effect for a fixed panel length.

Several other related results allow for a continuous support for the individual effects when the dependent variable is continuously distributed and/or a continuous proxy is available. This paper is most closely related to several other papers that apply nonparametric identification results from the measurement error literature (Hu, 2008; Hu and Schennach, 2008; Carroll et al., 2010) to show identification of a panel data model with unobserved effects (Hu and Shum, 2012; Shiu and Hu,

2013; Sasaki, 2015; Freyberger, 2018). Only Hu and Shum (2012) and Shiu and Hu (2013) allow the outcome to be discrete. These papers also allow the unobserved individual effect to be time-varying. However, they rely on sufficient dependence between a continuously distributed covariate and the individual effects. In contrast, this paper shows that identification can instead be attained if the latent index in a random utility model for the binary outcomes is a sufficient proxy for the individual effect. The result does not require dependence between the individual effects and the covariates.

3 Model

Let $V_i = (V_{i1}, \ldots, V_{iT})'$, $X_i = (X_{i1}', \ldots, X_{iT}')'$, $W_i = (W_{i1}', \ldots, W_{iT}')'$, and $Y_i = (Y_{i1}, \ldots, Y_{iT})'$. Throughout the discussion of the model and assumptions and the identification analysis I treat the joint distribution of (Y_i, W_i) as known and leave the dependence on "i" implicit in the notation. I also use a superscript to denote a partial history, that is, $Z^{(t)} = (Z_1, \ldots, Z_t)$ for any variable Z in the model. Let $\mathcal{Y}_t$, $\mathcal{V}_t$, $\mathcal{X}_t$, and $\mathcal{W}_t$ denote the supports of Y_t, V_t, X_t, and W_t, respectively, for each t, with $\mathcal{Y}_t \subset \mathbb{R}$, $\mathcal{V}_t \subseteq \mathbb{R}$ and $\mathcal{X}_t \subseteq \mathbb{R}^K$. Let $\mathcal{F} \subseteq \mathbb{R}$ denote the support of F and let $\mathcal{Y}, \mathcal{V}, \mathcal{X}, \mathcal{W}$ denote the supports of Y, V, X, and W, respectively. I consider only the case where $|\mathcal{Y}_t| < \infty$, meaning that Y_t is a discrete random variable.

For each $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, I define the observed choice probabilities, $p(y \mid w) = Pr(Y = y \mid W = w)$ and for each $y_t \in \mathcal{Y}_t$, $y_{t-1} \in \mathcal{Y}_{t-1}$ $w_t \in \mathcal{W}_t$, and $f \in \mathcal{F}$ I define the conditional choice probabilities, $p_t(y_t \mid y_{t-1}, w_t, f) = Pr(Y_t = y_t \mid Y_{t-1} = y_{t-1}, W_t = w_t, F = f)$.

Assumption 3.1. *For all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$,*

$$p(y \mid w) = \int \prod_{t=2}^{T} p_t(y_t \mid y_{t-1}, w_t, f) f_{Y_1, F \mid V_1, X}(y_1, f \mid v_1, x) df \tag{3.1}$$

This represents the conditional distribution observed in the data, $p(y \mid w)$, in terms of the underlying conditional choice probabilities, p_t, and $f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) = p_1(y_1 \mid v_1, x, f) f_{F|V_1,X}(f \mid v_1, x)$ where $p_1(y_1 \mid v_1, x, f) = Pr(Y_1 = y_1 \mid V_1 = v_1, X = x, F = f)$ represents the conditional distribution of the initial condition and $f_{F|V_1,X}$ is the conditional density of the unobserved individual effect given the initial value of the covariate V and the full history of the covariates X. In the panel

data literature it is standard to factor $p(y \mid w)$ as $\int \prod_{t=2}^{T} p_t(y_t \mid y_{t-1}, v_t, x_t, f) f_{Y_1,F|W}(y_1, f \mid w) df$, treating W and F as exogenous. Assumption 3.1 imposes two additional restrictions in addition to assuming that this standard factorization holds. First, the initial condition, p_1 depends only on V_1 and X and not on the remaining components of V. Second, the individual effect F is also independent of $V_2, \ldots, V_T$ conditional on V_1 and X. The discussion here will focus on these two additional restrictions. However, in Section A.1 in the appendix I discuss the implications of Assumption 3.1 in the context of a dynamic discrete choice model where (Y, W) are modeled jointly as a Markov process conditional on F.

First, consider the restriction on the individual effects implied by Assumption 3.1. This restriction is similar to the special regressor condition of Lewbel (1998) in that it requires at least one of the covariates to be conditionally independent of the individual effect.[4] The following proposition shows that it can be viewed as a restriction on the dynamic process for the covariates conditional on F.

Proposition 3.1. *Let $X_t = (X_{t1}', X_{t2}')'$. Suppose that for each $t \geq 2$,*

(i) $(V_t, X_{t1}) \perp\!\!\!\perp F | X_{t2}, W^{(t-1)}$

(ii) $X_{t2} \perp\!\!\!\perp (V^{(t-1)}, X_1^{(t-1)}) | X_2^{(t-1)}, F$

Then $F \perp\!\!\!\perp (V_2, X_{21}, \ldots, V_T, X_{T1}) | V_1, X_{11}, X_2^{(t-1)}$.

The proof is provided in Section A.1 in the appendix.

The implication of Proposition 3.1 is that it is sufficient to be able to split the covariates W_t into two types. The first type, V_t and X_{t1}, can depend on the current and past values of the other covariates but are conditionally independent of the individual effect F. The second type can depend on F as well as current and past values of all variables of this type but are conditionally independent of past values of the other type. This is a natural assumption in discrete choice models of demand where V_t and X_{t1} consist of price and other product characteristics and X_{t2} consists of consumer characteristics. Chintagunta et al. (2001), for example, use scanner data for a sample of consumers in a small U.S. city. Product level prices, advertisements, and the presence of store displays vary across consumers because different consumers in the sample shop at many

[4] Unlike in Honore and Lewbel (2002), who also apply a special regressor condition in a panel data model, this conditional independence is not conditional on lags of Y_t, thus avoiding the problem identified by Chen et al. (2018b).

different stores at different times over a period of two years. The individual characteristics include variables such as household income and household size. They motivate their analysis using a dynamic discrete choice model where F represents unobserved heterogeneity across consumers in tastes for the different products. In this setting the product prices are good candidates for the special regressor V_t. By Proposition 3.1, Assumption 3.1 can allow for unrestricted dependence between household characteristics and unobserved tastes and also for the possibility that changes over time in household income can effect the prices consumers face. The latter is important to account for as household income affects what neighborhood they live in as well as when and where they shop.

Next, the restriction on the distribution of the initial condition is easier to justify if period $t = 1$ indicates the initial period of the dynamic process and not just the first period observed by the econometrician. In that case it is standard to make the even stronger assumption that $Pr(Y_1 = y_1 \mid W, F) = Pr(Y_1 = y_1 \mid W_1, F)$, as would be the case if $Y_1 = r_1(W_1, F, U_1)$ for some function r_1 and unobservable U_1 that is independent of (W, F) (strict exogeneity in the initial period). Consider, however, the case where the dynamic process started at some point in the past and $t = 1$ is the first period observed. Let Y^0 denote the vector of past outcomes, going back to the initial period of the dynamic process. Then $Pr(Y_1 = y_1 \mid W, F) = \sum_{y^0} Pr(Y_1 = y_1 \mid Y^0 = y^0, W, F)P(Y^0 = y^0 \mid W, F)$. Next, let W^0 denote the vector of past covariates so that we can write $Pr(Y^0 = y^0 \mid W, F) = \int Pr(Y^0 = y^0 \mid W, F, W^0)dF_{W^0|W,F}$. Then, if we assume that $Pr(Y_1 = y_1 \mid Y^0 = y^0, W, F) = Pr(Y_1 = y_1 \mid Y^0 = y^0, W_1, F)$ and that $Pr(Y^0 = y^0 \mid W, F, W^0) = Pr(Y^0 = y^0 \mid F, W^0)$, it only remains to justify the assumption that $F_{W^0|W,F} = F_{W^0|V_1,X,F}$. This is satisfied, for example, under a first order Markov assumption that $F_{W_t|W^{(t-1)},W^0,F} = F_{W_t|W_{t-1},F}$ as this implies that the density, $f_{W^0|W,F}$, satisfies

$$\begin{aligned} f_{W^0|W,F} &= \frac{f_{W_2,\ldots,W_T|W_1,F} f_{W_1|W^0,F} f_{W^0|F}}{\int f_{W_2,\ldots,W_T|W_1,F} f_{W_1|W^0,F} f_{W^0|F} dW^0} \\ &= \frac{f_{W_1|W^0,F} f_{W^0|F}}{\int f_{W_1|W^0,F} f_{W^0|F} dW^0} = f_{W^0|W_1,F} \end{aligned} \tag{3.2}$$

While this result treats V and X interchangeably, conditions similar to those in Proposition 3.1 could also be used to show that $F_{W^0|W,F} = F_{W^0|V_1,X,F}$.

For the rest of the analysis consider the case with $T = 3$. The identification argument involves using $p(y \mid w)$ as the kernel of an integral operator and using Assumption 3.1 to derive an infinite-dimensional eigenvalue decomposition under sufficient "rank" conditions. Formulating the operator equivalences requires assumptions regarding the support of W, since $p(y \mid w)$ is only observed for $w \in \mathcal{W}$, and also requires the functions involved in the factorization of $p(y \mid w)$ to be bounded. Conveniently, the functions $p_t(y_t \mid y_{t-1}, v_t, x_t, f)$ and $p_1(y_1 \mid v_1, x, f)$ are bounded by definition.

Assumption 3.2. *$\mathcal{W} = \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{V}_3 \times \mathcal{X}$ and $|\mathcal{X}| < \infty$.*

Assumption 3.3. *The density $f_{F|V_1,X}$ is bounded.*

The requirement that V has rectangular support is necessary because identification is based on probabilities conditional on V rather than the density of V. The latter is defined for v outside of the support of V but the former is not. This is the reason that Hu and Schennach (2008), Hu and Shum (2012), and Shiu and Hu (2013) do not need such an assumption. On the other hand, Kasahara and Shimotsu (2009) require a similar support restriction. If instead, $\mathcal{V}$ contains a rectangular subspace then identification on this subspace is possible. However, this weaker condition still rules out time-invariant V_t. While V_1, V_2, and V_3 will be treated as continuously distributed random variables, Assumption 3.2 imposes that X is discrete. This simplifies the analysis but does not appear to be necessary.

Next, let $\mathcal{Y}_t = \{y_{t1}, y_{t2}, \ldots, y_{tJ_t}\}$ for each t. The next three assumptions will be stated relative to a baseline choice in periods 1 and 2, y_{11} and y_{21}.

Assumption 3.4. *For any $x_3 \in \mathcal{X}_3$,*

$$Pr\left(\exists f^* \in \mathcal{F} \text{ s.t. } p_3(Y_3 \mid y_{21}, V_3, x_3, f^*) = p_3(Y_3 \mid y_{21}, V_3, x_3, f)\right) < 1 \text{ for almost all } f \in \mathcal{F}.$$

This is a monotonicity assumption analogous to Assumption 4 in Hu and Schennach (2008) and Assumption 3.4 in Shiu and Hu (2013). It is satisfied if for any $x_3 \in \mathcal{X}_3$, any $v_3 \in \mathcal{V}_3$, and any $y_3 \in \mathcal{Y}_3$, $p_3(y_3 \mid y_{21}, v_3, x_3, \cdot)$ is a strictly monotonic function over $\mathcal{F}$. The statement of the assumption also allows for $p_3(y_3 \mid y_{21}, v_3, x_3, \cdot)$ to be flat for some values of v_3 and y_3 and accounts for the technicality that $p_3(y_3 \mid y_{21}, v_3, x_3, \cdot)$ can always be redefined on a set of measure 0 in $\mathcal{F}$

without affecting the distribution of the data. The role of this assumption in the identification argument is to prevent multiplicity of eigenvalues in the infinite-dimensional eigenvalue problem.

Assumption 3.5. *For each $y_1 \in \mathcal{Y}_1$ and $x_2 \in \mathcal{X}_2$, there exists a known $\bar{v}_2 \in \mathbb{R} \cup \{-\infty, \infty\}$ and a known $0 < \ell \leq 1$ such that $\lim_{v_2 \to \bar{v}_2} p_2(y_{21} \mid y_1, v_2, x_2, f) = \ell$ for all $f \in \mathcal{F}$. If $|\bar{v}_2| < \infty$ then $\bar{v}_2 \in \mathcal{V}_2$. If $\bar{v}_2 = \pm\infty$ then $\mathcal{V}_2$ is unbounded from above or below, respectively.*

Assumption 3.5 is similar to an identification at infinity condition. It is used to fix the scale of the eigenfunctions, which, unlike in the typical application, are not densities and hence do not integrate to 1. It requires that there is a limiting case in the support of V_2 at which the choice probability is known. In many cases this will be satisfied with $\ell = 1$, provided that the support of V_2 is large enough, meaning that in the limiting case the individual chooses $y_2 = y_{21}$ with certainty.[5] However, this assumption also allows for a case where V_2 represents time given to make a decision and as $V_2 \to 0$ the choice becomes "random", meaning that the choice probability converges to one over the number of alternatives. This assumption is not needed if F is discrete, as shown in Section A.3 in the appendix. The proof in that case suggests that it might also not be needed in the continuous case. Furthermore, in Section 5 for a linear latent index model I state lower level conditions for the "rank" conditions stated below in Assumption 3.7 and find that Assumption 3.5 is implied by these lower level conditions.

Assumption 3.6. *There exists $w_{20} = (v_{20}, x_{20}) \in \mathcal{W}_2$ and a known one-to-one function, $\pi : \mathbb{R} \to [0, 1]$, with $\pi(\mathbb{R}) = [0, 1]$ such that $\lim_{w_2 \to w_{20}} p_2(y_{21} \mid y_{11}, w_2, f) = \pi(f)$ for almost all $f \in \mathcal{F}$. The support $\mathcal{X}$ satisfies the condition that for each $x_3 \in \mathcal{X}_3$, there exists $x_1 \in \mathcal{X}_1$ such that $(x_1, x_{20}, x_3) \in \mathcal{X}$.*

The first part of Assumption 3.6 is a normalization of the function $p_2(y_{21} \mid y_{11}, w_{20}, \cdot)$.[6] A normalization is required because taking any monotonic transformation of F in the model produces an observationally equivalent model. Other normalizations are possible but this is the nonparametric version of the most common type of normalization in parametric versions of this model, as

[5] While this limiting case is often $\bar{v}_2 = \pm\infty$, if y_2 represents labor force participation and v_2 represents non-labor income then $\bar{v}_2 = 0$ may be a plausible alternative that does not require the support of non-labor income to be unbounded.

[6] As a conditional probability, $p_2(y_{21} \mid y_{11}, w_2, f)$ can have a removable discontinuity at w_{20} because it is only uniquely defined up to a set of measure 0. The assumption implicitly requires that it does not have any other type of discontinuity because it assumes that the limit exists.

discussed in Section 5. The second part of the assumption, the condition on the support of X, is needed so that the normalization can be applied as X_3 is varied. It rules out time-invariant covariates as we cannot distinguish a time-invariant covariate from the individual effect F in the model. Time-invariant covariates could be allowed if they are independent of F, though this would require a different identification argument (cf. Hausman and Taylor, 1981).

The last assumption is the injectivity of two integral operators.

Assumption 3.7.

(i) *For each* $y_1 \in \mathcal{Y}_1$ *and each* $x \in \mathcal{X}$*, if* $\psi \in \mathcal{L}^\infty(\mathcal{F})$ *and* $\int_{\mathcal{F}} f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x)\psi(f)df = 0$ *for all* $v_1 \in \mathcal{V}_1$ *then* $\psi \equiv 0$*.*

(ii) *For each* $y_1, y_2 \in support(Y_1, Y_2)$ *and each* $x_2 \in \mathcal{X}_2$*, if* $\psi \in \mathcal{L}^1(\mathcal{F})$ *and* $\int_{\mathcal{F}} p_2(y_2 \mid y_1, v_2, x_2, f)\psi(f)df = 0$ *for all* $v_2 \in \mathcal{V}_2$ *then* $\psi \equiv 0$*.*

This assumption is analogous to a full rank condition on two matrices in the finite-dimensional case where F is discrete. Operator injectivity conditions are common in nonseparable models of measurement error (Hu and Schennach, 2008; Carroll et al., 2010) and are closely related to completeness conditions in nonparametric IV models (Canay et al., 2013; Andrews, 2017; D'Haultfoeuille, 2011; Hu and Shiu, 2018). For example, condition (i) is satisfied if the family of conditional distributions $V_1 \mid Y_1, F, X$ is complete. On the other hand, if F is independent of V_1 conditional on X then $\int_{\mathcal{F}} f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x)\psi(f)df = \int_{\mathcal{F}} p_1(y_1 \mid f, v_1, x)f_{F|X}(f \mid x)\psi(f)df$ and in this case condition (i) can apparently not be stated as a completeness condition generally, though I show in Section 5 that it can be for the case of a linear latent index model such as that of equation (1.1), or more generally under a monotonicity condition. Similarly, I also show there that condition (ii) is implied by a completeness condition in these two cases as well, though it apparently cannot be in general.

4 Identification

This section provides the main identification result and an overview of its proof. The full proof is in Section A.2 in the appendix.

Theorem 4.1. *Under Assumptions 3.1-3.7, the functions p_1, p_2, p_3, and $f_{F|V_1,X}$ are uniquely determined by the function $p(y \mid w)$.*

Let $p(y_1, y_2 \mid w) = Pr(Y_1 = y_1, Y_2 = y_2 \mid W = w)$. By Assumption 3.1,

$$\begin{aligned} p(y_1, y_2 \mid w) &= \sum_{y_3 \in \mathcal{Y}_3} p(y \mid w) \\ &= \int p_2(y_2 \mid y_1, w_2, f) f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) df. \end{aligned} \tag{4.1}$$

Fix the values of y, v_3, and x and define the operators

$$\begin{aligned} [L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3} g](v_2) &= \int_{\mathcal{V}_1} p(y_1, y_2 \mid w) g(v_1) dv_1 \\ [L_{y;V_1,x_1,V_2,x_2,v_3,x_3} g](v_2) &= \int_{\mathcal{V}_1} p(y \mid w) g(v_1) dv_1 \end{aligned} \tag{4.2}$$

Both operators map any absolutely integrable function $g : \mathcal{V}_1 \to \mathbb{R}$ to a bounded real-valued function defined on $\mathcal{V}_2$.[7] If V_1 and V_2 were discrete random variables, with $\mathcal{V}_t = \{v_{t1}, \ldots, v_{tK_t}\}$ for $t = 1, 2$, then we could define two $K_2 \times K_1$ matrices, $(p(y_1, y_2 \mid v_{1k}, x_1, v_{2j}, x_2, v_3, x_3))_{j=1,\ldots,K_2,k=1,\ldots,K_1}$ and $(p(y \mid v_{1k}, x_1, v_{2j}, x_2, v_3, x_3))_{j=1,\ldots,K_2,k=1,\ldots,K_1}$. The operators $L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}$ and $L_{y;V_1,x_1,V_2,x_2,v_3,x_3}$ are the infinite-dimensional analogs of these matrices.[8] These operators are identified directly from the data and hence are treated as known. To simplify the exposition I leave the dependence on y, v_3, and x implicit in the notation where possible and refer to the operators $L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}$ and $L_{y;V_1,x_1,V_2,x_2,v_3,x_3}$ as L_1 and L_2, respectively.

Next, for any values of y, v_3, and x, define the operators

$$\begin{aligned} [\Lambda_{y_1;V_1,x,F} g](f) &= \int_{\mathcal{V}_1} f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) g(v_1) dv_1 \\ [\Lambda_{y_2;y_1,V_2,x_2,F} g](v_2) &= \int_{\mathcal{F}} p_2(y_2 \mid y_1, v_2, x_2, f) g(f) df \\ [\Delta_{y_3;y_2,v_3,x_3,F} g](f) &= p_3(y_3 \mid y_2, v_3, x_3, f) g(f). \end{aligned} \tag{4.3}$$

[7] This is because $p(y_1, y_2 \mid w)$ and $p(y \mid w)$ are bounded by 1, which implies that $\left|\int_{\mathcal{V}_1} p(y_1, y_2 \mid w) g(v_1) dv_1\right| \leq \int_{\mathcal{V}_1} |p(y_1, y_2 \mid w)||g(v_1)| dv_1 \leq \int_{\mathcal{V}_1} |g(v_1)| dv_1$ and $\left|\int_{\mathcal{V}_1} p(y \mid w) g(v_1) dv_1\right| \leq \int_{\mathcal{V}_1} |p(y \mid w)||g(v_1)| dv_1 \leq \int_{\mathcal{V}_1} |g(v_1)| dv_1$.

[8] See Section A.3 in the appendix for an analysis of identification in the discrete case.

The operator $\Lambda_{y_1;V_1,x,F}$ maps any absolutely integrable function $g : \mathcal{V}_1 \to \mathbb{R}$, to a bounded, absolutely integrable real-valued function defined on $\mathcal{F}$. The operator $\Lambda_{y_2;y_1,V_2,x_2,F}$ maps any absolutely integrable function $g : \mathcal{F} \to \mathbb{R}$ to a bounded real-valued function defined on $\mathcal{V}_2$. And $\Delta_{y_3;y_2,v_3,x_3,F}$ is a diagonal operator that takes any function $g : \mathcal{F} \to \mathbb{R}$ and returns another real-valued function defined on $\mathcal{F}$.[9] Again, I leave dependence on y, v_3, and x implicit in the notation where possible and refer to these operators as Λ_1, Λ_2, and Δ, respectively.

By Assumption 3.1, for any absolutely integrable function $g : \mathcal{V}_1 \to \mathbb{R}$,

$$\begin{aligned}
[L_2 g](v_2) &= \int_{\mathcal{V}_1} \left(\int_{\mathcal{F}} p_3(y_3 \mid y_2, w_3, f) p_2(y_2 \mid y_1, v_2, x_2, f) f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) df \right) g(v_1) dv_1 \\
&= \int_{\mathcal{F}} p_2(y_2 \mid y_1, v_2, x_2, f) p_3(y_3 \mid y_2, w_3, f) \left(\int_{\mathcal{V}_1} f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) g(v_1) dv_1 \right) df \\
&= \int_{\mathcal{F}} p_2(y_2 \mid y_1, v_2, x_2, f) p_3(y_3 \mid y_2, w_3, f) [\Lambda_1 g](f) df \qquad (4.4) \\
&= \int_{\mathcal{F}} p_2(y_2 \mid y_1, v_2, x_2, f) [\Delta \Lambda_1 g](f) df \\
&= [\Lambda_2 \Delta \Lambda_1 g](v_2),
\end{aligned}$$

using Fubini's theorem to interchange the order of integration. Summing over $y_3 \in \mathcal{Y}_3$ we also have that $[L_1 g](v_2) = [\Lambda_2 \Lambda_1 g](v_2)$. This defines two operator equivalences on $\mathcal{L}^1(\mathcal{V}_1) := \{g : \mathcal{V}_1 \to \mathbb{R} \text{ such that } \int_{\mathcal{V}_1} |g(v_1)| dv_1 < \infty\}$ which can be written as[10]

$$\begin{aligned}
L_1 &= \Lambda_2 \Lambda_1 \qquad (4.5) \\
L_2 &= \Lambda_2 \Delta \Lambda_1.
\end{aligned}$$

Thus, this model takes a form similar to the nonclassical measurement error model of Hu and Schennach (2008).

Following Hu and Schennach (2008), condition (ii) of Assumption 3.7 implies that Λ_2 is injective so that $\Lambda_1 = \Lambda_2^{-1} L_1$ and therefore $L_2 = \Lambda_2 \Delta \Lambda_2^{-1} L_1$. Furthermore, condition (i) of Assumption 3.7

[9] If g is absolutely integrable then Δg is as well. This result, and the other results in this paragraph, follow from (a) $f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) = p_1(y_1 \mid v_1, x, f) f_{F|V_1,X}(f \mid v_1, x)$, (b) $p_1(y_1 \mid v_1, x, f)$, $p_2(y_2 \mid y_1, w_2, f)$, and $p_3(y_3 \mid y_2, w_3, f)$ are bounded, and (c) $f_{F|V_1,X}(f \mid v_1, x)$ is absolutely integrable as a function of f, because it is a density, and is bounded by Assumption 3.3. The proofs follow the same arguments as in the footnote 6.

[10] Technically, $\mathcal{L}^1(\mathcal{V}_1)$ is defined as the space of all equivalence classes of absolutely integrable functions that are equal almost everywhere. Thus, we can apply results regarding linear operators on a Banach space.

implies that Λ_1 is surjective[11] which implies that the operator equivalence

$$L_2 L_1^{-1} = \Lambda_2 \Delta \Lambda_2^{-1}, \tag{4.6}$$

holds for all functions in the range of Λ_2.[12] Importantly, the left-hand side is known and the right-hand side involves the unknown parameters of the model.

The operator equivalence (4.6) can be viewed as a spectral decomposition, as in Hu and Schennach (2008). Applying the same spectral theorem used in Hu and Schennach (2008), I can conclude that this decomposition is unique up to (i) possible multiplicity of "eigenvalues", which are the elements of the set $\{p_3(y_3 \mid y_2, w_3, f) : f \in \mathcal{F}\}$, (ii) scaling of the "eigenfunctions", $p_2(y_2 \mid y_1, \cdot, x_2, f)$ for each $f \in \mathcal{F}$, and (iii) reordering or reindexing of the eigenvalues and associated eigenvectors. Assumption 3.4 prevents problems due to multiplicity of eigenvalues. Assumption 3.5 resolves the scale of the eigenfunctions. Assumption 3.6 is a normalization that rules out models obtained by reordering/reindexing the eigenvalues.

As noted by Hu and Schennach (2008), identification of the operators Λ_1, Λ_2 and Δ implies identification of the kernel functions $f_{Y_1,F|V_1,X}$, p_2 and p_3, respectively. Then the density $f_{F|V_1,X}$ is identified because $f_{F|V_1,X} = \sum_{y_1 \in \mathcal{Y}_1} f_{Y_1,F|V_1,X}$. And the initial condition distribution, p_1, is identified because $p_1 = f_{Y_1,F|V_1,X} / f_{F|V_1,X}$.

Importantly, Assumptions 3.4-3.7 do not impose restrictions that must hold for all values of y, v_3, x. It would be undesirable, for example, to impose the normalization in Assumption 3.6 for all values of x_2 and y_1. The reason this is not required is that as we vary y_1, y_2, x_1, x_2 we obtain different pairs of operator equivalences from (4.4) all with the same Δ, and as we vary y_3, v_3, x_1, x_3 we obtain different pairs of operator equivalences all with the same Λ_2, and as we vary y_2, y_3, v_3 we obtain different pairs of operator equivalences all with the same Λ_1. The proof proceeds by first applying the above argument for $y_1 = y_{11}$, $y_2 = y_{21}$, and $x_2 = x_{20}$. For $x_2 \neq x_{20}$ and $y_1 \neq y_{11}$, the eigenvalues in (4.6) have been identified, because p_3 does not vary with x_2 or y_1, but the eigenvectors have not. Since Assumption 3.5 holds for all x_2 and any y_1 it can be used to determine the scale of the eigenfunction associated with each eigenvalue and hence p_2 is identified at all $x_2 \in \mathcal{X}_2$ and all

[11] More formally, it implies that the range of the operator Λ_1 is dense in the domain of the operator Λ_2, $\mathcal{L}^1(\mathcal{F})$.

[12] And by extension, the equivalence is defined for all functions in the closure of the range of Λ_2, which is a Banach space, allowing us to apply the spectral theorem from Dunford and Schwartz (1971).

$y_1 \in \mathcal{Y}_1$. Thus, for any x and any y_1, $\Lambda_{y_1;V_1,x,F}$ is identified. Therefore $\Lambda_{y_2;y_{11},V_2,x_2,F}$ is identified for all $y_2 \in \mathcal{Y}_2$ since Assumption 3.7 implies that $\Lambda_{y_2;y_{11},V_2,x_2,F} = L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}\Lambda^{-1}_{y_1;V_1,x,F}$.

We will conclude this section with several remarks concerning Theorem 4.1.

Remark 1: *Theorem 4.1 does not contradict the impossibility result of Chamberlain (2010) because Assumption 3.5 either requires V_2 to have unbounded support or requires the structural error to have bounded support (see the model in the Section 5) . Chamberlain (2010), however, assumes that the structural error has unbounded support while the covariates all have bounded support.*

Remark 2: *If F and V are both discrete then we can use the matrix analogue of equations, $L_1 = \Lambda_2\Lambda_1$ and $L_2 = \Lambda_2\Delta\Lambda_1$, to show identification. See Section A.3 in the appendix. Kasahara and Shimotsu (2009) and Browning and Carro (2014) provide results for this case. However, neither paper shows identification for $T = 3$ with time-varying choice probabilities.*

Remark 3: *The result also implies identification of a static version of the model where $p_t(y_t \mid y_{t-1}, w_t, f) = p_t(y_t \mid w_t, f)$. See Section A.4 in the appendix for a more general result in the static case.*

Remark 4: *The identification result makes use of the fact that varying y_1 varies Λ_2 but not Δ. Therefore, the assumption that $Pr(Y_3 \mid Y_1, Y_2, W_3, F) = Pr(Y_3 \mid Y_2, W_3, F)$ is required when $T = 3$. However, for T sufficiently large, multiple lags of the dependent variable can be allowed, as shown in Section A.5 in the appendix.*

Remark 5: *The identification argument can be modified to allow $K_F := dim(F) > 1$ if $T \geq 2K_F + 1$. In that case, identification could be based off a version of equations (4.4) where the kernel of the operator Λ_1 is $f_{Y_1,\ldots,Y_{K_F},F|V_1,W}$, the kernel of the operator Λ_2 is $Pr(Y_{K_F+1}, \ldots, Y_{2K_F} \mid Y_{K_F-1}, W_{K_F+1}, \ldots, W_{2K_F})$ and Δ is the diagonal operator that multiplies by $Pr(Y_T \mid Y_{T-1}, W_T)$. Freyberger (2018) provides a different identification argument in a related model for $K_F \geq 1$ and $T \geq 2K_F + 1$.*

Remark 6: *As demonstrated by Shiu and Hu (2013), certain average effects can be identified without imposing the normalization in Assumption 3.6. Indeed, without Assumption 3.6, the model is equivalent to a model with $F^* = \pi^{-1}(p_2(y_{21} \mid y_{11}, w_{20}, F))$ that does satisfy Assumption 3.6. It is straightforward to show that in these two models, the average structural functions, $\int p_t(y_t \mid y_{t-1}, w_t, f) f_F(f) df$, are the same.*

Remark 7: *Suppose there is a variable Y^* that is influenced by the sequence $Y_1, \ldots, Y_3$. For example $Y_1, \ldots, Y_3$ may represent a sequence of decisions over time and Y^* an outcome of these decisions. In Section A.6 in the appendix, I give conditions under which $f_{Y^*|Y,W,X^*,F}$ is identified, where X^* is a vector of additional covariates not included in W. Identification of the model of Section 3 when $T > 3$, which does not immediately follow from Theorem 4.1, is also a special case of this result.*

Remark 8: *An alternative model for a dynamic discrete choice process posits that $Pr(Y_t \mid Y^{(t-1)}, W_t, F) = Pr(Y_t \mid \sum_{\tau=1}^{t-1} Y_\tau, W_t, F)$. This may be an appropriate model, for example, for the accumulation of human capital where Y_t indicates accumulation of an additional year of schooling in year t and hence $\sum_{\tau=1}^{t} Y_\tau$ represents the schooling level attained by year t (see, e.g., Cameron and Heckman, 2001; Heckman et al., 2016). The logic of Theorem 4.1 can readily be adapted to this case. Identification is still possible with $T = 3$ because, unlike the case where additional lags are included, we can still vary Λ_2 without varying Δ since support$(Y_1 \mid Y_1 + Y_2)$ is not a single point in $\mathcal{Y}_1$. See Aguirregabiria et al. (2018) for an application of this idea in a multinomial logit model.*

5 Further discussion of assumptions

The assumptions stated in Section 3 allow for a broad range of applications. However, these assumptions can be stated at a lower level given more structure on the model. I now study these assumptions in a binary choice panel model with a linear latent index and a binary choice model with a general nonseparable index under a monotonicity condition. I also discuss how the identification result in this paper extends or complements existing results for these particular models.

5.1 A linear latent index binary choice model

First, consider the linear latent index model,

$$Y_t = \mathbf{1}(\delta_t + \gamma_t Y_t + \beta_{t1} V_t + \beta_{t2}' X_t + \alpha_t F \geq U_t), t \geq 2, \tag{5.1}$$

where the errors $\{U_t\}_{t \geq 2}$ are mutually independent, independent of (W, F), with distribution F_{U_t}. The conditional choice probabilities in this model are given by $p_t(1 \mid y_{t-1}, v_t, x_t, f) = F_{U_t}(\delta_t + \gamma_t y_{t-1} + \beta_{t1} v_t + \beta_{t2}' x_t + \alpha_t f)$ for $t \geq 2$. Because of the conditions on U_t, Assumption 3.1 is satisfied

in this model if Y_1 is independent of (V_2, V_3) conditional on (F, V_1, X) and F is independent of (V_2, V_3) conditional on (V_1, X). The model of equation (5.1) does not provide any additional insight into whether the regularity and support conditions in Assumptions 3.2 and 3.3. However, Assumptions 3.4-3.6 can be stated at a much lower level given the structure provided by this model.

First, Assumption 3.4 is satisfied if $\alpha_3 \neq 0$ and F_{U_3} is a strictly increasing distribution function with full support on $\mathbb{R}$. However, even if U_3 does not have full support the assumption is satisfied if for each $x_3 \in \mathcal{X}_3$ and $f \in \mathcal{F}$, the support of $\delta_3 + \gamma_3 + \beta_{31} V_3 + \beta_{32}' x_3 + \alpha_3 f$ is large enough that it intersects the support of U_3. In that case, the conditional choice probability p_3 becomes flat for F large enough and V_3 fixed but the monotonicity condition is satisfied because V_3 can be sufficiently varied. Next, Assumption 3.5 is satisfied in this model with $\bar{v}_2 = sign(\beta_{21}) \cdot \infty$ and $\ell = 1$ for $y_{21} = 1$ if $\mathcal{V}_2$ is unbounded because $\lim_{v_2 \to \bar{v}_2} F_{U_2}(\delta_2 + \gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 + \alpha_2 f) = \lim_{u \to \infty} F_{U_2}(u) = 1$. Finally, Assumption 3.6 is satisfied with $w_{20} = 0$, $y_{21} = 1$ and $y_{11} = 0$ by normalizing $\delta_2 = 0$, $\alpha_2 = 1$, and $F_{U_2}(u) = \pi(u)$.

While the restrictions on α_2 and δ_2 are standard in an interactive fixed effects model, the restriction that F_{U_2} is known may seem to be a strong assumption given that Manski (1987), Honore and Kyriazidou (2000) and Honore and Lewbel (2002) achieve identification without any distributional assumptions on U_t or F. The difference is that these papers prove identification of the finite-dimensional coefficients but not of the distribution of F or U_t for $t > 2$. By contrast, Theorem 4.1 implies that, given the normalization of the distribution of U_2, the distributions of U_3 and F are nonparametrically identified.

The conditional logit approach of Honore and Kyriazidou (2000) and the random effects approach that is common in empirical applications further restrict the model of equation (5.1) by assuming that F_{U_t} is known for each t and assuming that some or all of the coefficients $\delta_t, \gamma_t, \beta_t, \alpha_t$ do not vary with t.[13][14] A random effects model, in addition, imposes a functional form for the initial conditions distribution and the distribution of the individual effects, $f_{F|W}$. In both cases, the parametric structure of the model allows for more general dependence between the covariates, W, and the individual effect, F. In the conditional logit model this dependence is fully unrestricted and in random effects models the conditional distribution of $F \mid W$ is often specified as a homoskedastic

[13] The conditional logit model also requires $T \geq 4$.

[14] In some applications of the random effects model, such as Hyslop (1999), U_t is permitted to follow an $AR(1)$ model as well.

normal with mean equal to a linear combination of some or all of the elements of W. It appears that the "special regressor" assumption $f_{F|W} = f_{F|V_1,X}$ is the cost of nonparametric identification of the model.[15]

The other cost of nonparametric identification is the operator injectivity condition of Assumption 3.7 in place of a rank condition in the (fully parametric) correlated random effects model. Consider first condition (ii) of Assumption 3.7. Suppose that F_{U_2} is differentiable. Maintaining the normalizations, $\delta_2 = 0$ and $\alpha_2 = 1$,

$$\begin{aligned} \frac{\partial p_2(1 \mid y_1, v_2, x_2, f)}{\partial v_2} &= \frac{\partial}{\partial v_2} F_{U_2}(\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 + f) \\ &= \beta_{21} f_{U_2}(\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 + f) \end{aligned} \tag{5.2}$$

Therefore, if $\int_{\mathcal{F}} p_2(1 \mid y_1, v_2, x_2, f)\psi(f) = 0$ for all $v_2 \in \mathcal{V}_2$ and an absolutely integrable function ψ then $\int_{\mathcal{F}} f_{U_2}(\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 + f)\psi(f) = 0$ for all $v_2 \in \mathcal{V}_2$ as well. This further implies that $\int_{\mathcal{F}} f_{-U_2}(u_2^* - f)\psi(f) = 0$ for all $u_2^* \in \{-(\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2) : v_2 \in \mathcal{V}_2\}$. This is a convolution equation. If the characteristic function of U_2 is non-vanishing and $\{\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 : v_2 \in \mathcal{V}_2\}$ contains the full support of U_2 then this implies that $\psi \equiv 0$. Therefore, in the linear latent index model of (5.1), injectivity condition (ii) follows from mild regularity conditions on the distribution of U_2 if V_2 has large enough support relative to the support of U_2. Note that this support condition implies the identification at infinity condition of Assumption 3.5, making that assumption redundant given Assumption 3.7 in this model.

To relate this result to a completeness condition, let $Y_2^*(y_1, x_2) := \beta_{21}^{-1}(U_2 - F - \gamma_2 y_1 - \beta_{22}' x_2)$. Then $Y_2 = \mathbf{1}(\delta_2 + \gamma_2 Y_1 + \beta_{21} V_2 + \beta_{22}' X_2 + \alpha_2 F \geq U_2) = \mathbf{1}(V_2 \geq Y_2^*(Y_1, X_2))$. In addition, if for any fixed y_1 and x_2 the family of conditional densities $\{f_{F|Y_2^*(y_1,x_2)}(f \mid v) : v \in \mathcal{V}_2\}$ is complete, meaning that $E(g(F) \mid Y_2^*(y_1, x_2) = v)$ for all $v \in \mathcal{V}_2$ for any g such that $E|g(F)| < \infty$ implies that $g = 0$ almost everywhere in $\mathcal{F}$, then condition (ii) of Assumption 3.7 is satisfied. This follows

[15] Shiu and Hu (2013) provide a different nonparametric identification result for this model that uses the observed covariates as proxies for F. By contrast, the result here does not require any dependence between the observed covariates and F.

because

$$\begin{aligned}\int_{\mathcal{F}} f_{U_2}(\gamma_2 y_1 + \beta_{21} v_2 + \beta_{22}' x_2 + f)\psi(f) &= \beta_{21}^{-1} \int_{\mathcal{F}} f_{Y_2^*(y_1,x_2)|F}(v_2 \mid f)\psi(f)df \\ &= \beta_{21}^{-1} \int_{\mathcal{F}} f_{F|Y_2^*(y_1,x_2)}(f \mid v_2) \frac{f_{Y_2^*(y_1,x_2)}(v_2)}{f_F(f)} \psi(f)df \qquad (5.3) \\ &= \frac{f_{Y_2^*(y_1,x_2)}(v_2)}{\beta_{21}} E\left(\frac{\psi(F)}{f_F(F)} \mid Y_2^*(y_1,x_2) = v_2 \right)\end{aligned}$$

Next, injectivity condition (i) requires that if $\int_{\mathcal{F}} p_1(y_1 \mid v_1, x, f) f_{F|V_1,X}(f \mid v_1, x)\psi(f)df = 0$ for all $v_1 \in \mathcal{V}_1$ for any bounded function ψ then $\psi \equiv 0$. Because the conditional choice probability p_1 and the conditional density $f_{F|V_1,X}$ both vary with v_1 and f, both functions can in principle contribute to injectivity of the operator. However, it is possible for this condition to be satisfied even if V_1 is (conditionally) independent of F so that $f_{F|V_1,X} = f_{F|X}$.

To see this in a simple case, suppose that $Y_1 = \mathbf{1}(\delta_1 + \beta_1 V_1 + \psi' X + \alpha_1 F \geq U_1)$ where U_1 is independent of (F, W). Then $p_1(1 \mid v_1, x, f) = F_{U_1}(\delta_1 + \beta_1 v_1 + \psi' x + \alpha_1 f)$. This can be rewritten as $p_1(1 \mid v_1, x, f) = F_{-U_1/\alpha_1}(u_1^* - f)$ where $u_1^* = -\alpha_1^{-1}(\delta_1 + \beta_1 v_1 + \psi' x)$, as long as $\alpha_1 \neq 0$. If, in addition, F_{U_1} is differentiable then it follows that $\int_{\mathcal{F}} p_1(y_1 \mid v_1, x, f) f_{F|V_1,X}(f \mid v_1, x)\psi(f)df = 0$ implies that $\int_{\mathcal{F}} f_{-U_1/\alpha_1}(u_1^* - f) f_{F|X}(f \mid x)\psi(f)df = 0$, which is again a convolution. Therefore, in this case, condition (i) of Assumption 3.7 holds if $\alpha_1 \neq 0$, F_{U_1} is differentiable and has a non-vanishing characteristic function, and $\{\delta_1 + \beta_1 v_1 + \psi' x : v_1 \in \mathcal{V}_1\}$ contains the full support of U_1. In addition, it can be shown that the condition follows from completeness of the family of distributions $\{f_{F|Y_1^*(x)}(f \mid v) : v \in \mathcal{V}_1\}$ where $Y_1^*(x) = \beta_1^{-1}(U_1 - \alpha_1 F - \psi' x - \delta_1)$.

If the dynamic process for $\{Y_t\}$ starts before period 1 then this model for the initial condition may be hard to motivate, as discussed in Section 3. In particular, U_1 would have a mixture distribution with the mixing probabilities varying with F. Nevertheless, injectivity condition (i) can be derived in this case from restrictions on this mixture distribution following a similar though more tedious argument.

5.2 A general latent index model

The model of equation (5.1) is popular in applied work because of its tractability. However, it often does not have a structural interpretation. Structural models can naturally imply nonlinearity.

For example, Aguirregabiria et al. (2018) give conditions under which the infinite horizon utility expected discounted utility maximization problem with per utility given by $\Pi_t(Y_t) = \alpha(y, W_t, F) + \beta(y, Y_{t-1}, W_t) + \varepsilon_t(Y_t)$ implies conditional choice probabilities of the form $p_t(y_t \mid y_{t-1}, w_t, f) = F_{U_t}(\tilde{\alpha}(W_t, F) + \tilde{\beta}(Y_{t-1}, W_t))$. They show that this result facilitates identification of $\beta(y, Y_{t-1}, W_t)$ when F_{U_t} is the cdf of the logistic distribution. If $\tilde{\alpha}(W_t, F) = \tilde{\alpha}_1' W_t + \tilde{\alpha}_2 F$ and $\tilde{\beta}(Y_{t-1}, W_t) = (\tilde{\beta}_0 + \tilde{\beta}_1' W_t) Y_{t-1}$ then this provides a structural justification for the use of the reduced form model of equation (5.1). However, these separability restrictions may be hard to justify as they do not follow from additive separability of the functions $\alpha(y, W_t, F)$ and $\beta(y, Y_{t-1}, W_t)$. The dynamic binary choice model for labor force participation in Hyslop (1999) is one example of this.

More generally, we can consider a latent index model where the latent index is nonseparable. Suppose that

$$Y_t = \mathbf{1}(r_t(Y_{t-1}, V_t, X_t, F) \geq U_t), t \geq 2 \tag{5.4}$$

and, to simplify discussion of the initial condition problem,

$$Y_1 = \mathbf{1}(r_1(V_1, X, F) \geq U_1). \tag{5.5}$$

where $\{U_t\}_{t \geq 1}$ are mutually independent, independent of (W, F), and have cumulative distribution functions F_{U_t}. As in the linear latent index model, if F is independent of V_2, V_3 conditional on V_1, X then Assumption 3.1 is satisfied, given that $V_2, \ldots, V_T$ are excluded from equation (5.5). Moreover, the conditional choice probabilities are

$$\begin{aligned} p_1(v_1, x, f) &= F_{U_1}(r_1(v_1, x, f)) \\ p_t(1 \mid y_{t-1}, v_t, x_t, f) &= F_{U_t}(r_t(y_{t-1}, v_t, x_t, f)), t = 2, 3 \end{aligned} \tag{5.6}$$

First, consider Assumption 3.4. For any (y_2, x_3), consider the random function $r^*(V_3, f) = r_3(y_2, V_3, x_3, f)$ defined on $\mathcal{F}$. Assumption 3.4 is satisfied if (i) F_{U_3} is a strictly increasing distribution function, (ii) for each $f \in \mathcal{F}$, the support of $r^*(V_3, f)$ intersects the support of U_3 and (iii) for almost all $f \in \mathcal{F}$, $Pr(\exists f^* \in \mathcal{F} \text{ s.t. } r^*(V_3, f^*) = r^*(V_3, f)) < 1$.

Second, consider Assumption 3.5. This is satisfied if there exists $\bar{v}_2$ for each (y_1, x_2, f) such that

$\lim_{v_2 \to \bar{v}_2} r_2(y_1, v_2, x_2, f)$ exceeds U_2 with probability 1. If r_2 is additively separable in v_2, that is $r_2(y_1, v_2, x_2, f) = r_{2a}(y_1, x_2, f) + r_{2b}(v_2)$, then this assumption will be satisfied if $\lim_{v_2 \to \bar{v}_2} r_{2b}(v_2) = \pm\infty$.

Assumption 3.6 is satisfied if F_{U_2} is assumed known and if $r_2(0, v_{20}, x_{20}, f) = f$ for some v_{20}, x_{20}.[16] It should be apparent that this restriction on r_2 is a generalization of the normalizations that $\delta_2 = 0$ and $\alpha_2 = 1$ in the linear latent index model. Given the structure of equation (5.4), Assumption 3.6 could instead be replaced by other functional restrictions on $r_2(0, v_2, x_2, f)$, perhaps involving integrating over the distribution of V_2, rather than evaluating at a single point in the support of V_2, along the lines of the normalizations discussed in Hu and Schennach (2008) and the subsequent literature using the same spectral decomposition approach.

Now consider Assumption 3.7 for this model, starting again with condition (ii). Suppose that $r_2(y_1, v_2, x_2, f)$ is strictly increasing in v_2 and let $r_{2V}^{-1}(u_2; y_2, x_2, f)$ denote the inverse in v_2. Then $Y_2 = \mathbf{1}(V_2 \geq r_{2V}^{-1}(U_2; Y_1, X_2, F))$. Defining $Y_2^*(y_1, x_2) = r_{2V}^{-1}(U_2; y_1, x_2, F)$, we also have that $Y_2 = \mathbf{1}(Y_2^*(Y_1, X_2) \leq V_2)$. Furthermore, because U_2 is independent of F,

$$F_{Y_2^*(y_1,x_2)|F}(y^* \mid f) = F_{U_2}(r_2(y_1, y^*, x_2, f)) \tag{5.7}$$

so that the conditional density is $f_{Y_2^*(y_1,x_2)|F}(y^* \mid f) = f_{U_2}(r_2(y_1, y^*, x_2, f)) \frac{\partial}{\partial y^*} r_2(y_1, y^*, x_2, f)$. Then we have the following result.

Proposition 5.1. *If equation (5.4) holds where $r_2(y_1, v_2, x_2, f)$ is strictly increasing in v_2 then condition (ii) of Assumption 3.7 holds if support$(Y_2^*(y_1, x_2)) \subseteq \mathcal{V}_2$ such that*

$$E(|\psi^*(F)|) < \infty \text{ and } E(\psi^*(F) \mid Y_2^*(y_1, x_2) = v_2) = 0 \text{ for all } v_2 \in \mathcal{R} \Longrightarrow \psi^*(f) \equiv 0 \tag{5.8}$$

Next, again consider condition (i) of Assumption 3.7 under the assumption that F is independent of V_1 conditional on X so that $f_{F|V_1,X} = f_{F|X}$. Suppose, that $r_1(v_1, x, f)$ is strictly increasing in v_1 and define the inverse, $r_{1V}^{-1}(u; x, f)$. Let $Y_1^*(x) = r_{1V}^{-1}(U_1; x, F)$. Then we also have that

[16] This would require also that the function r_2 is continuous in v_2 at v_{20}.

$Y_1 = \mathbf{1}(Y_1^*(X) \leq V_1)$ and

$$F_{Y_1^*(x)|F}(y^* \mid f) = F_{U_1}(r_1(y^*, x, f)) \tag{5.9}$$

so that the conditional density is $f_{Y_1^*(x)|F}(y^* \mid f) = f_{U_1}(r_1(y^*, x, f))\frac{\partial}{\partial y^*} r_1(y^*, x, f)$. Then we have the following result, from essentially the same argument as the previous proposition.

Proposition 5.2. *If F is independent of V_1 conditional on X, $r_1(v_1, x_1, f)$ is strictly increasing in v_1, and support$(Y_1^*(x)) \subseteq \mathcal{V}_1$ then condition (i) of Assumption 3.7 is implied by the condition*

$$E(|\psi^*(F)|) < \infty \text{ and } E(\psi^*(F) \mid Y_1^{*V}(x) = v_1) = 0 \text{ for all } v_1 \in \text{support}(Y_1^*(x)) \Rightarrow \psi^*(f) \equiv 0 \tag{5.10}$$

6 Estimation

Given an i.i.d. sample $\{Y_i, W_i\}_{i=1}^n$ from a distribution satisfying Assumption 3.1, a natural approach to estimation is based on the implied likelihood function,

$$\ell_i(\theta) := \int \prod_{t=2}^{T} p_t(Y_{it} \mid Y_{i(t-1)}, W_{it}, f; \theta) p_1(Y_{i1} \mid V_{i1}, X_i, f; \theta) f_{F|V_1,X}(f \mid V_{i1}, X_i; \theta) df \tag{6.1}$$

A sieve maximum likelihood estimator solves

$$\max_{\theta \in \Theta_n} \sum_{i=1}^{n} \log(\ell_i(\theta)) \tag{6.2}$$

where Θ_n is a sequence of finite-dimensional sieve spaces that approximates the parameter space Θ. For a fully nonparametric estimator, $\theta = (p_1, p_2, \ldots, p_T, f_{F|V_1,X})$, and Θ is the infinite-dimensional space of functions satisfying Assumptions 3.3-3.7. If the parameter space Θ is restricted to be finite-dimensional in some dimensions by imposing functional form assumptions then the solution to (6.2) is a semiparametric sieve MLE. Consistency and asymptotic normality follow under conditions given in Shen et al. (1997), Chen and Shen (1998), Ai and Chen (2003), and Bierens (2014). If the parameter space Θ is restricted to be finite-dimensional in all dimensions and $\Theta_n = \Theta$ then the solution to (6.2) is the standard correlated random effects MLE.

6.1 A semiparametric estimator for the binary choice model

Consider the model of equation (5.4). Suppose that $F_{U_t} = F_U$ is a known distribution function, such as the probit or logit function. Further, suppose that $r_t(Y_{t-1}, V_t, X_t, F) = r(Y_{t-1}, V_t, X_t, F; \beta_t)$ for $t \geq 2$ where the function r is known given the vector of parameters β_t. Then $\theta = (\beta, p_1, f_{F|V_1,X})$ where $\beta = (\beta_1', \ldots, \beta_T')'$ and the functions p_1 and $f_{F|V_1,X}$ are to be estimated nonparametrically. The likelihood function takes the form

$$\ell_i(\theta) := \int \prod_{t=2}^{T} p_t(Y_{it} \mid Y_{i(t-1)}, W_{it}, f; \beta_t) p_1(Y_{i1} \mid V_{i1}, X_i, f) f_{F|V_1,X}(f \mid V_{i1}, X_i) df \tag{6.3}$$

where $p_t(Y_{it} \mid Y_{i(t-1)}, W_{it}, f; \beta_t) = F_U(r_t(Y_{i(t-1)}, V_{it}, X_{it}, f; \beta_t))^{Y_{it}} \left(1 - F_U(r_t(Y_{i(t-1)}, V_{it}, X_{it}, f; \beta_t))\right)^{1-Y_{it}}$. Then defining Θ_n involves specifying a function space for p_1 that is restricted to functions bounded between 0 and 1 and a function space for $f_{F|V_1,X}$ that is restricted to bounded, positive functions that integrate to 1.

6.2 Monte Carlo simulations

In this section, I present the results of a Monte Carlo study of a semiparametric sieve MLE in a binary choice model. I generate an initial draw, $Y_{i,-4} \sim Bernoulli(1/2)$. Then for $t = -3, \ldots, 3$, I generate $Y_{i,t}$ according to

$$Y_{i,t} = \mathbf{1}(\delta_t + \gamma Y_{i,t-1} + \beta V_{i,t} + \alpha_t F_i \geq U_{i,t}) \tag{6.4}$$

I use time-invariant coefficients, $\delta_t = 0$ and $\alpha_t = 1$. The errors, $U_{i,-3}, \ldots, U_{i,3}$ are generated independently from the standard normal distribution and $V_{i,t} = V_{i,t-1} + \eta_{i,t}$ with $V_{i,-4}$ and $\eta_{i,-3}, \ldots, \eta_{i,3}$ also generated independently from the standard normal distribution. Finally, F is generated independently from either a $N(0,1)$ distribution (model 1) or a mixture of $N(-2, 0.04)$ and $N(2, 0.04)$ with mixing probability $1/2$ (model 2).

I discard data from before period $t = 1$ so that I only use 3 periods of data in the estimation. I implement two estimators. Both estimators are in the class of maximum likelihood estimators defined by equations (6.1) and (6.2). The first is a fully parametric random effects MLE with $p_t(1 \mid Y_{i,t-1}, V_{it}, f; \theta) = \Phi(\gamma Y_{i,t-1} + \beta V_{i,t} + F_i)$, $p_1(1 \mid V_{i1}, f; \theta) = \Phi(\delta_1 + \beta_1 V_{i1} + \alpha_1 F_i)$, and

$f_{F|V_{i1}}(f \mid v_1) = f_F(f) = \frac{1}{\sigma_F}\phi\left(\frac{f-\mu_F}{\sigma_F}\right)$. In model 1, the initial condition distribution is misspecified but not the individual effects distribution. In model 2, both distributions are misspecified. The second estimator is a semiparametric sieve MLE that imposes the same parametric form for the conditional choice probabilities p_2 and p_3 but treats both p_1 and f_F nonparametrically, though I maintain the assumption that $f_{F|V_{i1}}(f \mid v_1) = f_F(f)$. I use the sieve space of Hermite polynomials of degree J_n for f_F and the artificial neural network sieve space with logistic activation function and degree K_n. The estimators implemented for the interactive fixed effects model are the same except that $p_3(1 \mid Y_{i,t-1}, V_{it}, f; \theta) = \Phi(\delta_t + \gamma Y_{i,t-1} + \beta V_{i,t} + \alpha_t F_i)$. I use period $t = 2$ to normalize: $\delta_2 = 0$ and $\alpha_2 = 1$.

Table 1. Monte Carlo summary

		random effects MLE			semiparametric sieve MLE		
		Bias	Std. Dev.	MSE	Bias	Std. Dev.	MSE
model 1	γ	-0.010	0.125	0.016	-0.009	0.132	0.018
	β	0.009	0.052	0.003	0.007	0.061	0.004
	ATE	-0.003	0.050	0.003	-0.001	0.054	0.003
model 2	γ	-0.642	0.113	0.424	0.006	0.130	0.017
	β	0.047	0.035	0.003	0.037	0.048	0.004
	ATE	-0.055	0.009	0.003	0.006	0.029	0.001

Notes: These results were obtained from a Monte Carlo simulation with 250 iterations. In each iteration I used a sample size of n=2000. The true value of the ATE is 0.2576 in model 1 and 0.0801 in model 2.

Table 1 provides the bias, standard deviation, and MSE of coefficients in each model, as well as the average treatment effect at the mean value of $V_3 = 0$,

$$ATE = \int_{\mathcal{F}} \left(p_3(1 \mid 1, 0, f) - p_3(1 \mid 0, 0, f)\right) f_F(f) df.$$

The results in this table were obtained using $J_n = 4$ and $K_n = 2$ for the semiparametric sieve MLE. In the first model, both estimators have minimal bias while the sieve ML estimator is slightly less efficient. The good performance of the parametric random effects model in this case is consistent with simulation results in Arellano and Bonhomme (2009). In the second model, the parametric

random effects estimator is severely biased while the semiparametric sieve ML estimator exhibits very minimal bias. The degree of state dependence is severely underestimated in the normal random effects model.

7 Conclusion

This paper provides an important new identification result for nonparametric discrete choice panel models with unobserved individual effects and lagged dependent variables. The argument extends the special regressor method (Lewbel, 1998) to a nonparametric model. Contrary to other recent work, no proxy variable is required. While the paper shows that the special regressor in the first period can be used as a proxy, it is also shown that identification is possible even when this variable fails as a proxy. A key novel idea in the paper is that the latent threshold underlying the discrete outcomes, can be used implicitly as a proxy for the unobserved heterogeneity, even when all of the covariates are independent of the unobserved individual effect. The identification argument justifies the use of a semiparametric sieve MLE that generalizes the standard random effects probit or logit estimator.

References

Aguirregabiria, V., J. Gu, and Y. Luo (2018): "Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models," working paper, University of Toronto.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2001): "GMM estimation of linear panel data models with time-varying individual effects," *Journal of Econometrics*, 101, 219–255.

——— (2013): "Panel data models with multiple time-varying individual effects," *Journal of Econometrics*, 174, 1–14.

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

Alessie, R., S. Hochguertel, and A. v. Soest (2004): "Ownership of Stocks and Mutual Funds: A Panel Data Analysis," *The Review of Economics and Statistics*, 86, 783–796.

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

Anderson, T. W. and C. Hsiao (1982): "Formulation and estimation of dynamic models using panel data," *Journal of Econometrics*, 18, 47–82.

Ando, T. and J. Bai (2018): "Large scale panel choice models with unobserved heterogeneity: a Bayesian data augmentation approach," working paper.

Andrews, D. W. (2017): "Examples of L2-complete and boundedly-complete distributions," *Journal of Econometrics*, 199, 213–220.

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, 277–297.

Arellano, M. and S. Bonhomme (2009): "Robust priors in nonlinear panel data models," *Econometrica*, 77, 489–536.

Arellano, M. and O. Bover (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 68, 29–51.

Arellano, M. and R. Carrasco (2003): "Binary choice panel data models with predetermined variables," *Journal of Econometrics*, 115, 125–157.

Bierens, H. J. (2014): "Consistency and asymptotic normality of sieve ML estimators under low-level conditions," *Econometric Theory*, 30, 1021–1076.

Boneva, L. and O. Linton (2017): "A discrete-choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance," *Journal of Applied Econometrics*, 32, 1226–1243.

Bonhomme, S. (2012): "Functional differencing," *Econometrica*, 80, 1337–1385.

Browning, M. and J. Carro (2007): "Heterogeneity and microeconometrics modeling," in *Advances in Economics and Econometrics*, ed. by R. Blundell, W. Newey, and T. Persson, Cambridge University Press, vol. 3.

Browning, M. and J. M. Carro (2014): "Dynamic binary outcome models with maximal heterogeneity," *Journal of Econometrics*, 178, 805–823.

Cameron, S. V. and J. J. Heckman (1998): "Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males," *Journal of Political Economy*, 106, 262–333.

——— (2001): "The dynamics of educational attainment for black, hispanic, and white males," *Journal of Political Economy*, 109, 455–499.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the testability of identification in some nonparametric models with endogeneity," *Econometrica*, 81, 2535–2559.

Carroll, R. J., X. Chen, and Y. Hu (2010): "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors," *Journal of Nonparametric Statistics*, 22, 379–399.

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

Chen, M., I. Fernandez-Val, and M. Weidner (2018a): "Nonlinear Factor Models for Network and Panel Data," Working Paper CWP38/18, cemmap.

Chen, S., S. Khan, and X. Tang (2018b): "Exclusion Restrictions in Dynamic Binary Choice Panel Data Models," Working paper, Boston College Department of Economics.

Chen, X. and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 289–314.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

Chintagunta, P., E. Kyriazidou, and J. Perktold (2001): "Panel data analysis of household brand choices," *Journal of Econometrics*, 103, 111–153.

Contoyannis, P., A. M. Jones, and N. Rice (2004): "Simulation-based inference in dynamic panel probit models: an application to health," *Empirical Economics*, 29, 49–77.

D'Haultfoeuille, X. (2011): "On the completeness condition in nonparametric instrumental problems," *Econometric Theory*, 27, 460–471.

Dunford, N. and J. T. Schwartz (1971): *Linear Operators*, New York: Wiley.

Freyberger, J. (2018): "Non-parametric Panel Data Models with Interactive Fixed Effects," *The Review of Economic Studies*, 85, 1824–1851.

Gayle, W.-R. and S. D. Namoro (2013): "Estimation of a nonlinear panel data model with semiparametric individual effects," *Journal of Econometrics*, 175, 46–59.

Hausman, J. A. and W. E. Taylor (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377–1398.

Heckman, J. J. (1981a): "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process." in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, MIT Press.

——— (1981b): "Statistical models for discrete panel data," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, MIT Press.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2016): "Dynamic treatment effects," *Journal of Econometrics*, 191, 276–292.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988): "Estimating vector autoregressions with panel data," *Econometrica*, 1371–1395.

Honore, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Honore, B. E. and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063.

Honore, B. E. and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144, 27–61.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y. and J.-L. Shiu (2018): "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness," *Econometric Theory*, 34, 659–693.

Hu, Y. and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Hyslop, D. R. (1999): "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women," *Econometrica*, 67, 1255–1294.

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

Kerr, W. R., W. F. Lincoln, and P. Mishra (2014): "The Dynamics of Firm Lobbying," *American Economic Journal: Economic Policy*, 6, 343–79.

Khan, S., F. Ouyang, and E. Tamer (2016): "Adaptive Rank Inference in Semiparametric Multinomial Response Models," working paper.

Lewbel, A. (1998): "Semiparametric latent variable model estimation with endogenous or mismeasured regressors," *Econometrica*, 105–121.

Lewbel, A., D. McFadden, and O. Linton (2011): "Estimating features of a distribution from binomial data," *Journal of Econometrics*, 162, 170–188.

Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–62.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica: journal of the Econometric Society*, 69–85.

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

Neyman, J. and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, 1–32.

Pakes, A. and J. Porter (2016): "Moment inequalities for multinomial choice with fixed effects," Tech. rep., National Bureau of Economic Research.

Rasch, G. (1961): "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, University of California Press Berkeley, CA, vol. 4, 321–333.

Roberts, M. J. and J. R. Tybout (1997): "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," *The American Economic Review*, 87, 545–564.

Sasaki, Y. (2015): "Heterogeneity and selection in dynamic panel data," *Journal of Econometrics*, 188, 236–249.

Shen, X. et al. (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 25, 2555–2591.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Shiu, J.-L. and Y. Hu (2013): "Identification and estimation of nonlinear dynamic panel data models with unobserved covariates," *Journal of Econometrics*, 175, 116–131.

Wooldridge, J. (2005): "Unobserved heterogeneity and estimation of average partial effects," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. K. Andrews and J. H. Stock, Cambridge: Cambridge University Press, 27–55.

A Appendix

A.1 Assumption 3.1

Consider the follow set of conditions.

Assumption A.1.

(i) *Y is drawn from a first order Markov process conditional on (W, F). In other words,*

$$Pr(Y = y \mid W, F) = \prod_{t=2}^{T} Pr(Y_t = y_t \mid Y_{t-1}, W, F) Pr(Y_1 = y_1 \mid W, F)$$

(ii) *for $t \geq 2$, $Pr(Y_t = y_t \mid Y_{t-1}, W, F) = Pr(Y_t = y_t \mid Y_{t-1}, W_t, F)$*

(iii) $(V_2, \ldots, V_T) \perp\!\!\!\perp (Y_1, F) \mid V_1, X$

Condition (i) allows for dynamics in the form of lagged dependent variables. Condition (ii) imposes a common type of limited feed-back whereby innovations in period t cannot affect future values of the covariates. Condition (iii) defines V as the special regressor. In the discussion in Section 3 following the statement of Assumption 3.1, conditions (i) and (ii) are taken as given and the discussion centers around condition (iii). As is common in the panel data literature, the conditions in Assumption A.1 are imposed on the distribution of the dependent variables conditional on (W, F). However, Assumption 3.1 can also be derived from a first order Markov assumption on (Y, W) conditional on F. The following set of conditions is an alternative to Assumption A.1.

Assumption A.2.

(i) *For $t \geq 2$, $f_{Y_t|W^{(t)},Y^{(t-1)},F} = f_{Y_t|W_t,Y_{t-1},F}$*

(ii) $f_{X_t|V^{(t)},Y^{(t-1)},X^{(t-1)},F} = f_{X_t|V^{(t)},X^{(t-1)},F} = f_{X_t|X^{(t-1)},F}$,

(iii) $f_{V_t|V^{(t-1)},Y^{(t-1)},X^{(t-1)},F} = f_{V_t|V^{(t-1)},X^{(t-1)},F} = f_{V_t|V^{(t-1)}}$, *and*

(iv) $f_{Y_1|V_1,X,F} = f_{Y_1|V_1,X_1,F}$.

Assumption A.2 is more in line with the structure of the models of Kasahara and Shimotsu (2009) while Assumption A.1 is similar to the model of Browning and Carro (2014).

Proposition A.1. *Assumption 3.1 holds under either Assumption A.1 or Assumption A.2.*

Proof. First, under Assumption A.1,

$$\begin{aligned}
Pr(Y \mid W) &= \int \prod_{t=2}^{T} p_t(Y_t \mid Y_{t-1}, W, F) Pr(Y_1 = y_1 \mid W, F) f_{F|W} dF \\
&= \int \prod_{t=2}^{T} p_t(Y_t \mid Y_{t-1}, W_t, F) Pr(Y_1 = y_1 \mid W, F) f_{F|W} dF \\
&= \int \prod_{t=2}^{T} p_t(Y_t \mid Y_{t-1}, W_t, F) f_{Y_1,F|W} dF \\
&= \int \prod_{t=2}^{T} p_t(Y_t \mid Y_{t-1}, W_t, F) f_{Y_1,F|V_1,X} dF
\end{aligned}$$

Next, under Assumption A.2,

$$\begin{aligned}
f_{Y,W} &= \int f_{Y,W|F} f_F dF \\
&= \int \prod_{t=1}^{T} f_{Y_t|W^{(t)},Y^{(t-1)},F} f_{X_t|V^{(t)},Y^{(t-1)},X^{(t-1)},F} f_{V_t|V^{(t-1)},Y^{(t-1)},X^{(t-1)},F} f_F dF \\
&= \int \prod_{t=2}^{T} f_{Y_t|W_t,Y_{t-1},F} f_{X_t|X^{(t-1)},F} f_{V_t|V^{(t-1)}} f_{Y_1|W_1,F} f_{X_1|V_1,F} f_{V_1|F} f_F dF
\end{aligned}$$

and

$$\begin{aligned}
f_W &= \int f_{W|F} f_F dF \\
&= \int \prod_{t=1}^{T} f_{X_t|V^{(t)},X^{(t-1)},F} f_{V_t|V^{(t-1)},X^{(t-1)},F} f_F dF \\
&= \int \prod_{t=2}^{T} f_{X_t|X^{(t-1)},F} f_{V_t|V^{(t-1)}} f_{X_1|V_1,F} f_{V_1|F} f_F dF
\end{aligned}$$

Therefore,

$$
\begin{aligned}
Pr(Y \mid W) &= \frac{f_{Y,W}}{f_W} \\
&= \frac{\int \prod_{t=2}^{T} f_{Y_t|W_t,Y_{t-1},F} f_{X_t|X^{(t-1)},F} f_{V_t|V^{(t-1)}} f_{Y_1|W_1,F} f_{X_1|V_1,F} f_{V_1|F} f_F dF}{\int \prod_{t=2}^{T} f_{X_t|X^{(t-1)},F} f_{V_t|V^{(t-1)}} f_{X_1|V_1,F} f_{V_1|F} f_F dF} \\
&= \int \prod_{t=2}^{T} f_{Y_t|W_t,Y_{t-1},F} \frac{\prod_{t=2}^{T} f_{X_t|X^{(t-1)},F} f_{Y_1|W_1,F} f_{X_1|V_1,F} f_{V_1|F} f_F}{\int \prod_{t=2}^{T} f_{X_t|X^{(t-1)},F} f_{X_1|V_1,F} f_{V_1|F} f_F dF} dF \\
&= \int \prod_{t=2}^{T} f_{Y_t|W_t,Y_{t-1},F} \frac{f_{Y_1,X,V_1,F}}{f_{X,V_1}} dF \\
&= \int \prod_{t=2}^{T} f_{Y_t|W_t,Y_{t-1},F} f_{Y_1,F|X,V_1} dF
\end{aligned}
$$

□

Lastly, Proposition 3.1 provides sufficient conditions for condition (iii) in Assumption A.1. The proof of Proposition 3.1 follows.

Proof of Proposition 3.1. Let $\tilde{V}_t = (V_t, X_{t1}')'$ and $\tilde{X}_t = X_{t2}$. Then the density of $\tilde{V}_2, \ldots, \tilde{V}_T \mid \tilde{V}_1, \tilde{X}, F$ satisfies

$$
\begin{aligned}
f_{\tilde{V}_2,\ldots,\tilde{V}_T|\tilde{V}_1,\tilde{X},F} &= \frac{f_{\tilde{V},\tilde{X}|F}}{f_{\tilde{V}_1,\tilde{X}|F}} \\
&= \frac{\prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)},F} f_{\tilde{X}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t-1)},F} f_{\tilde{V}_1,\tilde{X}_1|F}}{\int \prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)},F} f_{\tilde{X}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t-1)},F} f_{\tilde{V}_1,\tilde{X}_1|F} d\tilde{V}_2 \ldots d\tilde{V}_T} \\
&= \frac{\prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)}} f_{\tilde{X}_t|\tilde{X}^{(t-1)},F} f_{\tilde{V}_1,\tilde{X}_1|F}}{\int \prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)}} f_{\tilde{X}_t|\tilde{X}^{(t-1)},F} f_{\tilde{V}_1,\tilde{X}_1|F} d\tilde{V}_2 \ldots d\tilde{V}_T} \\
&= \frac{\prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)}}}{\int \prod_{t=2}^{T} f_{\tilde{V}_t|\tilde{V}^{(t-1)},\tilde{X}^{(t)}} d\tilde{V}_2 \ldots d\tilde{V}_T},
\end{aligned}
\tag{A.1}
$$

where the third equality follows from conditions (i) and (ii) of the proposition. Since the final line doesn't depend on F, $f_{\tilde{V}_2,\ldots,\tilde{V}_T|\tilde{V}_1,\tilde{X},F} = f_{\tilde{V}_2,\ldots,\tilde{V}_T|\tilde{V}_1,\tilde{X}}$, which is the desired result. □

A.2 Theorem 4.1

First we state a lemma that extends some results in Hu and Schennach (2008).

Lemma A.1. *Suppose that A_1, A_2, B, C, and D are bounded linear operators such that $A_1 = BC$ and $A_2 = BDC$ where $C : \mathcal{L}^1(\mathcal{C}) \to \mathcal{L}^1(\mathcal{B})$, D is a diagonal operator (multiplication by the function δ), and $B : \mathcal{L}^1(\mathcal{B}) \to \mathcal{L}^\infty(\mathcal{A})$ for $\mathcal{A} \subseteq \mathbb{R}$, $\mathcal{B} \subseteq \mathbb{R}$, and $\mathcal{C} \subseteq \mathbb{R}$. Suppose that B and the adjoint of C, denoted C^*, are injective. Then A_1 has a right inverse, A_1^{-1}, and B has a left inverse, B^{-1}, and the operator equivalence $A_2 A_1^{-1} = BDB^{-1}$ holds over the range of B. Moreover, $P(M) = BI_M B^{-1}$ where for any $M \in \mathbb{R}$ $[I_M g](b) = \mathbf{1}(\delta(b) \in M)g(b)$ is a projection-valued measure supported on the spectrum $\sigma = \{\delta(b) : b \in \mathcal{B}\}$. Finally, $A_2 A_1^{-1}$ admits the unique decomposition $BDB^{-1} = \int_\sigma \mu P(d\mu)$.*

The proof of this lemma is given below. The proof uses many of the same arguments as Hu and Schennach (2008). The main difference between this result and what is shown in that paper is to clarify that only the adjoint of C needs to be injective, bu not the adjoint of A_1. Note that when C is an integral operator, $[Cg](b) = \int_{\mathcal{C}} k(b,c)g(c)dc$, the adjoint is the operator defined on the dual space $\mathcal{L}^1(\mathcal{B})^* = \mathcal{L}^\infty(\mathcal{B})$ given by $[C^*g](c) = \int_{\mathcal{B}} k(b,c)g(b)db$. Thus, injectivity of C^* requires that if $\int_{\mathcal{B}} k(b,c)g(b)db = 0$ for almost all $c \in \mathcal{C}$ for a function $g \in \mathcal{L}^\infty(\mathcal{B})$ then $g(b) = 0$ for almost all $b \in \mathcal{B}$.

We now present the proof of Theorem 4.1.

Proof of Theorem 4.1. First, as shown in Section 3, for any values of y, v_3, and x we have

$$[L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3} g](v_2) = \Lambda_{y_2;y_1,V_2,x_2,F} \Lambda_{y_1;V_1,x,F}$$
$$[L_{y;V_1,x_1,V_2,x_2,v_3,x_3} g] = \Lambda_{y_2;y_1,V_2,x_2,F} \Delta_{y_3;y_2,v_3,x_3,F} \Lambda_{y_1;V_1,x,F}$$

Consider $y_1 = y_{11}$, $y_2 = y_{21}$, any y_3, v_3, and x_3, and x_1 and $x_2 = x_{20}$ (provided by Assumption 3.6). By Assumption 3.7, $\Lambda_{y_{21};y_{11},V_2,x_{20},F}$ and $\Lambda^*_{y_{11};V_1,x,F}$ are both injective and therefore we can apply Lemma A.1.

As noted by Hu and Schennach (2008), in the spectral decomposition given by Lemma A.1 the spectrum consists of the values of the function $p_3(y_3 \mid y_{21}, v_3, x_3, f)$ as f varies in $\mathcal{F}$ and $P(M)$ can be defined via the subspace $\mathcal{S}(M) := span\{p_2(y_{21} \mid y_{11}, \cdot, x_{20}, f) : f \text{ such that } p_3(y_3 \mid y_{21}, v_3, x_3, f) \in M\}$. Therefore, the decomposition is unique up to (i) scaling of the eigenfunctions, $p_2(y_{21} \mid y_{11}, \cdot, x_{20}, f)$, (ii) possible multiplicity of eigenvalues, $p_3(y_3 \mid y_{21}, v_3, x_3^*, f)$, and (iii)

reordering of the eigenvalues and associated eigenvectors.

First, the scale of the eigenfunctions is fixed by Assumption 3.5 because these functions must be equal to ℓ at $v_2 = \bar{v}_2$.[17] Second, since the eigenfunctions do not vary with y_3 or v_3, I can vary (y_3, v_3) sufficiently to separately identify each eigenfunction (up to reordering) despite the possibility of multiplicity of eigenvalues for a fixed (y_3, v_3) under Assumption 3.4. This is possible because the functions $p_2(y_{21} \mid y_{11}, \cdot, x_{20}, f)$ do not depend on y_3 or v_3. Third, each of the eigenfunctions must be equal to $\pi(f)$, which is a one-to-one function on $\mathcal{F}$, at v_{20}, and, hence, no reordering of the eigenvalues is possible.[18]

I have shown that $\Lambda_{y_{21};y_{11},V_2,x_{20},F}$ is identified and $\Delta_{y_3;y_{21},v_3,x_3,F}$ is identified for any y_3, v_3. And $\Lambda_{y_{11};V_1,x,F}$ is also identified because $\Lambda_{y_{11};V_1,x,F} = \Lambda^{-1}_{y_{21};y_{11},V_2,x_{20},F} L_{y_{11},y_{21};V_1,x_1,V_2,x_{20},v_3,x_3} g](v_2)$.

Next consider $y_2 = y_{21}$ with any value of x, y_1, y_3, and v_3. Again, by Assumption 3.7, $\Lambda_{y_{21};y_1,V_2,x_2,F}$ and $\Lambda^*_{y_1;V_1,x,F}$ are both injective and therefore we can apply Lemma A.1. In the unique decomposition provided by this lemma,

$$L_{y_1,y_{21},y_3;V_1,x_1,V_2,x_2,v_3,x_3} L^{-1}_{y_1,y_{21};V_1,x_1,V_2,x_2,v_3,x_3} = \Lambda_{y_{21};y_{11},V_2,x_{20},F} \Delta_{y_3;y_{21},v_3,x_3,F} \Lambda^{-1}_{y_{21};y_{11},V_2,x_{20},F}, \quad \text{(A.2)}$$

the eigenvalues have already been identified in the previous step. Then, as in Step 1, the scale of the eigenfunctions is fixed by Assumption 3.5 because these functions must be equal to ℓ at $\bar{v}_2$ regardless of the value of x_2. And again I can vary (y_3, v_3) sufficiently to separately identify each eigenfunction (up to reordering) despite the possibility of multiplicity of eigenvalues for a fixed (y_3, v_3) under Assumption 3.4. Lastly, reordering is not possible because $\Delta_{y_3;y_{21},v_3,x_3,F}$ is already identified. That is, suppose I could obtain observationally equivalent models by swapping the eigenfunctions corresponding to $f_1, f_2 \in \mathcal{F}$. By Assumption 3.4, I can find (y_3, v_3) such that the eigenvalues associated with f_1 and f_2 are distinct. Since $\Delta_{y_3;y_{21},v_3,x_3,F}$ is already identified, these two eigenvalues are identified and hence their corresponding eigenfunctions are as well, contradicting the observational equivalence. This again implies identification of $\Lambda_{y_1;V_1,x,F}$ as well.

Finally, consider any value of $y_2 \neq y_{21}$. We apply Lemma A.1 again. Since $\Lambda_{y_1;V_1,x,F}$ doesn't de-

[17] Formally, suppose p_2 is observationally equivalent to some p_2^*. Then $p_2(y_{21} \mid y_{11}, \cdot, x_{20}, f) = p_2^*(y_{21} \mid y_{11}, \cdot, x_{20}, f)s(f)$. Hence $\ell = \lim_{v_2 \to \bar{v}_2} p_2(y_{21} \mid y_{11}, \cdot, x_{20}, f) = \lim_{v_2 \to \bar{v}_2} p_2^*(y_{21} \mid y_{11}, \cdot, x_{20}, f)s(f) = \ell s(f)$. Since $\ell > 0$, this implies that $s(f) = 1$.

[18] That is, to identify the eigenfunction associated with a particular $f^* \in \mathcal{F}$, I look for the eigenfunction that is equal to $\pi(f^*)$ at $v_2 = v_{20}$.

pend on y_2, it has already been identified so the equation $[L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}g](v_2) = \Lambda_{y_2;y_1,V_2,x_2,F}\Lambda_{y_1;V_1,x,F}$ can be solved for $\Lambda_{y_2;y_1,V_2,x_2,F}$. Then, in the operator equivalence.

$$L_{y;V_1,x_1,V_2,x_2,v_3,x_3}L^{-1}_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3} = \Lambda_{y_2;y_1,V_2,x_2,F}\Delta_{y_3;y_2,v_3,x_3,F}\Lambda^{-1}_{y_2;y_1,V_2,x_2,F} \tag{A.3}$$

we can solve for which implies that

$$\Delta_{y_3;y_2,v_3,x_3,F} = \Lambda^{-1}_{y_2;y_1,V_2,x_2,F}L_{y;V_1,x_1,V_2,x_2,v_3,x_3}L^{-1}_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}\Lambda_{y_2;y_1,V_2,x_2,F} \tag{A.4}$$

so $\Delta_{y_3;y_2,v_3,x_3,F}$ is identified.

Last, identification of Λ_2 and Δ imply identification of p_2 and p_3, respectively. Identification of Λ_1 implies identification of $f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x)$. Hence $f_{F|V_1,X}(f \mid v_1, x) = \sum_{y_1\in\mathcal{Y}_1} f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x)$ is identified too. Then p_1 is identified because $f_{Y_1,F|V_1,X}(y_1, f \mid v_1, x) = p_1(y_1 \mid f, v_1, x)f_{F|V_1,X}(f \mid v_1, x)$.

□

Proof of Lemma A.1. First, since B is injective it has a left inverse so that $B^{-1}A_1 = C$. Therefore, $A_2 = BDB^{-1}A_1$.

Next, by Lemma 1 in Hu and Schennach (2008), C^{-1} exists and is densely defined over $\mathcal{L}^1(\mathcal{B})$. Moreover, it can be extended to a bounded linear operator defined over $\mathcal{L}^1(\mathcal{B})$, which is the domain of B. Since B is injective, B^{-1} exists and is defined over the range of B. Therefore, $A_1^{-1} = C^{-1}B^{-1}$ defines a right inverse of A_1 over the range of B. Therefore, $A_2A_1^{-1} = BDB^{-1}$ defines an operator equivalence over the range of B. Further, the operator equivalence can be extended to the closure of the range of B, $r(\bar{B})$.

Next, $P(M) = BI_MB^{-1}$ is projection-valued because B^{-1} is defined on the range of B and therefore, for any M, $P(M)P(M) = BI_MB^{-1}\left(BI_MB^{-1}\right) = BI_MB^{-1} = P(M)$. Note that if C^{-1} were not densely defined over $\mathcal{L}^1(\mathcal{B})$, that is, if the range of C were not dense in $\mathcal{L}^1(\mathcal{B})$, then $A_2A_1^{-1} = BDB^{-1}$ would hold only over the range of A_1, which is not dense in the range of B, and therefore, $P(M)$ would not be projection-valued when restricted to the domain over which this equivalence holds. This is why it is crucial that C^* is injective. Lastly, $BDB^{-1} = \int_\sigma \mu P(d\mu)$, following the same argument as in the proof of Theorem 1 in Hu and Schennach (2008).

Since $r(\bar{B})$ is a closed linear subspace of a Banach space, it is also a Banach space. Since the spectrum σ is bounded, BDB^{-1} is a bounded linear operator. Therefore, Theorem XV.4.5 in Dunford and Schwartz (1971) can be applied to conclude that the decomposition is unique.

□

A.3 The discrete case

Suppose that $\mathcal{V}_t = \{v_{t,1}, \ldots, v_{t.K_t}\}$ for each t and $\mathcal{F} = \{1, \ldots, L\}$. Then we could define two $K_2 \times K_1$ matrices, $(p(y_1, y_2 \mid v_{1k}, x_1, v_{2j}, x_2, v_3, x_3))_{j=1,\ldots,K_2, k=1,\ldots,K_1}$ and $(p(y \mid v_{1k}, x_1, v_{2j}, x_2, v_3, x_3))_{j=1,\ldots,K_2, k=1,\ldots,K_1}$. Suppose that $K_t \geq L$ for $t = 1, 2$ and let $L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}$ and $L_{y;V_1,x_1,V_2,x_2,v_3,x_3}$ represent any $J \times J$ submatrices.

Then by Assumption 3.7,

$$\begin{aligned} & p(y \mid v_{1k}, x_1, v_{2j}, x_2, v_3, x_3) \\ & = \sum_{l=1}^{L} p_3(y_3 \mid y_2, v_3, x_3, l) p_2(y_2 \mid y_1, v_{2j}, x_2, l) p_1(y_1 \mid v_{1k}, x, l) Pr(F = l \mid V_1 = v_{1k}, X = x) \end{aligned}$$

And, therefore,

$$L_{y;V_1,x_1,V_2,x_2,v_3,x_3} = \Lambda_{y_2;y_1,V_2,x_2,F} \Delta_{y_3;y_2,v_3,x_3,F} \Lambda_{y_1;V_1,x,F}$$

where $\Lambda_{y_1;V_1,x,F}$ is the $L \times L$ submatrix of the $L \times K_1$ matrix $(p_1(y_1 \mid v_{1k}, x, l) Pr(F = l \mid V_1 = v_{1k}, X = x))_{l=1,\ldots,L, k=1,\ldots,K_1}$, $\Lambda_{y_2;y_1,V_2,x_2,F}$ is the $L \times L$ submatrix of the $K_2 \times L$ matrix $(p_2(y_2 \mid y_1, v_{2j}, x_2, l))_{j=1,\ldots,K_2, l=1,\ldots,L}$, and $\Delta_{y_3;y_2,v_3,x_3,F} = diag(p_3(y_3 \mid y_2, v_3, x_3, l), l = 1, \ldots, L)$. We can then obtain the following matrix equation by summing the above equation over $y_3 \in \{0, 1\}$.

$$L_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3} = \Lambda_{y_2;y_1,V_2,x_2,F} \Lambda_{y_1;V_1,x,F} \tag{A.5}$$

If the matrices $\Lambda_{y_2;y_1,V_2,x_2,F}$ and $\Lambda_{y_1;V_1,x,F}$ are nonsingular then we obtain the matrix equation

$$L_{y;V_1,x_1,V_2,x_2,v_3,x_3} L^{-1}_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3} = \Lambda_{y_2;y_1,V_2,x_2,F} \Delta_{y_3;y_2,v_3,x_3,F} \Lambda^{-1}_{y_2;y_1,V_2,x_2,F} \tag{A.6}$$

The role of Assumptions 3.4, 3.5, and 3.6 may be more clear in the discrete case. The above equa-

tion represents an eigenvalue decomposition of the observed matrix $L_{y;V_1,x_1,V_2,x_2,v_3,x_3} L^{-1}_{y_1,y_2;V_1,x_1,V_2,x_2,v_3,x_3}$. The eigenvalues identify $p_3(y_3 \mid y_2, v_3, x_3, l), l = 1, \ldots, L$. Assumptions 3.4 ensures that these L eigenvalues are distinct, at least for $y_2 = 1$ but the order of the eigenvalues is not uniquely determined. Thus, each eigenvector is equal to the vector $s(l)p_2(y_2 \mid y_1, v_{2j}, x_2, l), k = \ldots$ for some l. In other words, the matrix $\Lambda_{y_2;y_1,V_2,x_2,F}$ is identified up to multiplication on the right by PQ where P is a diagonal matrix and Q is a matrix that changes the order of the columns. Equation (A.5) then implies that $\Lambda_{y_1;V_1,x,F}$ is identified up to multiplication on the left by QP^{-1}. However, note that $\iota'\Lambda_{y_1;V_1,x,F} = L'_{y_1;V_1,x}$, where ι is a vector of ones. Therefore, $\iota'\Lambda_{y_1;V_1,x,F} = \iota'QP^{-1}\Lambda_{y_1;V_1,x,F}$. Since $\iota'Q = \iota'$ and $\Lambda_{y_1;V_1,x,F}$ is nonsingular, this implies that $P = I$. Therefore, it is clear that in the discrete case 3.5 is unnecessary.

Now, in place of Assumption 3.6, suppose that for some j and some x_{20}, $p_2(1 \mid 0, v_{2j}, x_{20}, 1) < p_2(1 \mid 0, v_{2j}, x_{20}, 2) < \ldots < p_2(1 \mid 0, v_{2j}, x_{20}, L)$.[19] Going back to the eigendecomposition in equation (A.6), taking $y_1 = 0, y_2 = 1$, the elements in the j^{th} row of $\Lambda_{1;0,V_2,x_{20},F}$ and $\Lambda_{1;0,V_2,x_{20},F}Q$ must both be in rank order, which implies that $Q = I$. Therefore, $\Lambda_{1;0,V_2,x_{20},F}$, $\Delta_{y_3;1,v_3,x_3,F}$, and $\Lambda_{0;V_1,x,F}$ are identified.

Because $\Lambda_{0;0,V_2,x_2,F} + \Lambda_{1;0,V_2,x_2,F} = \iota\iota'$ this also implies that $\Lambda_{0;0,V_2,x_2,F}$ is identified. Applying equation (A.6) for $y_1 = 0, y_2 = 0$, identification of $\Delta_{y_3;0,v_3,x_3,F}$ follows.

If we apply equation (A.6) for $x_2 \neq x_{20}$ and/or $y_1 \neq 0$, the order of the eigenvalues is already identified because $p_3(y_3 \mid y_2, v_3, x_3, l)$ does not vary with x_2 or y_1. Therefore, the eigenvectors are identified up to scale and the equation $\iota'\Lambda_{y_1;V_1,x,F} = L'_{y_1;V_1,x}$ can again be used to resolve the scale.

A.4 A static model

Consider the following assumption.

Assumption A.3. *For each $t = 1, \ldots, T$ there exists t' and t'' such that*

(i) $p(Y_t, Y_{t'}, Y_{t''} \mid W) = \int p_t(Y_t \mid V_t, X_1, F)p_{t'}(Y_{t'} \mid V_{t'}, X_2, F)p_{t''}(Y_{t''} \mid V_{t''}, X, F)f_{F|V_{t''},X}dF$,

(ii) $support(W_t, W_{t'}, W_{t''}) = \mathcal{V}_t \times \mathcal{V}_{t'} \times \mathcal{V}_{t''} \times support(X_t, X_{t'}, X_{t''})$,

(iii) the density $f_{F|V_{t''},X}$ is bounded,

[19] We can do this because we are assuming the points in the support, $\mathcal{F}$, are known to be the integers $1, \ldots, L$. If instead we maintain a normalization like that in Assumption 3.6 then these support points can be identified.

(iv) *for any* $x_t \in \mathcal{X}_t$, $Pr\left(\exists f^* \in \mathcal{F} \text{ s.t. } p_t(Y_t \mid V_t, x_t, f^*) = p_t(Y_t \mid V_t, x_t, f)\right) < 1$ *for almost all* $f \in \mathcal{F}$,

(v) *for each* $x_{t'} \in \mathcal{X}_{t'}$, *there exists a (known)* $\bar{v}_{t'} \in \mathbb{R} \cup \{-\infty, \infty\}$ *and a known* $0 < \ell \leq 1$ *such that* $\lim_{v_{t'} \to \bar{v}_{t'}} p_{t'}(y_{t'1} \mid v_{t'}, x_{t'}, f) = \ell$ *for all* $f \in \mathcal{F}$ *and if* $|\bar{v}_{t'}| < \infty$ *then* $\bar{v}_{t'} \in \mathcal{V}_{t'}$, *if* $\bar{v}_{t'} = \pm\infty$ *then* $\mathcal{V}_{t'}$ *is unbounded from above or below, respectively,*

(vi) *there exists* $w_{t'0} = (v_{t'0}, x_{t'0}) \in \mathcal{W}_{t'}$ *and a known one-to-one function,* $\pi : \mathbb{R} \to [0,1]$, *with* $\pi(\mathbb{R}) = [0,1]$ *such that* $\lim_{w_{t'} \to w_{t'0}} p_{t'}(y_{t'1} \mid w_{t'}, f) = \pi(f)$ *for all* $f \in \mathcal{F}$. $\mathcal{X} = Support(X_t, X_{t'}, X_{t''})$ *satisfies the condition that for each* $x_t \in \mathcal{X}_t$, *there exists* $x_{t''} \in \mathcal{X}_{t''}$ *such that* $(x_t, x_{t'0}, x_{t''}) \in \mathcal{X}$.

(vii) *for each* $y_{t''} \in \mathcal{Y}_{t''}$ *and* $x \in \mathcal{X}$, *if* $\psi \in \mathcal{L}^\infty(\mathcal{F})$ *and* $\int_{\mathcal{F}} f_{Y_{t''}, F|V_{t''}, X}(y_{t''}, f \mid v_{t''}, x)\psi(f)df = 0$ *for all* $v_{t''} \in \mathcal{V}_{t''}$ *then* $\psi \equiv 0$.

(viii) *For each* $y_{t'} \in \mathcal{Y}_{t'}$ *and each* $x_{t'} \in \mathcal{X}_{t'}$, *if* $\psi \in \mathcal{L}^1(\mathcal{F})$ *and* $\int_{\mathcal{F}} p_{t'}(y_{t'} \mid v_{t'}, x_{t'}, f)\psi(f)df = 0$ *for all* $v_{t'} \in \mathcal{V}_{t'}$ *then* $\psi \equiv 0$.

These conditions are very similar to the assumptions of the dynamic model in the paper. The main advantage is that in the static model with $T > 3$, the conditional independence assumption (condition (i)) may hold for triples other than $(t, t', t'') = (1, 2, 3)$. Then X_t does not need to be excluded from the conditional probability p_τ for all $\tau \neq t$, though the same is not true for V_t. One apparent difficulty is in applying the normalization (condition (vi)). Unless the conditional probabilities p_τ are stationary (i.e., do not depend on τ), it will typically only be plausible to assume the normalization for a particular t', say $t' = 1$. Then, in order to identify p_t for a given t, the components of X that affect Y_t (conditional on F, V_t) must be distinct from those that affect Y_1 (conditional on F, V_1).

Theorem A.1. *Under Assumption A.3, the choice probabilities,* $p_t, p_{t'}$, *and* $p_{t''}$, *are identified as is the distribution* $f_{F|V_{t''}, X}$.

Proof. The proof is identical to that of Theorem 4.1 with $t = 3$, $t' = 2$ and $t'' = 1$. □

This result follows the same proof as Theorem 4.1, starting with forming operator equivalence (4.6) where Y_1, Y_2, Y_3 are replaced by $Y_{t''}, Y_{t'}$, and Y_t, respectively. The identification argu-

ment could be modified when $T > 3$ in order to show identification of the static model under weaker conditions than those of Assumption A.3. This could be done, for example, by deriving operator equivalence (4.6) for other triples $(Y_{t_1}, Y_{t_2}, Y_{t_3})$ where $\{t_1, t_2, t_3\} \cap \{t, t', t''\} \neq \emptyset$ in order to get more information about $p_t, p_{t'}$, and $p_{t''}$. A full development of this model is beyond the scope of this paper.

A.5 More lags

To demonstrate the possibility of identification with more than one lagged dependent variable I provide here an example where there are two lags and $T = 5$. Suppose that $Pr(Y_t \mid Y^{(t-1)}, W, F) = Pr(Y_t \mid Y_{t-1}, Y_{t-2}, W, F)$.

Take a function $\omega : \mathcal{V}_2 \to \mathbb{R}$ such that $0 < \omega(v_3)$ for all $v_3 \in \mathcal{V}_2$ and $\sup_{v_3 \in \mathcal{V}_3} \omega(v_3) < \infty$ and define the operators

$$[L_1 g](v_3) = \int_{\mathcal{V}_1} \omega(v_3) p(y_1, y_2, y_3 \mid v_1, v_2, v_3, x_1, x_2) g(v_1) dv_1$$
$$[L_2 g](v_3) = \int_{\mathcal{V}_1} \omega(v_3) p(y \mid v_1, v_2, v_3, v_4, v_5, x) g(v_1) dv_1$$

Then define the operators $\Lambda_2 : \mathcal{L}^1_{bdd}(\mathcal{F}) \to \mathcal{L}^1_{bdd}(\mathcal{V}_3)$ such that

$$[\Lambda_2 g](v_3) = \int_{\mathcal{F}} \omega(v_3) p_3(y_3 \mid y_2, y_1, v_3, x_3, f) g(f) df,$$

$\Lambda_1 : \mathcal{L}^1_{bdd}(\mathcal{V}_1) \to \mathcal{L}^1_{bdd}(\mathcal{F})$ such that

$$[\Lambda_1 g](f) = \int_{\mathcal{V}_1} f_{Y_2, Y_1, F \mid V_1, X}(y_2, y_1, f \mid v_1, x) g(v_1) dv_1,$$

and the diagonal operator $\Delta : \mathcal{L}^1_{bdd}(\mathcal{F}) \to \mathcal{L}^1_{bdd}(\mathcal{F})$ such that

$$[\Delta g](f) = p_5(y_5 \mid y_4, y_3, v_5, x_5, f) p_4(y_4 \mid y_3, y_2, v_4, x_4, f) g(f).$$

Then under sufficient injectivity conditions, we obtain $L_2 = \Lambda_2 \Delta \Lambda_2^{-1}$. Suppose that $p_3(y_3 \mid 0, 0, v_{30}, x_{30}, f) = \pi(f)$. Then, following the proof of Theorem 4.1, we can identify $p_3(y_3 \mid 0, 0, v_{30}, x_{30}, f)$, $f_{Y_2, Y_1, F \mid V_1, X}(0, 0, f \mid v_1, x_0)$, and $p_5(y_5 \mid y_4, y_3, v_5, x_5, f) p_4(y_4 \mid y_3, 0, v_4, x_4, f)$. For $y_1 = 1$, we have

already identified Δ so identification of Λ_2 and Λ_1 follows readily.

However, for $y_2 = 1$, Δ has *not* been identified in the first step. This is why the identification argument in Theorem 4.1 fails, and why $T \geq 5$ is needed, when there are two lags. Instead, define

$$[\tilde{\Lambda}_2 g](v_3) = \int_{\mathcal{F}} \omega(v_3) p_4(y_4 \mid y_3, y_2, v_4, x_4, f) p_3(y_3 \mid y_2, y_1, v_3, x_3, f) g(f) df,$$

and

$$[\tilde{\Delta} g](f) = p_5(y_5 \mid y_4, y_3, v_5, x_5, f) g(f).$$

Then $\tilde{\Delta}$ was identified in the first step since $\sum_{y_5} p_5(y_5 \mid y_4, y_3, v_5, x_5, f) p_4(y_4 \mid y_3, y_2, v_4, x_4, f) = p_4(y_4 \mid y_3, y_2, v_4, x_4, f)$. Moreover, under sufficient injectitivity conditions, $L_2 = \tilde{\Lambda}_2 \tilde{\Delta} \tilde{\Lambda}_2^{-1}$. Applying this spectral decomposition with $y_2 = 1$, the eigenvalues have already been identified so no additional normalization is needed and identification of $\tilde{\Lambda}_2$ follows. Then $p_3(y_3 \mid 1, y_1, v_3, x_3, f) = \sum_{y_4} p_4(y_4 \mid y_3, 1, v_4, x_4, f) p_3(y_3 \mid 1, y_1, v_3, x_3, f)$.

A.6 Other outcomes

Suppose $T = 3$ and let $f_{Y^*, Y|W, X^*}$ denote the joint distribution of Y^* and $Y = (Y_1, Y_2, Y_3)$ conditional on W and X^*, where X^* is an additional vector of covariates. I make the follow assumption in place of Assumption 3.1

Assumption A.4. *For all $y^* \in \mathcal{Y}^*$, $y \in \mathcal{Y}$, $(w, x^*) \in support(W, X^*)$,*

$$f_{Y^*, Y|W, X^*}(y \mid w, x^*) = \int f_{Y^*|Y, W, X^*, F}(y^* \mid y, w, x^*, f) \prod_{t=2}^{T} p_t(y_t \mid y_{t-1}, v_t, x_t, f) f_{Y_1, F|V_1, X}(y_1, f \mid v_1, x) df$$

The important part of this assumption is that Y is independent of X^* conditional on W and F. Because of this, the assumption implies Assumption 3.1. Therefore, p_2, p_3 and $f_{Y_1, F|V_1, X}$ are identified by Theorem 4.1.

Assumption A.5. *For all $y^* \in \mathcal{Y}^*$, $y \in \mathcal{Y}$, $(w, x^*) \in support(W, X^*)$,*

$$f_{Y^*|Y, W, X^*, F}(y^* \mid y, w, x^*, f) = f_{Y^*|Y, V_3, X, X^*, F}(y^* \mid y, v_3, x, x^*, f)$$

This assumption allows for Y^* to depend on some, or all, components of X conditional on X^*, Y, and F, but not on V_1 or V_2. Next, define the operators

$$[L_3 g](v_2) = \int_{\mathcal{V}_1} f_{Y^*,Y|W,X^*}(y^*, y \mid v_1, v_2, v_3, x, x^*) g(v_1) dv_1$$
$$[\Delta^* g](f) = f_{Y^*|Y,V_3,X,X^*,F}(y^* \mid y, v_3, x, x^*, f) p_3(y_3 \mid y_2, v_3, x_3, f) g(f)$$

Following the same arguments as in the text, under Assumptions A.4 and A.5,

$$L_3 = \Lambda_2 \Delta^* \Lambda_1$$

Then, under Assumption 3.7, we can solve for $\Delta^* = \Lambda_2^{-1} L_3 \Lambda_1^{-1}$. This implies that $f_{Y^*|Y,W,X^*,F}(y^* \mid y, w, x^*, f) p_3(y_3 \mid y_2, v_3, x_3, f)$ is identified. If $p_3(y_3 \mid y_2, v_3, x_3, f) > 0$ then $f_{Y^*|Y,W,X^*,F}(y^* \mid y, w, x^*, f)$ is identified.