Identification and $\sqrt{N}$ Efficient Estimation of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor[*]

Xiaohong Chen
London School of Economics

James J. Heckman
University of Chicago
and American Bar Foundation

Edward Vytlacil
University of Chicago

First Draft, June, 1998
Second Draft, May, 1999

[*] Early versions of the paper were presented at the 1998 North American Summer Meetings of the Econometric Society, Canadian Econometric Study Group meeting in Western Ontario, Midwest Econometrics Group meeting in Indiana Bloomington, CREST in Paris, CEME conference in Pittsburgh, and seminars at UCSD, UCR, USC, UCLA, UCB, Indiana, LSE, Virginia, Toronto, Rutgers and the University of Bergen. We have benefited from comments received from participants in all those places, especially from the comments by Bo Honore, Joel Horowitz, Jim Powell, Jean-Marc Robin, Chris Taber and Quong Vuong. The research of the second and third authors was supported by NSF 97-09-873, NIH R01-HD-34958-01 and NIH R01-HD32058-03.

1 Introduction

Sequences of binary dependent variables arise in many contexts in analyzing economic panel data. As an example, consider the analysis of monthly fertility histories (Heckman and Willis, 1977). Each month, t, a woman either conceives a child ($D_t = 1$) or does not ($D_t = 0$). The woman is observed over sample period $t = 1, ..., T$. Many demographers, economists and sociologists study such strings to determine the causes of the timing and spacing of births. As another example, consider labor force histories. In each period t, persons either work ($D_t = 1$) or do not work ($D_t = 0$), and the goal of many studies is to determine the temporal patterns of employment. Panel data analyses of unemployment, welfare dependence and participation in crime have a similar character.

A common finding in many studies of panel data is dependence in outcomes. Typically $\Pr(D_t = 1 \mid D_{t-1} = 1) > \Pr(D_t = 1)$ whether or not we condition on observed characteristics X_t. One common explanation for this dependence is temporally persistent person-specific unobserved variables that cause persons who are more likely than average to occupy a state in one period to occupy it in another period. The unobserved variables give rise to temporally persistent heterogeneity in outcomes across persons. Failure to control for this bias is often said to produce heterogeneity bias. In the literature on fertility, a persistent unobserved component that gives rise to such bias is called fecundity (Sheps and Menken, 1973). In studies of mortality and survival it is called frailty. The problem of controlling for the effects of temporally persistent unobserved components and estimating their distribution is a central problem in science and social science.

Heckman (1981a) extends cross-section discrete choice theory and presents a class of parametric, binary, discrete-time, panel data models that allow for general forms of dependence and heterogeneity bias. His models generate binary discrete outcomes as a consequence of latent random variables crossing thresholds. These latent variables are utility (or value function) differences between potential states, and capture the essential idea that comparisons between alternative states generate choices. Heckman introduces a variety of stochastic processes for the unobservables to produce a rich array of micro stochastic processes for binary panel data that provide a framework for incorporating discrete dynamic choice theory into the discrete-data panel data analysis. These discrete-time, discrete-data models thus provide a convenient framework for choice theoretic models of binary panel data. This is in contrast with continuous time duration models which are typically difficult to justify using economic theory. (See the discussion in Heckman and Singer, 1984).

This paper is a first installment of an ongoing project that extends the models of Heckman to a semiparametric setting. In this paper, we consider random effects models where there exists a scalar factor representation for the unobserved random shock.[1] As discussed in Heckman (1981a) and Amemiya (1985), the factor structure allows for fairly flexible ser-

[1] In Chen et al. (1998), we generalize the scalar factor considered here to a multi-factor model.

ial correlation in the shocks while greatly reducing the dimensionality of the problem. The factor structure assumption is in contrast to the rigid permanent-transitory error scheme that is common in the literature and is critical for "fixed-effects" estimation. The factor structure model is more flexible than the permanent-transitory error scheme, and includes the permanent-transitory error scheme as a special case.

We consider three specific models: a repeated binary choice model, a single spell duration model, and a switching regression model with binary outcomes. For each model, we consider nonparametric identification of the model following an "identification in the limit" strategy. Having shown the conditions under which they are identified, we examine the conditions under which the models can be well estimated. In particular, we examine the semiparametric efficiency bounds for each model, and present conditions under which the structural parameters of each model are $\sqrt{N}$ estimable or not. We show that $\sqrt{N}$ estimation depends on the number of time periods observed and on particular features of the model. $\sqrt{N}$ estimableness is a much more fragile property then identification. We then examine the properties of the nonparametric MLE (NPMLE) estimator. We establish convergence rate of the NPMLE estimate for the density and the $\sqrt{N}-$normality and efficiency of the NPMLE estimate for the $\sqrt{N}-$estimable structural parameters. We present Monte Carlo results that support our theoretical analysis.

Our results have direct implications for the literature that uses NPMLE to estimate random effects, discrete-choice, discrete-time models with a nonparametric latent factor. Following the work of Heckman-Singer (1984), the NPMLE mixture approach has been applied to repeated logistic regressions by Follman (1985) and Follman and Lambert (1989), to discrete-time single spell duration model (based on Monte Carlo studies) by Baker and Melino (1997), and to more general discrete-time discrete-choice models by Cameron and Heckman (1987). These NPMLE discrete-time discrete-choice methods have been used in empirical work in a wide variety of areas. Despite the use of this methodology for applied work, there has been little theoretical research on the properties of these estimators. The only asymptotic analysis of these estimators is Follman (1985) who shows consistency in the case of repeated logistic regressions. There are no asymptotic distribution results known for these models. In applications, researchers typically conduct hypothesis testing under the assumption that the structural parameters are estimated at $\sqrt{N}$, despite there having been no theoretical justification for this practice. Our analysis is directly relevant for this literature. We show that the NPMLE estimates of the structural parameters in these discrete-choice mixture models are $\sqrt{N}$-normal only under very specific conditions, so that great care must be taken by researchers to determine if the structural parameters of their specific model will be estimated at $\sqrt{N}$.

The plan of this paper is as follows. In Section 2, we present the three models analyzed in this paper, and discuss recent applications of these models. In Section 3, we present identification results for each model with general error structure. In Section 4, we present identification results for each model with scalar factor error structure. In Section 5, we consider the $\sqrt{N}$ estimableness of the structure parameters of each model. Section 6 examines the convergence rates of the NPMLE estimator, and establishes conditions under which the structural parameters estimated by NPMLE will be $\sqrt{N}-$ normal and efficient. Section 7 presents Monte Carlo analyses. Section 8 concludes the paper by a short discussion of future research.

2 Models Considered in This Paper

2.1 Repeated Binary Choice Model

The simplest model we consider is a repeated binary choice model with serially correlated errors but no lagged dependence. In particular we investigate the following stochastic process. Let $D_t \in \{0,1\}$ denote the agent's discrete choice in period t, $t = 1, ..., \bar{T}$.[2] We specify that D_t is determined by an underlying index as follows:

$$\begin{aligned} D_t &= 1(I_t \geq 0) \\ I_t &= X_t\beta_t - \eta_t \end{aligned} \tag{1a}$$

where η_t is an unobserved random shock. $1(A)$ is an indicator variable that takes the value 1 if the event A occurs, and takes the value 0 otherwise. X_t are regressors entering the index for period t and are assumed to be strictly exogenous. X_t can include expectations of future outcomes in the case of forward looking behavior. The linearity of the index in X_t is not critical, and we relax this assumption in Chen et al. (1998). Let $X = (X_1, ..., X_{\bar{T}})$, and let $\eta = (\eta_1, ..., \eta_{\bar{T}})$. We assume that X is independent of η.

We assume that η_t has a factor model representation:

$$\eta_t = -\alpha_t\Theta + U_t \qquad \alpha_1 = 1 \tag{1b}$$

Θ is an unobserved factor with unknown distribution. It is an individual-specific, time-invariant effect representing "unobserved heterogeneity". Setting $\alpha_1 = 1$ is an innocuous normalization. We assume that $U_t \perp\!\!\!\perp U_{t'}$, for $t \neq t'$, and that $U_t \perp\!\!\!\perp \Theta$. We assume that $(\Theta, U_1, ..., U_{\bar{T}})$ is jointly independent of X.

We will use the following notation. Let $\beta = (\beta_1, ..., \beta_{\bar{T}})$. Since there is no lagged dependence by assumption for this model, there is no problem of initial conditions. However, to maintain notational uniformity with the single-spell duration model, we let D_0 be the initial condition, which we assume that we observe. We will use D^{t-1} to denote the sequence of choices up to period t: $(D_0, ..., D_{t-1})$, and let D denote the full sequence: $D = (D_0, ..., D_{\bar{T}})$. We will let F_{η_j} be the distribution of η_j, F_{η_j,η_k} be the distribution of (η_j,η_k), and will let F_η be the joint distribution of $\eta \equiv (\eta_1, ..., \eta_{\bar{T}})$.

[2] While we consider only the case of a binary decision, the extension of our results to a multinomial decision process is straightforward.

We note in passing that if we assume a permanent-transitory model for the error term, $\eta_t = \Theta + U_t$ for all t, then the model can be analyzed using standard fixed-effect approaches (see, e.g., Arellano and Honoré, 1999, for a review of fixed-effect methodology). The fixed-effect approach allows for arbitrary dependence between the regressors and the disturbances, and in that way is more general than the repeated binary choice model considered here. However, a critical trade-off is that the random effects assumption allows us to identify and estimate the full joint distribution of the disturbances, and we can thus identify marginal effects and make out-of-sample predictions.[3]

2.2 Single Spell Duration Model

Consider the following one-spell model of duration. Assume individuals start out in a state and exit it at time $T = t$, so that T is a random variable representing total completed spell length. Let $D_t = 1$ if the individual survives to time t and $D_t = 0$ otherwise. The event $D_{t-1} = 0$ signifies that an individual has dropped out of the initial state by date t. There is no meaningful event corresponding to the outcome $D_t = 1$ and $D_{t-1} = 0$. Let $X_t = x_t$ denote regressors determining transitions from time $t-1$ to time t, let $\bar{T}$ be the upper limit on the survival time, and impose the initial condition that $D_0 = 1$.

We represent our duration model as arising from the threshold-crossing behavior of a sequence on underlying latent indices given by:

$$\left.\begin{array}{rcl} D_t & = & 1(I_t \geq 0) \\ I_t & = & X_t\beta_t - \eta_t \end{array}\right\} \text{ if } D^{t-1} = (1, ..., 1) \tag{2a}$$

where the definitions and assumptions are the same as in model 1. However, note here that the D_t outcome is observed only if $D_{t-1} = 1$ (which is equivalent to $D^{t-1} = (1, ..., 1)$), and thus all period t parameters are implicitly conditional on all past choices. Within this model, the decision rule for D_t is not well defined if $D_{t-1} = 0$. We again assume that η_t has a factor model representation:

$$\left.\begin{array}{rcl} \eta_t & = & -\alpha_t\Theta + U_t \\ \alpha_1 & = & 1 \end{array}\right\} \text{ if } D^{t-1} = (1, ..., 1) \tag{2b}$$

where we assume that $(\Theta, U_1, ..., U_{\bar{T}})$ is jointly independent of X.

Single-spell duration models are common in the applied literature – for example, time till end of a spell of unemployment, time till end of a strike, and time till dropping out of schooling. As an example, consider Cameron and Heckman's (1998) analysis of schooling

[3] See Chen et al. (1998) and Honoré and Lewbel (1998) for analysis that extends the results of Lewbel (1998) to a panel data context. These approaches allow for arbitrary dependence between the disturbances and all but one of the regressors, while allowing for a factor-structure for the error term and while recovering the joint distribution of the disturbances.

transitions. In their application, D_t is an indicator for the agent completing grade t. If $D_t = 0$, then the agent has dropped out of school by grade t, and is not eligible to transit to grade $t+1$. If $D_t = 1$, then the agent has completed at least grade t and is at risk of completing grade $t+1$. A central question for this education literature is whether the effects of family background and family resources on educational grade transitions diminish at higher levels of education. This question was interpreted by Mare to be whether the corresponding β_t coefficients are lower for higher education levels (higher t values). Cameron and Heckman impose a factor structure assumption with the distribution of U_t assumed to be logistic and the distribution of Θ unknown, and estimate the model using NPMLE.

We note in passing that single-spell duration models cannot be estimated using fixed effect approaches. No individuals leave and re-enter the sample in single-spell duration models, and such behavior is required by fixed effect approaches.

2.3 Switching Regression Model

The next model that we consider is a switching regression model with binary outcome variables. We will formally view the switching regression model as a two period model where the second period depends on the first period outcome. Of course, the model need not have any intertemporal aspect and the two periods need not correspond to two points in calendar time.

Let $D_1 \in \{0,1\}$ denote the agent's discrete choice in period 1, D_1 is determined by an underlying index as before:

$$\begin{aligned} D_1 &= 1(I_1 \geq 0) \\ I_1 &= X_1\beta_1 - \eta_1 \end{aligned} \tag{3a.1}$$

For the second period, $D_2 \in \{0,1\}$ is determined by:

$$D_2 = (1 - D_1)\, D_{20} + D_1 D_{21} \tag{3a.2}$$

where

$$\left.\begin{aligned} D_{2j} &= 1(I_{2j} \geq 0) \\ I_{2j} &= X_2\beta_{2j} - \eta_{2j} \end{aligned}\right\} j = 0,1 \tag{3a.3}$$

Note that we could equivalently define the model for D_2 as follows, which is more symmetric with the notation of model (1):

$$\begin{aligned} D_2 &= 1(I_2 \geq 0) \\ I_2 &= X_2\beta_{20} + D_1\left[X_2(\beta_{21} - \beta_{20}) + (\eta_{20} - \eta_{21})\right] - \eta_{20} \end{aligned} \tag{3a.4}$$

We will again assume a factor model representation for the shocks:

$$\begin{aligned} \eta_1 &= -\Theta + U_1 \\ \eta_{2j} &= -\alpha_{2j}\Theta + U_{2j} \qquad j = 0,1 \end{aligned} \tag{3b}$$

where we assume that $(\Theta, U_1, U_{21}, U_{20})$ is jointly independent of X. D_{20} is the outcome the individual would have had if he or she had chosen $D_1 = 0$ in period 1, and D_{21} is the outcome the individual would have had if he or she had chosen $D_1 = 1$ in period 1. D_{20} and D_{21} are thus sometimes called "potential outcomes." Which outcome is observed depends on the choice actually made in the first period. As is made clear by expression (3a.4), the model can be viewed as a random coefficient model. In addition, the model can be viewed as a two period, heterogeneous, time inhomogeneous Markov chain model.

We will use the following notation. Let $\beta = (\beta_1, \beta_2)$ where $\beta_2 = (\beta_{20}, \beta_{21})$. Since there is no lagged dependence for the D_1 choice, there is no problem of initial conditions for this model. However, to maintain notational uniformity with the single-spell duration model, we let D_0 be the initial condition, which we assume that we observe.

Switching regression models are common in the applied literature. As an example, consider Aakvik, Heckman and Vytlacil (1998) who apply this model to study the effects of job training on employment outcomes.. For them, D_1 is an indicator for whether the person enters the training program, and D_2 is an indicator for whether the individual is employed three years later. Aakvik et al. assume that the distribution of the U terms is known, but that the distribution of Θ is unknown. This framework allows for heterogeneous training effects, and allows for program participation to be based in part on the individual's idiosyncratic training effect. Note that for this model it is critical that β be allowed to vary with time, since it is a fundamentally different choice in the second period than in the first period. The full β vector being allowed to vary with past outcomes $(\beta_{00} \neq \beta_{01})$ allows the effect of treatment to vary with the agent's observable characteristics. Using the factor structure with the factor loadings dependant on the treatment decision $(\alpha_{00} \neq \alpha_{01})$ allows the effect of training to vary with the agent's unobservable characteristics. Aakvik et al. apply their model to evaluate the effectiveness of the Norwegian Vocational Rehabilitation Program. They impose the scalar factor assumption, and use NPMLE to estimate the structural model. Given the estimation of the structural model, they derive estimates of various treatment parameters such as the effect of treatment on the treated, average treatment effect, and local average treatment effect. They find that the treatment effect varies substantially with both the observable and unobservable characteristics, and that ignoring this heterogeneity in treatment effects results in vastly overstating the program's effectiveness.

We note in passing that this model cannot be estimated using fixed effect approaches. Fixed effect approaches can be used if we assume multiple (D_1, D_2) observations for each individual and stationarity conditions are imposed.[4]

[4] See the discussions in Honorè and Kyriazidou (1998b) and Arellano and Honorè (1999).

3 Identification of the Models with General Error Structures

We now consider identification of each model without imposing the factor structure, i.e., we work with the models defined by equations (1a), (2a), and (3a) without imposing the factor structure assumption of (1b), (2b), and (3b). Given identification of the distribution of the shocks, we will then discuss conditions under which these distributions imply a unique factor model representation for the shocks[5].

We assume that we observe a large sample of i.i.d (D, X) observations. Thus, for any d^{t-1} s.t. $Pr(D^{t-1} = d^{t-1}) > 0$, we can nonparametrically identify $Pr(D_t = d_t|X, D^{t-1} = d^{t-1})$ a.e. $F_{X|D^{t-1}=d^{t-1}}$ where $F_{X|D^{t-1}=d^{t-1}}$ is the distribution of X conditional on previous choices. We assume that we know that $(\beta, F_\eta) \in \mathcal{B} \times \mathcal{H}$, where $\mathcal{B} \times \mathcal{H}$ is the parameter space. Our goal is to establish conditions under which knowledge of $Pr(D_t = d_t|X, D^{t-1} = d^{t-1})$ a.e. $F_{X|D^{t-1}=d^{t-1}}$ allows us to identify a unique element of $\mathcal{B} \times \mathcal{H}$. We define identification of the parameters as follows.

Definition 1 *Let $P_{\beta,F_\eta}(D_t = 1|X = x, D^{t-1} = d^{t-1})$ be the probability of observing the choice $D_t = 1$ conditional on observables $X = x$ and past choices $D^{t-1} = d^{t-1}$ under the particular model (1a, 2a or 3a) when the parameter values are given by (β, F_η). Let $\mathcal{B} \times \mathcal{H}$ be the space of permissible parameter values. We will say that $(\beta, F_\eta) \in \mathcal{B} \times \mathcal{H}$ is identified iff for all $(\beta^*, F^*_\eta) \in \mathcal{B} \times \mathcal{H} \backslash (\beta, F_\eta)$, there exists a sequence of past choices, d^{t-1}, $Pr(D^{t-1} = d^{t-1}) > 0$, s.t.*

$$\Pr{}_{X|D^{t-1}=d^{t-1}} \left\{ P_{\beta,F_\eta}(D_t = 1|X, D^{t-1} = d^{t-1}) \neq P_{\beta^*,F^*_\eta}(D_t = 1|X, D^{t-1} = d^{t-1}) \right\} > 0$$

Theorem 1 shows conditions for identification of the first model, the repeated binary choice model without lagged dependence. This is the easiest model of the three to identify. Identification follows from applying Manski to identify the β_t and F_{η_t} parameters for each t, and then using an assumption that we can independently vary each index to trace out the full joint distribution of F_η.

Theorem 1 (repeated binary): For the model defined by equation (1a), provided that:

(i) $\eta \equiv (\eta_1, ..., \eta_{\overline{T}})$ is statistically independent of $X \equiv (X_1, ..., X_{\overline{T}})$.

(ii) F_η is absolutely continuous with respect to Lebesgue measure on $R^{\overline{T}}$ with support $\prod_{t=1}^{\bar{T}} (L_t, U_t)$, where $-\infty \leq L_t < U_t \leq +\infty$ for all $t = 1, ..., \overline{T}$ does not depend on β.

[5] We do not here pursue the alternative identification strategy of identifying the factor structure directly. Cameron and Heckman (1998) and Chen et al. (1998) show that if sufficient distributional assumptions are imposed on Θ and $(U_1, ..., U_{\bar{T}})$, then it is possible to identify the factor structure and thus identify F_η even when the distribution of $F_{\tilde{\eta}}$ is nonparametrically unidentified.

(iii) For all $t = 1, \ldots, \bar{T}$, X_t is a K_t-dimensional random variable and there exists no proper linear subspace of R^{K_t} having probability 1 under F_{X_t}.

(iv) $Supp\,(X_t\beta_t|X_1\beta_1 = g_1, ..., X_{t-1}\beta_{t-1} = g_{t-1}) \supseteq (L_t, U_t)$ for almost every $(g_{t-1}, \ldots, g_1) \in \prod_{i=1}^{t-1}(L_i, U_i)$, for $t = 1, \ldots, \bar{T}$, where the boundary points $\{L_t, U_t : t = 1, ..., \bar{T}\}$ are not functions of β_t for $t = 1, ..., \bar{T}$.

Then: F_η and $(\beta_1, ..., \beta_{\overline{T}})$ are identified given location and scale normalizations. (e.g., either if X_t is constrained not to contain an intercept and $\|\beta_t\| = 1$ for $t = 1, ..., \bar{T}$, or if $F_{\eta_t}(\cdot)$ has median zero and variance one for $t = 1, ..., \bar{T}$, then F_η and $(\beta_1, ..., \beta_{\overline{T}})$ are identified).

Proof. We observe (D_t, X_t) for each individual for each $t = 1, ..., \bar{T}$, and can thus identify the left hand sides of:

$$Pr\,(D_t = 1|X_t = x_t) = F_{\eta_t}(x_t\beta_t)$$

for each $t = 1, ..., \bar{T}$. Each equation is a standard binary discrete choice model. Using the results of Manski (1988, Proposition 2, Corollary 5), under conditions (i), (ii), (iii) and the support condition on $X_t\beta_t$ in (iv) of Theorem 1, we have that we identify β_t and the distribution of η_t up to scale and location for each $t = 1, ..., \bar{T}$. For example, we may normalize the location and scale by constraining X_t not to have an intercept and constraining $||\beta_1|| = 1$. We thus recover $x_t\beta_t$, $t = 1, \ldots, \bar{T}$.

We also identify the left hand side of the following equation:

$$\begin{aligned} & Pr\,(D_1 = 1, D_2 = 1, \ldots, D_{\bar{T}} = 1|X_1 = x_1, \ldots, X_{\bar{T}} = x_{\bar{T}}) \\ = \; & F_\eta\,(x_1\beta_1, \ldots, x_{\bar{T}}\beta_{\bar{T}})) \,. \end{aligned}$$

where $F_\eta = F_{\eta_1,\ldots,\eta_{\bar{T}}}$. Since we identify $x_t\beta_t$ for each t, and using (iv), we can vary the components of $(x_1\beta_1, ..., x_{\bar{T}}\beta_{\bar{T}})$ to trace out the joint distribution F_η. ■

The first three assumptions in Theorem 1 are very standard – the first two are independence assumptions typical for discrete choice models, and the third assumption is simply a full rank condition. We can trivially relax the first two conditions (i) and (ii) to a weaker index sufficiency assumption and still identify the β parameters, though we would no longer be able to identify F_η. If condition (iii) is weakened so that X_t lies in a proper subspace of R^{K_t} w.p. 1, we can only identify linear combinations of the β_t.

Assumption (iv) is the least standard assumption in Theorem 1. As in Manski (1988), we require that the support of $X_t\beta_t$ is at least as large as that for η_t. There is an important distinction here between components of X_t that are continuous and those that are discrete. If all components of X_t are discrete, condition (iv) cannot possibly be satisfied because there are no intervals in the support of $X_t\beta_t$. However, (iv) is stronger than was required by Manski – it imposes not only that the support of $X_t\beta_t$ is at least as large as that for η_t, but also that the support of $X_t\beta_t$ is at least as large as that for η_t given levels of the preceding indices. It is this stronger support condition which allows us to independently vary the indices to trace out F_η. If we relax this assumption by assuming that the support of $X_t\beta_t$ is at least as large as that for η_t but not conditional on the preceding indexes, then we will still identify β and $(F_{\eta_1}, ..., F_{\eta_T})$, but will no longer be able to identify F_η.

We now consider identification of the single-spell duration model. Identification of this model is harder than for identification of the repeated binary model. We no longer directly identify $Pr\left(D_t = 1|X_t = x_t\right)$ for $t \geq 2$, but instead directly identify $Pr\left(D_t = 1|D_{t-1} = 1, X = x\right)$. We therefore follow an identification in the limit strategy that allows us to recover $Pr(D_t = 1|X_t = x_t)$ by conditioning on large values of the proceeding indices. The full support condition (iv) of Theorem 1 allows us to do this. We then need to augment the assumptions of Theorem 1 to have that X_t has full rank when conditioning on the values of the previous indices.

Theorem 2 (single spell): For the model defined by equation (2a), assume conditions (i), (ii), and (iv) of Theorem 1, and strengthen condition (iii) of Theorem 1 to the following condition:

(iii′) For all $t = 1, \ldots, \bar{T}$, X_t is a K_t-dimensional random variable. There exists no proper linear subspace of R^{K_1} having probability 1 under F_{X_1}. The exists a $\check{g} = (\check{g_1}, ..., \check{g_{t-1}})$ s.t. for almost every $g = (g_1, \ldots, g_{t-1}) \in \prod_{j=1}^{t-1} (L_j, U_j)$ with $g \geq \check{g}$, there exists no proper linear subspace of R^{K_t} having probability 1 under $F_{X_t|X_1\beta_1 \geq g_1, \ldots, X_{t-1}\beta_{t-1} \geq g_{t-1}}$.

Then: F_η and $(\beta_1, ..., \beta_{\overline{T}})$ are identified given location and scale normalizations.

Proof. See Appendix A. ■

The proof of Theorem 2 is much longer than the other proofs in this sections, and we therefore place it in the appendix. The theorem is an adaptation and refinement of Theorem 2 in Cameron and Heckman (1998).

The assumptions of Theorem 2 can clearly be satisfied if X contains some kind of exclusion restriction or some component of X that varies across transitions. However, the following corollary tells us that the assumptions can be satisfied even when X_t are the same across all time periods if sufficient structure is placed on how the β_t vary with t.

Corollary 1: For the model defined by equation (2a), suppose

(v) The first $\bar{T}$ coordinates of $\overline{X}$ are continuous random variables $\left(\bar{T} \leq K\right)$. The support of $\prod_{i=1}^{\bar{T}} X^i$ is $\prod_{i=1}^{\bar{T}} (-\infty, \infty)$ where X^i is the ith coordinate of $\overline{X}$

(vi) $\beta_1, \ldots, \beta_{\bar{T}}$ are linearly independent, and β_t^i ,($i = 1, ..., \bar{T}$), the first $\bar{T}$ coordinates of β_t, are non-zero for all $t = 1, ..., \bar{T}$.

Then assumptions (iii′) and (iv) of Theorem 2 are satisfied with $L_i = -\infty, U_i = \infty$. Thus, under assumptions (i) and (ii) of Theorem 1 and assumptions (v)-(vi) above, F_η and $\{\beta_t : t = 1, \ldots, \bar{T}\}$ are identified up to scale and location normalizations.

Finally, we turn to identification of the switching regression model.

Theorem 3 (switching regression): For the model defined by equations (3a.1-3a.4), assume that for $j = 0, 1$:

(i) $\left(\eta_1, \eta_{2j}\right)$ is statistically independent of (X_1, X_2).

(ii) $F_{\eta_1,\eta_{2j}}$ is absolutely continuous with respect to Lebesgue measure on R^2 with support $(L_1, U_1) \times (L_{2j}, U_{2j})$, where $-\infty \leq L_1 < U_1 \leq +\infty$, and $-\infty \leq L_{2j} < U_{2j} \leq +\infty$ does not depend on β.

(iii) $X_1 \in \overset{=}{X}_1 \subseteq R^{K_1}$ is a K_1- dimensional random variable. There exists no proper linear subspace of R^{K_1} having probability 1 under F_{X_1}. $X_2 \in \overset{=}{X}_2 \subseteq R^{K_2}$ is a K_2- dimensional random variable. For almost every $g_1 \in (L_1, U_1)$, there exists no proper linear subspace of R^{K_2} having probability 1 under $F_{X_2|X_1\beta_1=g_1}$.

(iv) $Supp(X_1\beta_1) \supseteq (L_1, U_1)$, $Supp\left(X_2\beta_{2j}|X_1\beta_1 = g_1\right) \supseteq (L_{2j}, U_{2j})$, for almost every $g_1 \in (L_1, U_1)$ where the boundary points $\{(L_1, U_1), (L_{2j}, U_{2j})\}$ are not functions of β.

Then: $(F_{\eta_1,\eta_{20}}, F_{\eta_1,\eta_{21}})$ and (β_1, β_2) are identified given scale and location normalizations.

Proof. By hypothesis, we know the left hand sides of the following 3 equations:

$$Pr\left(D_1 = 1|X_1 = x_1\right) = F_{\eta_1}\left(x_1\beta_1\right)$$

$$Pr\left(D_1 = 1, D_2 = 1|X_1 = x_1, X_2 = x_2\right) = F_{\eta_1,\eta_{21}}\left(x_1\beta_1, x_2\beta_{21}\right)$$

$$Pr\left(D_1 = 0, D_2 = 1|X_1 = x_1, X_2 = x_2\right) = F_{-\eta_1,\eta_{20}}\left(-x_1\beta_1, x_2\beta_{20}\right)$$

Using the first and second equations, we can follow the same argument as for Theorem 2 to show that ($F_{\eta_1,\eta_{21}}$, (β_1, β_{21})) parameters are identified given location and scale normalizations. Proceeding in the same fashion by using the first and third equations we can show that the ($F_{\eta_1,\eta_{20}}$, (β_1, β_{20})) parameters are identified given location and scale normalizations. ■

The theorem informs us that we can nonparametrically identify the joint distribution of (η_1, η_{20}), the joint distribution of (η_1, η_{21}), and the β parameters up to scale. However, for each individual, we observe either D_{20} or D_{21} but never both. Thus, we do not identify $Pr(D_{20} = 1, D_{21} = 1|X)$, and cannot nonparametrically identify $F_{\eta_{20},\eta_{21}}$. All we can say about $F_{\eta_{20},\eta_{21}}$ is that it belongs to the set of all distributions consistent with $F_{\eta_1,\eta_{20}}$ and $F_{\eta_1,\eta_{21}}$. Identification of $F_{\eta_{20},\eta_{21}}$ may be of substantive interest.[6] One can identify F_{η_1,η_0,η_D} given additional identifying assumptions.[7]

Theorem 3 suggests a general identification strategy. Any sequence of choices can be mapped into the form of a single-spell duration model, and thus conditions for identification of the parameters of any sequence of choices can be found through the conditions required for identification of the single-spell duration model. Parameters of a particular form of dependence, such as a lagged dependent variable, can then be identified through the difference of the parameters of two sequences. Chen et al. (1998) use this approach to formulate identification conditions for binary choice models with very general forms of state dependence. However, as illustrated by the switching regression results, identification of the joint distribution of the shocks of two different sequences of choices is not identified unless more structure is imposed.

4 Identification of Models with Scalar Latent Factor

In the last section, we presented sufficient conditions to identify β and the joint distribution F_η up to scale and location normalizations without imposing structure on F_η. As Heckman (1981) and Amemiya (1985) pointed out, it is computationally too intensive to estimate the $\overline{T}$ -dimensional joint distribution F_η nonparametrically for $\overline{T} \geq 3$. In practice, people impose a permanent-transitory error scheme or the more general factor structure (e.g., (1b), (2b) and (3b)) to reduce dimensionality.[8] In this section, we consider identification when imposing the factor structure.

Identification and estimation of the factor structure is fundamentally connected to the identification of the scale, so we first discuss scale and location normalizations. Here, and throughout the rest of the paper, we will make the following location normalization:

$$X_t = \left(1, X_t^{(-1)}\right) \text{ , median}(\widetilde{\eta}_t) = 0 \text{ for } t = 1, .., \bar{T}. \tag{LN}$$

[6] For example, in Aakvik et al. (1998), the treatment effect is $D_{21} - D_{20}$, and the distribution of treatment effects is identified only if $F_{\eta_{20},\eta_{21}}$ is identified. The distribution of treatment effects and not just the average treatment effect may be the parameter of interest – see Heckman and Smith (1998), and Heckman, Smith and Clements (1997).

[7] One source of identifying information is to impose additional information on the first period decision rule. In particular, identifies $F_{\eta_{20},\eta_{21}}$ under the Roy model assumption that the participation is based only on the gains from participation (Heckman and Honoré, 1990, and Heckman and Smith, 1998). Additional structure on $F_{\eta_{20},\eta_{21}}$ will also allow identification. For example, a common effects assumption or a scalar factor structure assumption are sufficient for identification of $F_{\eta_{20},\eta_{21}}$. See Aakvik et al. (1998).

[8] The permanent-transitory error scheme is $\eta_t = \Theta + U_t$, which is the special case of the factor structure with $\alpha_t = 1$ for all t.

where the notation $X_t = \left(1, X_t^{(-1)}\right)$ means that the first element of X_t is one, and the rest of elements of X_t are real-valued continuous or discrete-valued non-degenerate randome variables. Correspondingly, $\beta_t^{(1)}$ denotes the intercept parameter and $\beta_t^{(-1)}$ denotes the vector of slope parameters.

For the scale normalizations, a common scale normalization when not imposing a factor structure is:

$$||\beta_t|| = 1 \quad \text{for } t = 1, ..., \bar{T}. \tag{SN1}$$

For the rest of the paper, we will use $F_{\tilde{\eta}}$ and $\tilde{\beta}$ to denote the parameter values defined through normalizations (LN) and (SN1). (LN) and (SN1) are always sufficient to tie down the location and scale of the model, and thus $\tilde{\beta}$ and $F_{\tilde{\eta}}$ are immediately identified under the conditions of Theorems 1 to 3. Let $\tilde{\alpha}$, $F_{\tilde{\Theta}}$ and $F_{\tilde{U}_t}$ denote the parameters of the factor structure consistent with (SN1) and (LN). Identification of these parameters is discussed below.

The normalization (SN1) is not a typical normalization when imposing a factor structure. The standard normalization when imposing a factor structure is

$$Var(U_t) \text{ is known.} \tag{SN2}$$

Normalization (SN2) is not always sufficient to tie down the scale of η_t and β_t. For example, consider the case of $\bar{T} = 1$, $\Theta \sim N(0, \sigma^2_\Theta)$, and $U_1 \sim N(0, 1)$. Then choices determined by decision rule

$$D_t = 1\left(X_t\beta_t \geq U_t\right)$$

are observationally equivalent to choices determined by the decision rule

$$D_t = 1\left(X_t(\sigma^2_\Theta + 1)\beta_t \geq -\Theta + U_t\right)$$

for any $\sigma^2_\Theta \geq 0$, and thus a model with parameters $\left((\sigma^2_\Theta + 1)\beta, \sigma^2_\Theta\right)$ is not identified versus parameters $\left((\sigma^{2\prime}_\Theta + 1)\beta, \sigma^{2\prime}_\Theta\right)$ for any σ^2_Θ, $\sigma^{2\prime}_\Theta \geq 0$. An important question is thus under what conditions is (SN2) sufficient to tie down the scale.

For the rest of the paper, we will use F_η and β to be the parameter values defined through normalizations (LN) and (SN2). Note that there is a simple relationship between the parameters defined by (SN1) and (SN2). Let

$$\gamma_t = 1/\left\|\beta_t\right\|.$$

Then we immediately have that

$$\begin{array}{rcl} \tilde{\Theta} & = & \gamma_1 \Theta \\ \tilde{\alpha}_t & = & \frac{\gamma_t}{\gamma_1}\alpha_t \end{array} \tag{1}$$

$$\begin{pmatrix} \tilde{\beta}_t \\ \tilde{\eta}_t \\ \tilde{U}_t \end{pmatrix} = \gamma_t \begin{pmatrix} \beta_t \\ \eta_t \\ U_t \end{pmatrix} \qquad (2)$$

γ_t determines the relationship between (SN1) and (SN2). If we impose $\sigma^2_{U_t} = 1$, then γ^2_t is just the variance of $\tilde{U}_t$, it is the variance of the shock that is consistent with (SN1). Given (SN1) and the conditions of one of Theorems 1 to 3, $Var(\tilde{\eta}_t)$ is immediately identified. Thus, if $(\gamma_1, ..., \gamma_{\bar{T}})$ is identified, then using the above equalities one can trivially show that all scale parameters of the factor structure are identified (σ^2_Θ and $(\alpha_2, ..., \alpha_{\bar{T}})$, as well as $\sigma^2_{\tilde{\Theta}}$, $\sigma^2_{\tilde{U}_t}$ and $(\tilde{\alpha}_2, ..., \tilde{\alpha}_{\bar{T}})$).

In this section, we present two kinds of identification results for the scalar factor structure: the first kind nonparametrically identifies the distributions of Θ and of $U_t, t = 1, ..., \overline{T}$ from $F_{\tilde{\eta}}$; the second kind nonparametrically identifies the distribution of Θ while assuming the distributions of $U_t, t = 1, ..., \overline{T}$ are completely known. Notice that the second kind is widely applied in practice for the ease of computation. For both sets of results, the identification analysis will proceed as follows. We start by taking $F_{\tilde{\eta}}$ and $\tilde{\beta}$ as identified from the analysis of section 3 and normalizations (LN) and (SN1). We then show identification of the model with the factor structure and normalizations (LN) and (SN2) by first showing identification of $(\gamma_1, ..., \gamma_{\bar{T}})$ and thus identification of the scale terms of the factor structure, and by second showing identification of the distribution parameters of the factor structure.

4.1 Repeated Binary Choice and Single Spell Duration Models

We now consider the identification of factor structure for the repeated binary choice and the single spell duration models. We will assume that $F_{\tilde{\eta}}$ is identified, as would be the case under the assumptions of Theorem 1 (for the repeated binary choice model) or Theorem 2 (for single spell duration model) while imposing normalizations (LN) and (SN1).

Theorem 4 (nonparametric): Assume:

(i) $F_{\tilde{\eta}}$ is identified.

(ii) $(U_1, ..., U_{\bar{T}})$ are mutually independent with zero medians, finite known variances, and the unknown distribution (G_t) of U_t has non-vanishing characteristic function for $t = 1, \ldots, \bar{T}$.

(iii) Θ_i is independent of $(U_1, ..., U_{\bar{T}})$ with zero median, finite unknown variance, and its unknown distribution (H) has non-vanishing characteristic function.

Then:

(a) $(\alpha_2, ..., \alpha_{\bar{T}}, \gamma_1, ..., \gamma_{\bar{T}}, Var(\Theta))$ and the distributions $(G_1, ..., G_{\overline{T}}, H)$ are identified when $\bar{T} \geq 3$ and $0 < |\alpha_t|, \gamma_t < \infty$ for all $t = 1, ..., \overline{T}$.

(b) When $\bar{T}=2$ and $0<|\alpha_2|,\gamma_1,\gamma_2<\infty$, if any one of $\alpha_2,\gamma_1,\gamma_2,Var(\Theta)$ is assumed to be known, then the rest of $\alpha_2,\gamma_1,\gamma_2,Var(\Theta)$, and the distributions (G_1,G_2,H) are identified

Proof.

The factor structure $\widetilde{\eta}_t=\gamma_t(-\alpha_t\Theta_i+U_t)$, $\alpha_1=1$ for $t=1,\ldots,\bar{T}$ gives us $\bar{T}(\bar{T}+1)/2$ variance-covariance equations

$$Var(\widetilde{\eta}_t)=\gamma_t^2[\alpha_t^2Var(\Theta)+Var(U_t)]\ ,\ t=1,\ldots,\bar{T}$$

$$Cov(\widetilde{\eta}_t,\widetilde{\eta}_{t'})=\gamma_t\gamma_{t'}\alpha_t^2Var(\Theta)\ ,\ t=1,\ldots,\bar{T}-1\ ,\ t<t'$$

and $2\bar{T}$ unknowns $(\alpha_2,...,\alpha_{\bar{T}},\gamma_1,...,\gamma_{\bar{T}},Var(\Theta))$ for (a) and $2\bar{T}-1$ unknowns for (b), which implies the identification of $(\alpha_2,...,\alpha_{\bar{T}},\gamma_1,...,\gamma_{\bar{T}},Var(\Theta))$ for (a) when $\bar{T}\geq 3$ and for (b) when $\bar{T}=2$.

Since $0<|\alpha_t|,\gamma_t<\infty$ for all $t=1,...,\overline{T}$, and each element of $(G_1,...,G_{\overline{T}},H)$ has a non-vanishing characteristic function, we have that $F_{\widetilde{\eta}_1,...,\widetilde{\eta}_{\bar{T}}}$ has a non-vanishing characteristic function. The identification of the distributions $(G_1,...,G_{\overline{T}},H)$ follows by applying the result of Rao (1971) (cf. Kagan et al. (1973))[9] to the linear combination $\widetilde{\eta}_t=\gamma_t(\alpha_t\Theta_i-U_t)$ for $t=1,\ldots,\bar{T}$ with $\bar{T}\geq 2$. ■

Remark 1: In Theorem 4 and the similar results in the rest of this section, the assumption that each element of $(G_1,...,G_{\overline{T}},H)$ has a non-vanishing characteristic function can be replaced by the assumption that $F_{\widetilde{\eta}_1,...,\widetilde{\eta}_{\bar{T}}}$ has a non-vanishing characteristic function.

Following the same proof, it is easy to obtain the following result:

Corollary 2: Assuming all the conditions of Theorem 4 except that $Var(U_t)$ is now unknown, if $Var(U_t)=Var(U_{t'})$ for all t,t', then $(\alpha_2,...,\alpha_{\bar{T}},\gamma_1,...,\gamma_{\bar{T}},Var(\Theta),Var(U_t))$ and the distributions $(G_1,...,G_{\overline{T}},H)$ are identified when $\overline{T}\geq 4$ and $0<|\alpha_t|,\gamma_t<\infty$ for all $t=1,...,\overline{T}$.

Theorem 4 differs from the alternative factor identification results in Heckman-Taber (1994), and Cameron-Taber (1994). They do not require the existence of variances $Var(U_t)$ nor that the characteristic function for Θ is non-vanishing, but assume that the distributions of $U_{i,t}$, $t=1,\ldots,\bar{T}$, are known. They then rely on deconvolution to identify the distribution of Θ. The following theorem summarizes their results:

Theorem 5 (semiparametric): Assume

(i) $F_{\widetilde{\eta}}$ is identified.

[9] Also see the theorem 2.1.4 (page 12) and the remark 2.1.8 (page 16) in Prakasa Rao (1992).

(ii) $(U_1, ..., U_{\bar{T}})$ are mutually independent and have zero medians. U_t has completely known distribution (G_t) with non-vanishing characteristic function for $t = 1, \ldots, \bar{T}$.

(iii) Θ is independent of $(U_1, ..., U_{\bar{T}})$, has median zero, and its unknown distribution (H) has finite unknown variance.

Then:

(a) $(\alpha_2, ..., \alpha_{\bar{T}}, \gamma_1, ..., \gamma_{\bar{T}}, Var(\Theta))$ and the distribution H are identified when $\bar{T} \geq 3$ and $0 < |\alpha_t|, \gamma_t < \infty$ for all $t = 1, ..., \overline{T}$.

(b) If any one of $\alpha_2, \gamma_1, \gamma_2, Var(\Theta)$ is assumed to be known, then the rest of $\alpha_2, \gamma_1, \gamma_2, Var(\Theta)$, and the distribution H are identified when $\bar{T} = 2$ and $0 < |\alpha_2|, \gamma_1, \gamma_2 < \infty$.

(c) If U_1 and $X_1\beta_1$ both have full support R^1, and H has thinner tail than G_1, then γ_1 and the distribution H are identified when $\bar{T} = 1$ and $0 < \gamma_1 < \infty$

Proof. See Heckman-Taber (1994), Cameron-Taber (1994) ■

Remark 2: When $\bar{T} = 1$, Theorem 5 relies on the support condition to identify scale term γ_1, then uses deconvolution (since G_t is known) to identify the distribution H. Since Theorem 4 relies on the existence of variances to identify all the finite-dimensional parameters, and then applies Rao's (1971) identification results for linear combinations of independent random variables to recover the distributions $(G_1, ..., G_{\overline{T}}, H)$ jointly, one needs at least $\bar{T} \geq 2$ to apply Theorem 4.

4.2 Switching Regression Model

Now consider the identification of the factor structure for the switching regression model.

Corollary 3 (nonparametric): For the model defined by equations (3a) and (3b), assume that for $j = 0, 1$:

(i) $F_{\tilde{\eta}_1, \tilde{\eta}_{2j}}$ is identified.

(ii) (U_1, U_{2j}) are mutually independent with zero medians, finite known variances, and the unknown distributions (G_1, G_{2j}) have non-vanishing characteristic functions.

(iii) Θ is independent of (U_1, U_{2j}) with zero median, finite unknown variance, and its unknown distribution (H) has non-vanishing characteristic function.

Then: if any one of $\alpha_{2j}, \gamma_1, \gamma_{2j}, Var(\Theta)$ is known, then the rest of $\alpha_{2j}, \gamma_1, \gamma_{2j}, Var(\Theta)$, and the distributions (G_1, G_{2j}, H) are identified when $0 < |\alpha_{2j}|, \gamma_1, \gamma_{2j} < \infty$ for $j = 0, 1$.

Similarly, we can obtain the following

Corollary 4 (semiparametric): For the model defined by equations (3a) and (3b), assume that for $j = 0, 1$,

(i) $F_{\tilde{\eta}_1,\tilde{\eta}_{2j}}$ is identified.

(ii) (U_1, U_{2j}) are mutually independent with zero medians, known distributions with non-vanishing characteristic functions.

(iii) Θ is independent of (U_1, U_{2j}), and its unknown distribution (H) has zero median and finite unknown variance.

Then: if any one of $\alpha_{2j}, \gamma_1, \gamma_{2j}, Var(\Theta)$ is known, then the rest of $\alpha_{2j}, \gamma_1, \gamma_{2j}, Var(\Theta)$, and the distribution H are identified when $0 < |\alpha_{2j}|, \gamma_1, \gamma_{2j} < \infty$ for $j = 0, 1$.

Proof. The proofs for Corollaries 3 and 4 follow from the same strategies as those for Theorems 4 and 5. ■

5 $\sqrt{N}$ Estimableness of Structure Parameters for Models with Scalar Latent Factor

>From the last two sections and under some sufficient conditions, we know that the distributions $F_{\widetilde{\eta}_1,...,\widetilde{\eta}_{\overline{T}}}$, $G_1, ..., G_{\overline{T}}$, and H are nonparametrically identified. Thus, one could estimate $(\beta_1, ..., \beta_{\overline{T}}, \alpha_2, ..., \alpha_{\overline{T}}, G_1, ..., G_{\overline{T}}, H)$ by NPMLE and investigate the $\sqrt{N}$ estimableness of $(\beta_1, ..., \beta_{\overline{T}}, \alpha_2, ..., \alpha_{\overline{T}})$ by treating $(G_1, ..., G_{\overline{T}}, H)$ as the infinite-dimensional nuisance parameters. However, this is computationally too complicated and unstable when $\overline{T}$ is large (say $\overline{T} \geq 3$). In practice, people assume that the distributions $(G_1, ..., G_{\overline{T}})$ are known (typically standard normal or logistic), and estimate $(\beta_1, ..., \beta_{\overline{T}}, \alpha_2, ..., \alpha_{\overline{T}}, H)$ by NPMLE. In this section, we study the $\sqrt{N}$ estimableness of $(\beta_1, ..., \beta_{\overline{T}}, \alpha_2, ..., \alpha_{\overline{T}})$ by treating H as the only infinite-dimensional nuisance parameter for the three models with different $\overline{T}$.

5.1 Semiparametric Efficient Scores for Mixture Models

Let $P_{\beta,\alpha,h}$ denote the probability associated with individual i's data $(D_i, X_i) \equiv \{D_{i,t}, X_{i,t} : t = 1, ..., \bar{T}\}$ if it were generated by parameter (β, α, H). P_o is the probability associated with individual i's data $\{D_{i,t}, X_{i,t} : t = 1, ..., \bar{T}\}$ when it is generated by the true parameter (β_o, α_o, H_o), and μ is a dominating measure (e.g., a product of counting measures and Lebesgue measures). For the above three discrete-choice models with a nonparametric latent facor, we assume that $X_i = (X_{i,1}, ..., X_{i,\overline{T}})$ has the same distribution as that of X, denoted as F_X with density f_X , which is not a function of (β, α, H). Then we have the following probability density structures:

$$\frac{dP_{\beta,\alpha,h}}{d\mu}(D_i, X_i) = \int \wp(\beta, \alpha, D_i, X_i | \Theta_i = \theta) dH(\theta),$$

where the latent facor distribution H is called "mixing distribution" which is unknown, and $(d, x) \rightarrow \wp(\beta, \alpha, D_i, X_i|\Theta_i = \theta)$ is called "kernel or mixture density" which is known up to finite-dimentional parameters (β, α) and the unknown marginal density f_X of X_i. Under our asssumptions on X_i, we have

$$\wp(\beta, \alpha, D_i, X_i|\Theta_i = \theta) \equiv f_X(X_i)\wp(\beta, \alpha, D_i|X_i, \Theta_i = \theta)$$

where $\wp(\beta, \alpha, D_i|X_i, \Theta_i = \theta)$ is the conditional density of $D_i|X_i, \Theta_i$, which has known functional forms up to unknown parameters (β, α) in our three semiparametric mixture discrete-choice panel models with a latent factor.
The unconditional log-likelihood for person i is denoted as:

$$l(\beta, \alpha, H, D_i, X_i) \equiv \log \left(\int \wp(\beta, \alpha, D_i, X_i|\Theta_i = \theta) dH(\theta) \right).$$

The ordinary score functions are:

$$\begin{aligned} & l'_\beta(\beta_o, \alpha_o, H_o; D_i, X_i) \\ = \quad & \frac{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) \frac{\partial}{\partial\beta}[\log(\wp(\beta_o, \alpha_o, D_i, X_i|\theta))] dH_o(\theta)}{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) dH_o(\theta)} \end{aligned}$$

$$\begin{aligned} & l'_\alpha(\beta_o, \alpha_o, H_o; D_i, X_i) \\ = \quad & \frac{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) \frac{\partial}{\partial\alpha}[\log(\wp(\beta_o, \alpha_o, D_i, X_i|\theta))] dH_o(\theta)}{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) dH_o(\theta)} \end{aligned}$$

$$l'_H(\beta_o, \alpha_o, H_o; D_i, X_i)f = \frac{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) f(\theta) dH_o(\theta)}{\int \wp(\beta_o, \alpha_o, D_i, X_i|\theta) dH_o(\theta)}.$$

Here l'_H is the score for the infinite-dimensional nuisance parameter H. It transforms scores for the unknown probability measure (H) into scores for the model of observations $(\{D_i, X_i\})$. Denote $l'_{H_o}(D_i, X_i) \equiv l'_H(\beta_o, \alpha_o, H_o; D_i, X_i)$, $l'_{\beta_o}(D_i, X_i) \equiv l'_\beta(\beta_o, \alpha_o, H_o; D_i, X_i)$, $l'_{\alpha_o}(D_i, X_i) \equiv l'_\alpha(\beta_o, \alpha_o, H_o; D_i, X_i)$.

The *efficient score* function for (β_o, α_o) is defined as the ordinary score function for (β_o, α_o) minus its L_2-orthogonal projection onto the closed linear span (clsp) of the score functions for the nuisance parameter H:

$$\begin{bmatrix} \mathcal{S}_{\beta_o}(D_i, X_i) \\ \mathcal{S}_{\alpha_o}(D_i, X_i) \end{bmatrix} \equiv \begin{bmatrix} l'_{\beta_o}(D_i, X_i) \\ l'_{\alpha_o}(D_i, X_i) \end{bmatrix} - E_o \left\{ \begin{bmatrix} l'_{\beta_o}(D_i, X_i) \\ l'_{\alpha_o}(D_i, X_i) \end{bmatrix} |\text{clsp}\left\{ l'_{H_o}(D_i, X_i) \right\} \right\}.$$

Here $E_o\left\{\left[\begin{array}{c} l'_{\beta_o}(D_i,X_i)\\ l'_{\alpha_o}(D_i,X_i)\end{array}\right] |\text{clsp}\left\{l'_{H_o}(D_i,X_i)\right\}\right\}$ solves the infinite-dimensional optimization problem (the $*$ notation means the transpose):

$$\inf_{M\in\text{clsp}\left\{l'_{H_o}(D_i,X_i)f:f\in\mathcal{F}_o\right\}} E_o\left\{\left(\left[\begin{array}{c} l'_{\beta_o}(D_i,X_i)\\ l'_{\alpha_o}(D_i,X_i)\end{array}\right]-M\right)^{*}\left(\left[\begin{array}{c} l'_{\beta_o}(D_i,X_i)\\ l'_{\alpha_o}(D_i,X_i)\end{array}\right]-M\right)\right\},$$

where $\mathcal{F}_o$ denotes a Hilbert space of bounded functions (with respect to $L_2(dH_o)$) with $\int f dH_o = 0$ and $E_o(\cdot)$ denotes the expectation under true parameter (β_o,α_o,H_o). The *efficient information matrix* (evaluated at (β_o,α_o,H_o)) is simply the expectation of the outer product of the efficient score matrix,

$$\mathcal{I} = E_o\left\{\left[\begin{array}{c} \mathcal{S}_{\beta_o}(D_i,X_i)\\ \mathcal{S}_{\alpha_o}(D_i,X_i)\end{array}\right]\left[\begin{array}{cc} \mathcal{S}^{*}_{\beta_o}(D_i,X_i) & \mathcal{S}^{*}_{\alpha_o}(D_i,X_i)\end{array}\right]\right\}.$$

(β_o,α_o) is $\sqrt{N}$-efficiently *estimable if and only if* $\mathcal{I}$ *is non-singular*, i.e., if $\mathcal{S}_{\beta_o}$, $\mathcal{S}_{\alpha_o}$ are linearly independent (see Van der Vaart, 1991).

It is generally difficult to find explicit expressions for $\mathcal{S}_{\beta_o}$ and $\mathcal{S}_{\alpha_o}$. We next focus on an important class of mixture models by assuming that there exists a "quasi-statistics" $\psi(D,X;\beta,\alpha)$ which is sufficient for Θ (i.e. H) given a fixed value of (β,α). Now by Van der Vaart (1996, page 868), we have the following expressions for efficient scores:

$$\mathcal{S}_{\beta_o} = \frac{\int \wp_o(D,X|\theta)\left\{\frac{\partial}{\partial\beta}[\log(\wp_o(D,X|\theta))] - E\left(\frac{\partial}{\partial\beta}[\log(\wp_o(D,X|\theta))]|\psi(D,X;\beta,\alpha)\right)\right\}dH_o(\theta)}{\int \wp(\beta_o,\alpha_o,D_i,X_i|\theta)dH_o(\theta)} \tag{3}$$

$$\mathcal{S}_{\alpha_o} = \frac{\int \wp_o(D,X|\theta)\left\{\frac{\partial}{\partial\alpha}[\log(\wp_o(D,X|\theta))] - E\left(\frac{\partial}{\partial\alpha}[\log(\wp_o(D,X|\theta))]|\psi(D,X;\beta,\alpha)\right)\right\}dH_o(\theta)}{\int \wp(\beta_o,\alpha_o,D_i,X_i|\theta)dH_o(\theta)} \tag{4}$$

Since $\psi(D,X;\beta,\alpha)$ depends on unknown parameters of interest (β,α), it is not the conventional sufficient statistics for the nuicance parameter H. Nevertheless, the above expressions will allows us to present simple sufficient conditions to ensure positive definiteness of the Fisher information matrix $\mathcal{I}$ for widely applied econometrics models. As an illustration, we consider the three models with U_t being logistic $G_t(y) = \frac{\exp(y)}{1+\exp(y)}$ for all $t = 1,...,\overline{T}$. The more general case may be found in Chen et al. (1998).

5.2 Repeated Binary Choice Model

The mixture density associated with the repeated binary choice model (1a) and (1b) is:

$$\wp(\beta,\alpha,D_i,X_i|\Theta_i=\theta) \equiv f_X(X_i)\prod_{t=1}^{\bar{T}}\left([G_t(X_{i,t}\beta_t+\alpha_t\theta)]^{D_{i,t}}[1-G_t(X_{i,t}\beta_t+\alpha_t\theta)]^{1-D_{i,t}}\right)$$

In the following we denote $G_t(X_{i,t}\beta_t + \alpha_t\theta) = G_t$ and $\wp_o(D|X,\theta) = \wp(\beta_o, \alpha_o, D|X, \Theta = \theta)$, where

$$\wp(\beta, \alpha, D_i|X_i, \Theta_i = \theta) \equiv \prod_{t=1}^{\bar{T}} \left([G_t(X_{i,t}\beta_t + \alpha_t\theta)]^{D_{i,t}} [1 - G_t(X_{i,t}\beta_t + \alpha_t\theta)]^{1-D_{i,t}} \right)$$

Lemma 1 (repeated binary): For the model defined by equations (1a) and (1b) under the normalization (LN) and (SN2), let all the conditions of Theorem 5 be satisfied with $G_t(y) = \frac{\exp(y)}{1+\exp(y)}$ for $t = 1, ..., \overline{T}$. For simplicity we assume that $\widetilde{\beta}_t \neq \widetilde{\beta}_{t'}$, $\gamma_t \neq \gamma_{t'}$ for $t \neq t'$. Then the efficient scores for $(\widetilde{\beta}_1, ..., \widetilde{\beta}_{\overline{T}}, \gamma_1, ..., \gamma_{\overline{T}}, \alpha_2, ..., \alpha_{\bar{T}})$ are:

$$\mathcal{S}_{\widetilde{\beta}^{(1)}_{t,o}}(D_i, X_i) = \frac{1}{\gamma_{t,o}} \left(D_t - E[D_t \mid \sum_{j=1}^{\overline{T}} \alpha_j D_j, \ X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right)$$

$$\begin{aligned} \mathcal{S}_{\widetilde{\beta}^{(-1)}_{t,o}}(D_i, X_i) \quad &= \quad \frac{1}{\gamma_{t,o}} \left(D_t X_t^{(-1)} - E[D_t X_t^{(-1)} \mid \sum_{j=1}^{\overline{T}} \alpha_j D_j, \ X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right) \\ & \quad - \frac{1}{\gamma_{t,o}} \frac{\int G_t \wp_o(D|X,\theta) dH_o(\theta)}{\int \wp_o(D|X,\theta) dH_o(\theta)} \left(X_t^{(-1)} - E[X_t^{(-1)} | \sum_{j=1}^{\overline{T}} \alpha_j D_j, \ X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right) \end{aligned}$$

$$\mathcal{S}_{\gamma_{t,o}}(D_i, X_i) = -\frac{1}{\gamma_{t,o}^2} X_t \widetilde{\beta}_{t,o} \left\{ D_t - E[D_t \mid \sum_{j=1}^{\overline{T}} \alpha_j D_j, \ X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right\}.$$

$$\mathcal{S}_{\alpha_{t,o}}(D, X) = \frac{\int \theta \wp_o(D|X,\theta) dH_o(\theta)}{\int \wp_o(D|X,\theta) dH_o(\theta)} \left\{ D_t - E[D_t \mid \sum_{j=1}^{\overline{T}} \alpha_j D_j, \ X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right\}.$$

Proof. Since

$$\begin{aligned} & \wp(\beta, \alpha, D, X|\Theta) \\ = \quad & \exp\left\{ \sum_{t=1}^{\overline{T}} D_t X_t \widetilde{\beta}_t \gamma_t^{-1} + \Theta \sum_{t=1}^{\overline{T}} D_t \alpha_t - \sum_{t=1}^{\overline{T}} \log[1 + \exp(X_t\widetilde{\beta}_t\gamma_t^{-1} + \alpha_t\Theta)] \right\} f_X(X) \end{aligned}$$

then $\left\{ \sum_{t=1}^{\overline{T}} D_t\alpha_t, \ [X_1\widetilde{\beta}_1, ..., X_{\overline{T}}\widetilde{\beta}_{\overline{T}}] \right\}$ may be chosen as the “quasi sufficient statistics $\psi(D, X; \beta, \alpha)$ for Θ. Now the result follows from (3) and (4) ■

Theorem 6 (repeated binary, $\overline{T} = 1$): Assume all the conditions for Lemma 1 hold for model (1a)-(1b) when $\overline{T} = 1$. Then:

(a) $\widetilde{\beta}_1^{(-1)}$ is $\sqrt{N}$ estimable if and only if $E[X_1^{(-1)}|X_1\widetilde{\beta}_{1,o}, D_1] \neq X_1^{(-1)}$, which is satisfied if $\dim(X_1^{(-1)}) \geq 2$.

(b) $\widetilde{\beta}_1^{(1)}$ and γ_1 are not $\sqrt{N}$ estimable.

Remark 3. When $\overline{T} = 1$, although the intercept $(\widetilde{\beta}_1^{(1)})$ is identified, the scale (γ_1) is not identified in general but is identified if H has bounded support or has thinner tails than G_1 (see Theorem 5 part(c)). Nevertheless, neither intercept nor scale are $\sqrt{N}$ estimable according to Theorem 6, and we need to impose exactly the same set of conditions (i.e., normalizing scale (γ_1) and intercept $(\widetilde{\beta}_1^{(1)})$) as those in Chamberlain (1986) or Cosslett (1987) to ensure $\sqrt{N}$ consistent rates for all the slope coefficients.

Theorem 7 (repeated binary, $\overline{T} = 2$): Assume all the conditions for Lemma 1 hold for model (1a)-(1b) when $\overline{T} = 2$. Then:

(a) when $|\alpha_2| = 1$, $(\widetilde{\beta}_1^{(-1)}, \widetilde{\beta}_2^{(-1)})$ are $\sqrt{N}$ estimable.

(b) when $|\alpha_2| \neq 1$, $\widetilde{\beta}_t^{(-1)}$ are $\sqrt{N}$ estimable if and only if $E[X_t^{(-1)}|X_1\widetilde{\beta}_{1,o}, X_2\widetilde{\beta}_{2,o}, D_1 + \alpha_2 D_2] \neq X_t^{(-1)}$ for $t = 1, 2$, which is satisfied if $\dim(X_t^{(-1)}) \geq 2$.

(c) if $\widetilde{\beta}_1^{(1)}$ (or $\widetilde{\beta}_2^{(1)}$) is known, then $\widetilde{\beta}_2^{(1)}$ (or $\widetilde{\beta}_1^{(1)}$) is $\sqrt{N}$ estimable if and only if $|\alpha_2| = 1$.

(d) (γ_1, γ_2) are $\sqrt{N}$ estimable if and only if $|\alpha_2| = 1$.

(e) assuming $E[\Theta|D, X] \neq 0$ almost surely, then α_2 is $\sqrt{N}$ estimable if and only if $|\alpha_2| = 1$.

Remark 4. When $\overline{T} = 2$, by Theorem 5 part (b), we need to assume one of γ_1, γ_2, α_2 is known or to assume H has thinner tails than G_1 for the sake of identification. By Theorem 7, the most commonly used normalization $\alpha_2 = 1$ also implies $\sqrt{N}$ convergence rates for all the slope coefficients without scale normalization.

Theorem 8 (repeated binary, $\overline{T} \geq 3$): Assume all the conditions for Lemma 1 hold for model (1a)-(1b) when $\overline{T} \geq 3$. Then:

(a) $(\widetilde{\beta}_1^{(-1)}, ..., \widetilde{\beta}_{\overline{T}}^{(-1)})$ are $\sqrt{N}$ estimable.

(b) if one of $\widetilde{\beta}_t^{(1)}$ is known, then all the other $\widetilde{\beta}_{t'}^{(1)}$, $t' \neq t$, are $\sqrt{N}$ estimable for all $t, t' = 1, ..., \overline{T}$

(c) $(\gamma_1, ..., \gamma_{\overline{T}})$ are $\sqrt{N}$ estimable.

(e) assuming $E[\Theta|D, X] \neq 0$ almost surely, then $(\alpha_2, ..., \alpha_{\overline{T}})$ are $\sqrt{N}$ estimable.

Remark 5. By well-known existing results such as Chamberlain (1986) or Cosslett (1987) on cross-sectional binary choice models, one can easily guess that all the normalized slope coefficients $(\widetilde{\beta}_t^{(-1)})$ at each time periods can be estimated at root-N rates. However what is surprising from Theorems 6, 7 and 8 is that the scale parameters, the factor loading parameters can also be estimated at root-N rates when $\overline{T} \geq 3$, and all the other time periods intercept parameters can be estimated at root-N rates after normalizing the first period intercept.

Proof. First consider $\overline{T} = 1$. Recall that $\alpha_1 = 1$. From the expression for the efficient scores, we have $\mathcal{S}_{\widetilde{\beta}_1^{(1)}} = 0$ and $\mathcal{S}_{\gamma_{1,o}} = 0$, hence $\widetilde{\beta}_1^{(1)}$ and γ_1 are not $\sqrt{N}$ estimable. Since

$$\mathcal{S}_{\widetilde{\beta}_1^{(-1)}} = \frac{1}{\gamma_{1,o}} \left[D_1 - \frac{\int G_1 \wp_o(D|X, \theta) dH_o(\theta)}{\int \wp_o(D|X, \theta) dH_o(\theta)} \right] \left\{ X_1^{(-1)} - E[X_1^{(-1)} \mid X_1 \widetilde{\beta}_{1,o}, D_1] \right\},$$

we have that $E_o[\mathcal{S}_{\widetilde{\beta}_1^{(-1)}} \mathcal{S}^*_{\widetilde{\beta}_1^{(-1)}}]$ is positive definite when $\dim(X_1^{(-1)}) \geq 2$, hence $\widetilde{\beta}_1^{(-1)}$ is $\sqrt{N}$ estimable.

Now for any $\overline{T} \geq 2$ we notice that

$$\mathcal{S}_{\widetilde{\beta}_{t,o}^{(1)}}(D_i, X_i) = \frac{1}{\gamma_{t,o}} \left(D_t - E[D_t \mid \sum_{j=1}^{\overline{T}} \alpha_j D_j,\ X_1 \widetilde{\beta}_1, ..., X_{\overline{T}} \widetilde{\beta}_{\overline{T}}] \right)$$

hence

$$\sum_{j=1}^{\overline{T}} \gamma_{j,o} \alpha_j \mathcal{S}_{\widetilde{\beta}_{j,o}^{(1)}} = 0$$

thus we need to assume that at least one of $\widetilde{\beta}_t^{(1)}$, $t = 1, 2, ..., \overline{T}$ is known to get rid of the linearly dependence.

Now when $\overline{T} = 2$, $D_t - E[D_t \mid X_t \widetilde{\beta}_{t,o}, D_1 + \alpha_2 D_2] \neq 0$ if and only if $|\alpha_2| = 1$. Under the assumed conditions and $|\alpha_2| = 1$, we have that $\mathcal{S}_{\widetilde{\beta}_1^{(-1)}}, \mathcal{S}_{\widetilde{\beta}_2^{(-1)}}, \mathcal{S}_{\gamma_1,o}, \mathcal{S}_{\gamma_2,o}, \mathcal{S}_{\alpha_2,o}, \mathcal{S}_{\widetilde{\beta}_2^{(1)}}$ are linearly independent, hence $(\widetilde{\beta}_1^{(-1)}, \widetilde{\beta}_2^{(-1)}, \gamma_1, \gamma_2, \alpha_2, \widetilde{\beta}_2^{(1)})$ are $\sqrt{N}$ estimable when $|\alpha_2| = 1$. Now when $|\alpha_2| \neq 1$, $\mathcal{S}_{\gamma_1,o}, \mathcal{S}_{\gamma_2,o}, \mathcal{S}_{\alpha_2,o}, \mathcal{S}_{\widetilde{\beta}_2^{(1)}} = 0$, and thus $(\gamma_1, \gamma_2, \alpha_2, \widetilde{\beta}_2^{(1)})$ are not $\sqrt{N}$ estimable when $|\alpha_2| \neq 1$. However in this case, for $t = 1, 2$,

$$\mathcal{S}_{\widetilde{\beta}_t^{(-1)}} = \frac{1}{\gamma_{t,o}} \left[D_t - \frac{\int G_t \wp_o(D|X, \theta) dH_o(\theta)}{\int \wp_o(D|X, \theta) dH_o(\theta)} \right] \left\{ X_t^{(-1)} - E[X_t^{(-1)} \mid X_1 \widetilde{\beta}_{1,o}, X_2 \widetilde{\beta}_{2,o}, D_1 + \alpha_2 D_2] \right\}.$$

Thus $\mathcal{S}_{\widetilde{\beta}_1^{(-1)}}, \mathcal{S}_{\widetilde{\beta}_2^{(-1)}}$ are still linearly independent, hence $(\widetilde{\beta}_1^{(-1)}, \widetilde{\beta}_2^{(-1)})$ are still $\sqrt{N}$ estimable when $|\alpha_2| \neq 1$.

When $\overline{T} \geq 3$, in general we have $\mathcal{S}_{\widetilde{\beta}_1^{(-1)}}, ..., \mathcal{S}_{\widetilde{\beta}_{\overline{T}}^{(-1)}}, \mathcal{S}_{\gamma_1,o}, ..., \mathcal{S}_{\gamma_{\overline{T}},o}, \mathcal{S}_{\alpha_2,o}, ..., \mathcal{S}_{\alpha_{\overline{T}},o}, \mathcal{S}_{\widetilde{\beta}_2^{(1)}}, ..., , \mathcal{S}_{\widetilde{\beta}_{\overline{T}}^{(1)}}$ are linearly independent, hence $(\widetilde{\beta}_1^{(-1)}, ..., \widetilde{\beta}_{\overline{T}}^{(-1)}, \gamma_1, ..., \gamma_{\overline{T}}, \alpha_2, ..., \alpha_{\overline{T}}, \widetilde{\beta}_2^{(1)}, ..., \widetilde{\beta}_{\overline{T}}^{(1)})$ are all $\sqrt{N}$ estimable. ■

5.3 Single Spell Duration Model

Denote $T_i = \sum_{t=1}^{\overline{T}} D_{i,t}$ as total completed spell length for individual i. Let $P_{\beta,\alpha,h}$ denote the probability associated with individual i's data $\{T_i, X_{i,1}, ..., X_{i,\bar{T}}\}$ if it were generated by parameter (β, α, H). P_o is the probability associated with individual i's data $\{T_i, X_{i,1}, ..., X_{i,\bar{T}}\}$ when it is generated by the true parameter (β_o, α_o, H_o), and μ is a dominating measure (i.e., a product of counting measures and Lebesgue measures). Recall that X_i has distribution F_X (and density f_X) which is not a function of (β, α, h).

Let $\delta_k(t) = 1$ if $t = k$; $\delta_k(t) = 0$ if $t \neq k$. Then for $k \in \{0, 1, .., \bar{T}\}$, $x \in \text{supp}(f_X)$,

$$\frac{dP_{\beta,\alpha,h}}{d\mu}(k, x) = \prod_{t=0}^{\bar{T}} [\Pr(T = t | X \leq x)]^{\delta_k(t)} f_X(x).$$

Person i 's probability of surviving $T_i = t$ periods conditional on X_i is:

$$\Pr(T_i = t | X_i) = \int Pr(T_i = t | X_i, \theta) dH(\theta).$$

We adopt the convention that all individuals are in the sample at time 0,

$$\prod_{j=1}^{0} G_j(x_{i,j}\beta_j + \alpha_j \theta) \equiv 1,$$

so that

$$Pr(T_i = 0 | X_i, \Theta_i) \equiv Pr(D_{i,1} = 0 | D_{i,0} = 1, X_{i,1}, \Theta_i) \equiv 1 - G_1(X_{i,1}\beta_1 + \alpha_1 \Theta_i).$$

We also adopt the convention

$$G_{\overline{T}+1}(x_{i,\overline{T}+1}\beta_{\overline{T}+1} + \alpha_{\overline{T}+1}\theta) \equiv 0,$$

since

$$Pr(D_{i,\overline{T}+1} = 0 | D_{i,\overline{T}} = 1, X_{i,\overline{T}+1}, \Theta_i) \equiv 1.$$

Person i 's probability of surviving $T_i = t$ periods conditional on $(X_i = x_i \ , \ \Theta_i = \theta)$ is then

$$\begin{aligned}
& Pr(T_i = t|X_i = x_i, \Theta_i = \theta) \\
= & \left[\prod_{j=1}^{t} Pr(D_{i,j} = 1|D_{i,j-1} = 1, x_{i,j}, \theta)\right] Pr(D_{i,t+1} = 0|D_{i,t} = 1, x_{i,t+1}, \theta) \\
= & \left[\prod_{j=1}^{t} G_j(x_{i,j}\beta_j + \alpha_j\theta)\right] (1 - G_{t+1}(x_{i,t+1}\beta_{t+1} + \alpha_{t+1}\theta)).
\end{aligned}$$

The probability (or density) of T_i conditional on $\{X_i, \Theta_i = \theta\}$ is denoted as:

$$\begin{aligned}
\wp(\beta, \alpha, T_i|X_i, \Theta_i = \theta) &= \prod_{t=0}^{\bar{T}} [Pr(T_i = t|X_i, \Theta_i = \theta)]^{\delta_{T_i}(t)} \\
&= \prod_{t=0}^{\bar{T}} \left[\prod_{j=1}^{t} G_j(X_{i,j}\beta_j + \alpha_j\theta)[1 - G_{t+1}(X_{i,t+1}\beta_{t+1} + \alpha_{t+1}\theta)]\right]^{\delta_{T_i}(t)} .
\end{aligned}$$

When $G_t(y) = \frac{\exp(y)}{1+\exp(y)}$ for $t = 1, ..., \overline{T}$, we can again write down the efficient score functions explicitly. Since the formula is complicated, we put it as Lemma 2 in the Appendix. We can obtain the following results as an immediate application of Lemma 2:

Theorem 7 (single spell): For the model defined by equations (2a) and (2b), let the conditions for Theorem 5 be satisfied with $G_t(y) = \frac{\exp(y)}{1+\exp(y)}$ for $t = 1, ..., \overline{T}$. For simplicity we assume that all β_t are distinct and that X_t is independent across time or first-order Markov process. Then:

(a) when $\overline{T} \geq 1$, $(\widetilde{\beta}_1, ..., \widetilde{\beta}_{\overline{T}})$ are root-N estimable.

(b) when $\overline{T} = 1$, γ_1 is not root-N estimable.

(c) when $\overline{T} = 2$, (γ_1, γ_2) are root-N estimable if and only if $\alpha_2 = -1$

(d) when $\overline{T} = 3$, $(\gamma_1, \gamma_2, \gamma_3, \alpha_2, \alpha_3)$ are root-N estimable if any one of the following is satisfied:

$$1 + \alpha_2 = 0 \text{ or } 1 + \alpha_2 + \alpha_3 = 0 \text{ or } \alpha_2 + \alpha_3 = 0.$$

Proof. Similar proof as that for Theorem 6 except using Lemma 2 in the Appendix. ■

5.4 Switching Binary Regression Model

Let $P_{\beta,\alpha,h}$ denote the probability associated with individual i's data $(D_i, X_i) \equiv \{D_{i,t}, X_{i,t} : t = 1, 2\}$ if it were generated by parameter (β, α, H). P_o is the probability associated with individual i's data $\{D_{i,t}, X_{i,t} : t = 1, 2\}$ when it is generated by the true parameter (β_o, α_o, H_o), and μ is a dominating measure (e.g., a product of counting measures and Lebesgue measures). We have for any $d \in \prod_{t=1}^{2} \{0, 1\}$, $x \in \text{supp}(f_X)$,

$$\frac{dP_{\beta,\alpha,h}}{d\mu}(d, x) = f_X(x) \int \{\wp(\beta, \alpha, D_i|X_i = x, \Theta_i = \theta)\} \, dH(\theta),$$

where $\wp(\beta, \alpha, D_i|X_i = x, \Theta_i = \theta)$ is the probability (or density) of D_i conditional on $\{X_i, \Theta_i\}$ and is given by:

$$\begin{aligned} \wp(\beta, \alpha, D_i|X_i &= x_i, \Theta_i = \theta) \\ &\equiv \left[G_1(x_{i,1}\beta_1 + \theta)G_{21}(x_{i,2}\beta_{2,1} + \alpha_{2,1}\theta)\right]^{D_{i,1}D_{i,2}} \\ &\quad \times \left[G_1(x_{i,1}\beta_1 + \theta)\left(1 - G_{21}(x_{i,2}\beta_{2,1} + \alpha_{2,1}\theta)\right)\right]^{D_{i,1}(1-D_{i,2})} \\ &\quad \times \left[(1 - G_1(x_{i,1}\beta_1 + \theta))\, G_{20}(x_{i,2}\beta_{2,0} + \alpha_{2,0}\theta)\right]^{(1-D_{i,1})D_{i,2}} \\ &\quad \times \left[(1 - G_1(x_{i,1}\beta_1 + \theta))\left(1 - G_{20}(x_{i,2}\beta_{2,0} + \alpha_{2,0}\theta)\right)\right]^{(1-D_{i,1})(1-D_{i,2})} . \end{aligned}$$

Again when $G_1(y), G_{21}(y), G_{20}(y) = \frac{\exp(y)}{1+\exp(y)}$, we can write down the efficient score functions explicitly. We again collect the formula as Lemma 3 into the Appendix, and obtain the following results as a consequence of Lemma 3:

Theorem 8 (switching regression): For the model defined by equations (3a) and (3b), let the conditions for Corollary 4 be satisfied with $G_1(y), G_{21}(y), G_{20}(y) = \frac{\exp(y)}{1+\exp(y)}$. For simplicity we assume that X_t is independent across time or first-order Markov process. Then:

(a) $(\widetilde{\beta}_1, \widetilde{\beta}_{21}, \widetilde{\beta}_{20})$ are root-N estimable.

(b) $(\gamma_1, \gamma_{21}, \gamma_{20}, \alpha_{21}, \alpha_{20})$ are root-N estimable if any one of the following is satisfied:

$$\alpha_{21} = -1 \text{ or } \alpha_{20} = 1 \text{ or } \alpha_{20} = 1 + \alpha_{21}.$$

Proof. Similar proof as that for Theorem 6 except using Lemma 3 in the Appendix. ■

6 Properties of NPMLE for Models with Scalar Factor

We now consider estimation of the true parameters $(\beta_o, \alpha_o, H_o) \in \mathcal{B} \times \mathcal{A} \times \mathcal{H}$ for the three models with scalar latent factor, where $\mathcal{B}$ and $\mathcal{A}$ are finite-dimensional Euclidean spaces, and $\mathcal{H}$ is the space of distributions with zero mean and finite variances. We will first consider the following distance on the parameter spaces $\mathcal{B} \times \mathcal{A} \times \mathcal{H}$:

$$\begin{aligned}\|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|^2 &= \frac{1}{2}\int\left(\sqrt{dP_{\beta,\alpha,h}}-\sqrt{dP_o}\right)^2=\frac{1}{2}\int\left(\sqrt{\frac{dP_{\beta,\alpha,h}}{d\mu}}-\sqrt{\frac{dP_o}{d\mu}}\right)^2 d\mu\\ &= 1-E_o\sqrt{\frac{dP_{\beta,\alpha,h}}{dP_o}},\end{aligned}$$

which is the Hellinger distance directly on the observed probabilities. Denote the L_1-norm between $P_{\beta,\alpha,h}$ and P_o as

$$\|P_{\beta,\alpha,h}-P_o\|_1=\sup_{g:|g|\leq 1}\left|\int g(dP_{\beta,\alpha,h}-dP_o)\right|.$$

Then (see e.g., Le Cam and Yang 1990, page 25):

$$\|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|^2\leq\frac{1}{2}\|P_{\beta,\alpha,h}-P_o\|_1\leq\|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|\sqrt{2}.$$

Denote the Kullback-Leibler distance as

$$\begin{aligned}KL((\beta_o,\alpha_o,H_o),(\beta,\alpha,H)) &= E_o\left[l(\beta_o,\alpha_o,H_o;D_i,X_i)-l(\beta,\alpha,H;D_i,X_i)\right]\\ &= E_o\left[\log\left(\frac{dP_o}{dP_{\beta,\alpha,h}}\right)\right].\end{aligned}$$

Since $\log(1+y)\leq y$ for all $y\geq -1$, we have by taking $y=\sqrt{\frac{dP_{\beta,\alpha,h}}{dP_o}}-1$,

$$KL((\beta_o,\alpha_o,H_o),(\beta,\alpha,H))\geq 2\|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|^2.$$

6.1 Convergence Rate of NPMLE

Recall that the nonparametric MLE (NPMLE) optimizes the sample likelihood over the entire parameter space without any smoothing, i.e., the NPMLE $(\widehat{\beta}_N,\widehat{\alpha}_N,\widehat{H}_N)$ solves

$$\max_{\beta\in\mathcal{B},\alpha\in\mathcal{A},H\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}l(\beta,\alpha,H;D_i|X_i).$$

NPMLE is known to be consistent and has optimal rate if the observed density is bounded away from zero and if the $\mathcal{H}$ space is not too large.

Theorem 9: Suppose that conditions for Theorems 6, 7 and 8 are satisfied. Let $\mathcal{B}$ and $\mathcal{A}$ be finite-dimensional compact sets, and $\mathcal{H}$ be the space of probability distributions with bounded supports. Then: $\left\|(\widehat{\beta}_N,\widehat{\alpha}_N,\widehat{H}_N)-(\beta_o,\alpha_o,H_o)\right\|=O_p(N^{-1/3})$.

Proof. See Appendix C. ■

6.2 $\sqrt{N}-$Asymptotic Normality

Under the conditions for Theorems 6, 7 and 8, we know that the finite-dimensional parameter (β_o, α_o) is $\sqrt{N}-$estimable. We now present sufficient conditions to establish that the NPMLE estimator $(\widehat{\beta}_N, \widehat{\alpha}_N)$ is $\sqrt{N}-$ asymptotic normal centered around (β_o, α_o) with asymptotic variance $\mathcal{I}^{-1}$, where $\mathcal{I}$ is the expectation of outer-product of efficient scores for each of the models, (see Section 4), hence it is efficient estimator also.

There are many ways to establish such results. Here we follow the approaches taken by Shen (1997) and Chen and Shen (1998). We first define the following norm at the neighborhood of (β_o, α_o, H_o):

$$\begin{aligned}
& \|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|_e^2 \\
= \ & E_o\left[l'_{\beta,o}(D,X)(\beta-\beta_o)+l'_{\alpha,o}(D,X)(\alpha-\alpha_o)+l'_{H,o}(D,X)[H-H_o]\right]^2.
\end{aligned}$$

Let $[H-H_o]=-\left[\begin{array}{c} W_\beta(\beta-\beta_o) \\ W_\alpha(\alpha-\alpha_o)\end{array}\right]$, then

$$\begin{aligned}
& \|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|_e^2 \\
= \ & \left[\begin{array}{c} \beta-\beta_o \\ \alpha-\alpha_o \end{array}\right]^* E_o\left(\left[\begin{array}{c} l'_{\beta_o}(D,X)-l'_{H_o}(D,X)W_\beta \\ l'_{\alpha_o}(D,X)-l'_{H_o}(D,X)W_\alpha \end{array}\right]\left[\begin{array}{c} l'_{\beta_o}(D,X)-l'_{H_o}(D,X)W_\beta \\ l'_{\alpha_o}(D,X)-l'_{H_o}(D,X)W_\alpha \end{array}\right]^*\right)\left[\begin{array}{c} \beta-\beta_o \\ \alpha-\alpha_o \end{array}\right].
\end{aligned}$$

Consider the smooth functional $S(\beta,\alpha,H)\equiv\lambda_\beta\beta+\lambda_\alpha\alpha$ for any fixed $\lambda_\beta, \lambda_\alpha$ such that $|\lambda_\beta|+|\lambda_\alpha|=1$. Then

$$\begin{aligned}
S(\beta,\alpha,H)-S(\beta_o,\alpha_o,H_o) &= \lambda_\beta(\beta-\beta_o)+\lambda_\alpha(\alpha-\alpha_o) \\
&\equiv S'_{(\beta_o,\alpha_o,H_o)}[\beta-\beta_o,\alpha-\alpha_o]
\end{aligned}$$

$$\begin{aligned}
& \left\|S'_{(\beta_o,\alpha_o,H_o)}\right\|^2 \\
\equiv \ & \sup_{\{(\beta,\alpha,H)\in\mathcal{B}\times\mathcal{A}\times\mathcal{H}\}}\frac{\left|S'_{(\beta_o,\alpha_o,H_o)}[\beta-\beta_o,\alpha-\alpha_o]\right|^2}{\|(\beta,\alpha,H)-(\beta_o,\alpha_o,H_o)\|_e^2} \\
= \ & \sup_{\{(b,a,W)\in\mathcal{B}\times\mathcal{A}\times\mathcal{F}_o\}}\frac{\left[\begin{array}{c} b \\ a \end{array}\right]^*\left[\begin{array}{c} \lambda_\beta \\ \lambda_\alpha \end{array}\right]\left[\begin{array}{c} \lambda_\beta \\ \lambda_\alpha \end{array}\right]^*\left[\begin{array}{c} b \\ a \end{array}\right]}{\left[\begin{array}{c} b \\ a \end{array}\right]^* E_o\left(\left[\begin{array}{c} l'_{\beta_o}-l'_{H_o}W_\beta \\ l'_{\alpha_o}-l'_{H_o}W_\alpha \end{array}\right]\left[\begin{array}{c} l'_{\beta_o}-l'_{H_o}W_\beta \\ l'_{\alpha_o}-l'_{H_o}W_\alpha \end{array}\right]^*\right)\left[\begin{array}{c} b \\ a \end{array}\right]} \\
= \ & \left[\begin{array}{c} \lambda_\beta \\ \lambda_\alpha \end{array}\right]^*\left\{\inf_{W\in\mathcal{F}_o} E_o\left[\left[\begin{array}{c} l'_{\beta_o}(D,X)-l'_{H_o}(D,X)W_\beta \\ l'_{\alpha_o}(D,X)-l'_{H_o}(D,X)W_\alpha \end{array}\right]\left[\begin{array}{c} l'_{\beta_o}(D,X)-l'_{H_o}(D,X)W_\beta \\ l'_{\alpha_o}(D,X)-l'_{H_o}(D,X)W_\alpha \end{array}\right]^*\right]\right\}^{-1}\left[\begin{array}{c} \lambda_\beta \\ \lambda_\alpha \end{array}\right]
\end{aligned}$$

$$= \quad \begin{bmatrix} \lambda_\beta \\ \lambda_\alpha \end{bmatrix}^* \mathcal{I}^{-1} \begin{bmatrix} \lambda_\beta \\ \lambda_\alpha \end{bmatrix} \equiv \|v_o\|^2 < \infty \quad \text{iff} \quad \mathcal{I} \text{ is positive definite.}$$

Thus, we may choose the Riesz representor $v_o = (v_{o\beta}, v_{o\alpha}, v_{oh})^*$ as

$$\begin{bmatrix} v_{o\beta} \\ v_{o\alpha} \end{bmatrix} = \mathcal{I}^{-1} \begin{bmatrix} \lambda_\beta \\ \lambda_\alpha \end{bmatrix}, \qquad v_{oh} = - \begin{bmatrix} W_{o\beta}\mathcal{I}^{-1}\lambda_\beta \\ W_{o\alpha}\mathcal{I}^{-1}\lambda_\alpha \end{bmatrix}.$$

Assumption B

In the following we denote $\gamma_o \equiv (\beta_o, \alpha_o, H_o)$ and $\gamma \equiv (\beta, \alpha, H)$.

$$\begin{aligned} & r[\gamma - \gamma_o, D_i, X_i] \\ = \; & l(\gamma, D_i, X_i) - l(\gamma_o, D_i, X_i) \\ & - \Big[l'_\beta(\gamma_o, D_i, X_i)[\beta - \beta_o] + l'_\alpha(\gamma_o, D_i, X_i)[\alpha - \alpha_o] + l'_H(\gamma_o, D_i, X_i)[H - H_o] \Big] \end{aligned}$$

$$\text{(i)} \quad \sup_{\left\{\gamma : \|\gamma - \gamma_o\|_e \le \varepsilon_N\right\}} \sum_{i=1}^{N} \left(r[\gamma - \gamma_o, D_i, X_i] - E(r[\gamma - \gamma_o, D_i, X_i]) \right) = o_P(1)$$

$$\text{(ii)} \quad \sup_{\left\{\gamma : \|\gamma - \gamma_o\|_e \le \varepsilon_N\right\}} \left(KL(\gamma_o, \gamma) - \frac{1}{2} \|\gamma - \gamma_o\|_e^2 \right) = o(\frac{1}{N})$$

$$\text{(i)} \quad \sup_{\left\{\gamma : \|\gamma - \gamma_o\|_e \le \varepsilon_N\right\}} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(l'_{\gamma_o}[\gamma - \gamma_o, D_i, X_i] - E(l'_{\gamma_o}[\gamma - \gamma_o, D_i, X_i]) \right) = o_P(1)$$

Theorem 10: Assume that $\mathcal{I}$ is non-singular, Assumption B is satisfied. Then the NPMLE estimators $\widehat{\beta}_N$ and $\widehat{\alpha}_N$ are $\sqrt{N}$−efficient, that is,

$$\sqrt{N} \begin{bmatrix} \widehat{\beta}_N - \beta_o \\ \widehat{\alpha}_N - \alpha_o \end{bmatrix} \Longrightarrow \mathcal{N}(0, \mathcal{I}^{-1})$$

7 Monte Carlo

8 Conclusion

In this paper, we have discussed the identification, root-N estimableness, convergence rate and asymptotic normality of three typical panel data models with binary dependent variables. All three models allow for unobserved heterogeneity, and two of them in addition allow for true state-dependence. We show that the latent scalar factor structure is not only to reduce dimensionality to easy computational burden, but also impose much more restriction onto the models. It is thus important to test for such a factor structure. Our models do not depend on the factor model assumption for identification, and the factor model does restrict the joint dependence in the shocks. The factor model thus does impose testable restrictions in these models, and we will consider testing for the factor representation in subsequent research.

References

[1] Aakvik, A., J. Heckman, and E. Vytlacil (1998): "Training Effects on Employment when the Training Effects are Heterogeneous: An Application to Norwegian Vocational Rehabilitation Programs", working paper, University of Chicago.

[2] Amemiya, T. (1985): *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

[3] Arellano, M., and B. Honoré (1999): "Panel Data Models. Some Recent Developments"

[4] Baker, M. and A. Melino (1997): "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study", working paper, University of Toronto.

[5] Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, Maryland: John Hopkins University Press

[6] Cameron, S. and J. Heckman (1987): "Son of CTM: The DCPA Approach Based on Discrete Factor Structure Models", working paper, University of Chicago.

[7] Cameron, S. and J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias", *Journal of Political Economy* 106:2, 262-333.

[8] Cameron, S. and C. Taber (1994): "Evaluation and Identification of Semiparametric Maximum Likelihood Models of Dynamic Discrete Choice", working paper. University of Chicago.

[9] Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics* 32, 189-218

[10] Chen, X., J. Heckman, and E. Vytlacil (1998): "Semiparametric Estimation of Panel Data Binary Dependent Variable Models With Latent Factors", working paper, University of Chicago.

[11] Chen, X., and X. Shen (1998): "Asymptotic Properties of Sieve Extremum Estimates for Weakly Dependent Data with Applications", *Econometrica* 66:2, 289-314.

[12] Cosslett, S. (1983): "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* 51:3, 765-782.

[13] Follman, D. (1985): "Nonparametric Mixtures of Logistic Regression Models", unpublished Ph.D. Dissertation, Carnegie Mellon, Department of Statistics.

[14] Follman, D. and D. Lambert (1989): "Generalizing Logistic Regression by Nonparametric Mixing", *Journal of American Statistical Association*, 84, 295-300.

[15] Grenader, U. (1981): *Abstract Inference*, New York: Wiley Series.

[16] Hahn, J. (1994): "The Efficiency Bound of the Mixed Proportional Hazard Model", *Review of Economic Studies*, 61(4), 607-629.

[17] Hahn, J. (1998): "Information Bound of Dynamic Panel Logit Model with Fixed Effects," unpublished manuscript, University of Pennsylvania.

[18] Heckman, J. (1981a): "Statistical Models for Discrete Panel Data", in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data*, MIT Press.

[19] ______ (1981b): "The Incidental Parameter Problem and the Problem of Initial Conditions", in C. Manski and D. McFadden, eds, *Structural Analysis of Discrete Data*, MIT Press.

[20] Heckman, J. and B. Honoré(1990): "The Empirical Content of the Roy Model", *Econometrica* 58(5): 1121-1149.

[21] Heckman, J. and Singer, B. (1984): "A Method for Minimizing the Impact of Distributional Assumptions", *Econometrica* 52.

[22] Heckman, J. and J. Smith (1998): "Evaluating the Welfare State", in S. Strom, ed., *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Econometric Society Monograph Series, Cambridge University Press.

[23] Heckman, J., J. Smith, and N. Clemence (1997): "Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts", *Review of Economic Studies*, 64: 487-535.

[24] Heckman, J., and C. Taber (1994): "Econometric Mixture Models and More General Models for Unobservables in Duration Analysis", unpublished manuscript, University of Chicago.

[25] Heckman, J., and R. Willis (1977): "A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women", *Journal of Political Economy* 85, 27-58.

[26] Honoré, B., and E. Kyriazidou (1998a): "Panel Data Discrete Choice Models with Lagged Dependent Variables", working paper, Princeton University.

[27] Honoré, B., and E. Kyriazidou (1998b): "Estimation of Tobit-Type Models with Individual Specific Effects", working paper, Princeton University.

[28] Honoré, B., and A. Lewbel (1998): "Semiparametric Binary Choice Panel Data Models without Strictly Exogonous Regressors", working paper, Princeton University.

[29] Huh, K. and R. Sickles (1994): "Estimation of the Duration Model by Non-Parametric Maximum Likelihood, Maximum Penalized Likelihood and Probability Simulators", *Review of Economics and Statistics*, 76, 683-694.

[30] Kagan, A., Y. Linnik, and C.R. Rao (1973): *Characterization Problems in Mathematical Statistics* John Wiley & Sons.

[31] Kiefer, J. and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *Annals of Mathematical Statistics*, 27, 363-366.

[32] Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models", *Econometrica*, 61(2), 387-421.

[33] Lewbel, A., (1998): "Semiparametric Qualitative Response Model Estimation with Instrumental Variables and Unknown Heteroscedasticity", unpublished manuscript, Brandeis University.

[34] Lindsey, B. (1983a): "The Geometry of Mixture Likelihoods, Part I", *Annals of Statistics*, 11, 86-94.

[35] Lindsey, B. (1983b): "The Geometry of Mixture Likelihoods, Part II", *Annals of Statistics*, 11, 783-792.

[36] Manski, C. (1988): "Identification of Binary Response Models," *Journal of American Statistical Association*, 83, 729-737.

[37] Prakasa Rao, B.L.S. (1992): *Identifiability in Stochastic Models, Characterization of Probability Distributions*, Academic Press.

[38] Rao, C.R. (1971): "Characterization of Probability Laws by Linear Functions", *Sankhyā Ser A*, 33, 265-270.

[39] Sheps, M. and J. Menken (1973): *Mathematical Models of Conception and Birth* Chicago: University of Chicago Press.

[40] Van der Vaart, A. (1996): "Efficient Maximum Likelihood Estimation in Semiparametric Mixture Models", *The Annals of Statistics*, 24(2), 862-.

Appendix A: Identification Proofs.

Proof. (Theorem 2): By hypothesis, we know the left hand sides of the following $\bar{T}$ equations:

$$Pr\left(D_1 = 1|X_1 = x_1\right) = F_{\eta_1}\left(x_1\beta_1\right) \tag{A-1}$$

$$Pr\left(D_1 = 1, D_2 = 1|X_1 = x_1, X_2 = x_2\right) = F_{\eta_1,\eta_2}\left(x_1\beta_1, x_2\beta_2\right) \tag{A-2}$$

$$\cdots$$

$$\begin{aligned} & Pr\left(D_1 = 1, D_2 = 1, \ldots, D_{\bar{T}} = 1|X_1 = x_1, \ldots, X_{\bar{T}} = x_{\bar{T}}\right) \qquad \text{(A-3)} \\ = \quad & F_{\eta_1,\ldots,\eta_{\bar{T}}}\left(x_1\beta_1, \ldots, x_{\bar{T}}\beta_{\bar{T}}\right)). \end{aligned}$$

We may treat (A-1) as a binary discrete choice model and again following the analysis of Manski (1988, Proposition 2, Corollary 5) attain that we identify β_1 and F_{η_1} up to scale and location. For example, we may normalize the location and scale by constraining X_1 not to have an intercept and constraining $||\beta_1|| = 1$.

However, unlike the case for model one, we cannot directly apply Manski for $T \geq 2$. We do not directly observe $Pr(D_2 = 1|X_2)$, since the D_2 outcome is not observed for individuals with $D_1 = 0$. We therefore proceed with a recursive "identification at the limit" arguement.

If the true parameter values are $(F_{\eta_2^0},\ \beta_2^0)$, then given the identification of the first period parameters from the first step, the second period parameters are identified, iff for any alternative parameter values $(F_{\eta_2^*},\ \beta_2^*) \in \mathcal{H}_2 \times \mathcal{B}_2$ with $(F_{\eta_2^*}, \beta_2^*) \neq (F_{\eta_2^0}, \beta_2^0)$, there exists some $\epsilon > 0$ s.t.

$$Pr(|F_{\eta_1^0,\eta_2^0}\left(X_1\beta_1^0, X_2\beta_2^0\right) - F_{\eta_1^0,\eta_2^*}\left(X_1\beta_1^0, X_t\beta_2^*\right)| > \epsilon) > 0. \tag{A-4}$$

Pick any $(F_{\eta_2^*},\ \beta_2^*) \in \mathcal{H}_2 \times \mathcal{B}_2/(F_{\eta_2^0},\ \beta_2^0)$. We will show (A-4) holds for some $\epsilon > 0$. By continuity of $F_{\eta_1^0}$, we have that for any $\varepsilon > 0$ we can pick $\tilde{g}_1 \in (L_1, U_1)$ such that

$$1 - F_{\eta_1^0}\left(g_1\right) \leq \varepsilon/2 \text{ for all } g_1 \geq \tilde{g}_1 \Longrightarrow \sup_{g_2}|F_{\eta_1^0,\eta_2^0}\left(g_1, g_2\right) - F_{\eta_2^0}\left(g_2\right)| \leq \varepsilon/2$$

and

$$\sup_{g_2}|F_{\eta_1^0,\eta_2^*}\left(g_1, g_2\right) - F_{\eta_2^*}\left(g_2\right)| \leq \varepsilon/2$$

for all $g_1 \geq \tilde{g}$. For any $\epsilon > 0$, we have

$$\begin{aligned} & Pr(|F_{\eta_1^0,\eta_2^0}\left(X_1\beta_1, X_2\beta_2^0\right) - F_{\eta_1^0,\eta_2^*}\left(X_1\beta_1, X_2\beta_2^*\right)| > \epsilon|X_1\beta_1 \geq \max(\tilde{g}_1, \check{g}_1) \\ \geq \quad & Pr(|F_{\eta_2^0}\left(X_2\beta_2^0\right) - F_{\eta_2^*}\left(X_2\beta_2^*\right)| > \epsilon + \varepsilon|X_1\beta_1 \geq \max(\tilde{g}_1, \check{g}_1). \end{aligned}$$

Using (iii′) and (iv), we have that $Pr\left(F_{\eta_2^0}(X_2\beta_2^0) = F_{\eta_2^*}(X_2\beta_2^*)|X_1\beta_1 \geq \max(\tilde{g}_1, \breve{g}_1\right) = 1$ iff $(F_{\eta_2^*},\ \beta_2^*)$ =$(F_{\eta_2^0},\ \beta_2^0)$. Since $(F_{\eta_2^*},\ \beta_2^*) \neq$$(F_{\eta_2^0},\ \beta_2^0)$, and since we can set ε arbitrarily small, we have that there exists ϵ values such that the last probability is strictly positive so that, for such ϵ values,

$$Pr\left(|F_{\eta_1^0,\eta_2^0}\left(X_1\beta_1, x_2\beta_2^0\right) - F_{\eta_1^0,\eta_2^*}\left(X_1\beta_1, x_2\beta_2^*\right)| > \epsilon|X_1\beta_1 \geq \max(\tilde{g}_1, \breve{g}_1)\right) > 0$$

Using (iv), we have

$$Pr(X_1\beta_1 \geq \max(\tilde{g}_1, \breve{g}_1)) > 0,$$

so that (A-4) holds. We have shown that $(F_{\eta_2^*},\ \beta_2^*) \neq$$(F_{\eta_2^0},\ \beta_2^0)$ implies (A-4), and thus the $(F_{\eta_2^0},\ \beta_2^0)$ parameters are identified. Proceeding in this fashion, we can recover $x_t\beta_t$, $t = 1, \ldots, \bar{T}$. Since we identify $x_t\beta_t$ and using (iv), we can recover the joint distribution of $(\eta_1, \ldots, \eta_{\bar{T}})$ varying the components of $(x_1\beta_1, ..., x_{\bar{T}}\beta_{\bar{T}})$ to trace out the joint distribution $F_{\eta_1,\ldots,\eta_{\bar{T}}}$. ■

Proof. (Corollary 1). Let

$$x\beta_1 = g_1$$

recalling that $\gamma_t = 0$ for $t = 1, ..., \bar{T}$, as a normalization and that the first $\bar{T}$ coordinates of x correspond to continuous regressors. By assumption (iv), $\beta_{11} \neq 0$, and we can write

$$x_1 = \frac{g_1}{\beta_{11}} - x_2\frac{\beta_{12}}{\beta_{11}} - \cdots - x_K\frac{\beta_{1K}}{\beta_{11}}$$

where in this expression lower case x_i is the i^{th} coordinate of x.

In the index $x\beta_2$, use standard Gaussian elimination and substitute for x_1, from the preceding equation and obtain

$$\left(\frac{g_1}{\beta_{11}} - x_2\frac{\beta_{12}}{\beta_{11}} - \cdots - x_K\frac{\beta_{1K}}{\beta_{11}}\right)\beta_{21} + \beta_{22}x_2 + \cdots + \beta_{2K}x_K.$$

These variables can be freely varied given $x\beta_1 = g_1$. Proceeding recursively, in the $(j+1)^{th}$ argument, $(j < \bar{T})$, we obtain an expression that substitutes out for $(x_1, ..., x_j)$ leaving at least $\bar{T} - j$ free continuous variables.

Array the β_j into a matrix B with the j^{th} row of B being β_j. B is an $\bar{T}$ $\times K$ matrix. Let $B(r, n)$ be the $r \times n$ submatrix of B consisting of the first r rows and n columns, and let $B(r, K - n)$ be the matrix consisting of the first r rows and the last $K - n$ columns of B. Partition β_j into the first e elements $\left(\beta_j\,(e)\right)$ and the last $K - e$ elements $\beta_j(K - e)$.

In this notation, successive Gaussian elimination produces

$$\bar{\beta}_{j+1} = \beta_j\,(K - j) - \beta_{j+1}\,(j)\,[B\,(j,j)]^{-1}\,B\,(j, K - j)$$

a $K-j$ dimensional vector. In order for $[B(j,j)]^{-1}$ to exist, it is necessary that $\beta_1,\ldots,\beta_j$ be linearly independent vectors. Condition (v) assures us that this requirement is satisfied for $j\leq m$. Define $\widetilde{\beta}_{j+1}\left(\bar{T}-j\right)$ as the first $(\bar{T}-j)$ elements of $\widetilde{\beta}_{j+1}$ associated with the continuous regressors. In order to satisfy (vi), at least one component of $\widetilde{\beta}_{j+1}(\bar{T}-j)$ must be non-zero.

Again consider

$$g_1 = x\beta_1$$

$$g_{2=}\tilde{\gamma}_2(g_1)+\hat{x}^2\widetilde{\beta}_{j+1}$$

where $\tilde{\gamma}_2(g_1)=\left(\frac{g_1}{\beta_{11}}\beta_{21}\right)$ is obtained via the same linear transformation that is used to obtain $\widetilde{\beta}_{j+1}$. Since $\tilde{\gamma}_2(g_1)$ is a function of g_1, the second period index is a function of g_1 and for fixed $\hat{x}^2$ we have that $g_1\to\infty\Longrightarrow g_2\to\infty$. However, note that using assumptions (iii)-(v), we can send $g_1\to\infty$ while varying $\hat{x}^2$ to keep g_2 fixed. In particular, we can use x_1 to send $g_1\to\infty$ and set x_2 to compensate for x_1 in the second period index so as to hold g_2 fixed. Thus, $supp(X\beta_2|X\beta_1=g_1)=R$ and X such that $X\beta_1=g_1$ will have rank $K-1$ for a.e. $g_1\in R$. Moreover, we have, for a.e. $g_1\in R$, $supp(X\beta_2|X\beta_1\geq g_1)=R$ and X such that $X\beta_1\geq g_1$ has full rank (there exists no proper linear subspace of R^K having probability 1 under $F_{X|X\beta_1\geq g_1}$). We can repeat this argument sequentially, using sequential Gaussian elimination as described above, to show

$$Supp\left(X\beta_t|X\beta_1=g_1,...,X\beta_{t-1}=g_{t-1}\right)=R$$

and there exists no proper linear subspace of R^K having probability 1 under $F_{X|X\beta_1\geq g_1,\ldots,X\beta_t\geq g_t}$ for almost every $(g_{t-1},\ldots,g_1)\in R^{t-1}$ for $t=2,...,\bar{T}$. ■

Appendix B: $\sqrt{N}$ Estimableness Proofs.

Lemma 2 (Single Spell): For model (2a) or (2b), let all the conditions for Theorem 5 be satisfied with $G_t(y)=\frac{\exp(y)}{1+\exp(y)}$ for $t=1,...,\overline{T}$. For simplicity we assume that all β_t are distinct and that X_t is independent across time or first-order Markov process. Then:

$$\begin{aligned}&\wp(\beta,\alpha,T|X,\Theta)\\ =\;&\exp\left\{\sum_{k=1}^{\overline{T}}\delta_T(k)\Big[\sum_{j=1}^{k}X_j\beta_j+\Theta\sum_{j=1}^{k}\alpha_j-\sum_{j=1}^{k+1}\log[1+\exp(X_j\beta_j+\alpha_j\Theta)]\Big]\right\}\end{aligned}$$

The sufficient statistics is $\left\{\sum_{k=1}^{\overline{T}}\delta_T(k)\sum_{j=1}^{k}\alpha_k,\ [X_1\widetilde{\beta}_1,...,X_{\overline{T}}\widetilde{\beta}_{\overline{T}}]\right\}$.

Let $[logG_t]'$ denote $\frac{\partial \log(G_t(y))}{\partial y}$. Thus,

$$\mathcal{S}_{\beta_o} = \frac{\int \wp(\beta_o, \alpha_o, T|X, \theta)\left(\frac{\partial}{\partial \beta}[\log(\wp(\beta_o, \alpha_o, T|X, \theta))] - f_{\beta_o}(\theta)\right) dH_o(\theta)}{\int \wp(\beta_o, \alpha_o, T|X, \theta) dH_o(\theta)}$$

$$\mathcal{S}_{\alpha_o} = \frac{\int \wp(\beta_o, \alpha_o, T|X, \theta)\left(\frac{\partial}{\partial \alpha}[\log(\wp(\beta_o, \alpha_o, T|X, \theta))] - f_{\alpha_o}(\theta)\right) dH_o(\theta)}{\int \wp(\beta_o, \alpha_o, T|X, \theta) dH_o(\theta)},$$

where for $t = 1, 2, ..., \overline{T}$,

$$\begin{aligned} f_{\beta_{t,o}}(\theta) \;&=\; [\log G_t]' E\left[1(T_i \geq t) X_t \mid [X_1 \widetilde{\beta}_1, ..., X_{\overline{T}} \widetilde{\beta}_{\overline{T}}], \sum_{k=1}^{\overline{T}} \delta_{T_i}(k) \sum_{j=1}^{k} \alpha_j\right] \\ &\quad + [\log(1 - G_t)]' E\left[1(T_i + 1 = t) X_t \mid [X_1 \widetilde{\beta}_1, ..., X_{\overline{T}} \widetilde{\beta}_{\overline{T}}], \sum_{k=1}^{\overline{T}} \delta_{T_i}(k) \sum_{j=1}^{k} \alpha_j\right], \end{aligned}$$

and for $t = 2, ..., \overline{T}$,

$$\begin{aligned} f_{\alpha_{t,o}}(\theta) \;&=\; \theta [\log G_t]' E\left[1(T_i \geq t) \mid [X_1 \widetilde{\beta}_1, ..., X_{\overline{T}} \widetilde{\beta}_{\overline{T}}], \sum_{k=1}^{\overline{T}} \delta_{T_i}(k) \sum_{j=1}^{k} \alpha_j\right] \\ &\quad + \theta [\log(1 - G_t)]' E\left[1(T_i + 1 = t) \mid [X_1 \widetilde{\beta}_1, ..., X_{\overline{T}} \widetilde{\beta}_{\overline{T}}], \sum_{k=1}^{\overline{T}} \delta_{T_i}(k) \sum_{j=1}^{k} \alpha_j\right]. \end{aligned}$$

Proof. (**Theorem 7, Single Spell):** Similar proof as that for Theorem 6 except applying Lemma 2. For example, when $\overline{T} = 2$, we have $\sum_{k=1}^{\overline{T}} \delta_{T_i}(k) \sum_{j=1}^{k} \alpha_j = \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2$. Thus,

$$\begin{aligned} & \frac{\partial \log \wp_o(\theta)}{\partial \beta_1} - f_{\beta_{1,o}}(\theta) \\ = \;& [\log G_1]' \left(1(T_i \geq 1) X_1 - E\left[1(T_i \geq 1) X_1 \mid X_1 \beta_1, X_2 \beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right) + \\ & [\log(1 - G_1)]' \left(1(T_i = 0) X_1 - E\left[1(T_i = 0) X_1 \mid X_1 \beta_1, X_2 \beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right), \end{aligned}$$

$$\begin{aligned} & \frac{\partial \log \wp_o(\theta)}{\partial \beta_2} - f_{\beta_{2,o}}(\theta) \\ = \;& [\log G_2]' \left(1(T_i \geq 2) X_2 - E\left[1(T_i \geq 2) X_2 \mid X_1 \beta_1, X_2 \beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right) + \\ & [\log(1 - G_2)]' \left(1(T_i = 1) X_2 - E\left[1(T_i = 1) X_2 \mid X_1 \beta_1, X_2 \beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right), \end{aligned}$$

and

$$\begin{aligned} & \frac{\partial \log \wp_o(\theta)}{\partial \alpha_2} - f_{\alpha_{2,o}}(\theta) \\ = \;& \theta [\log G_2]' \left(1(T_i \geq 2) - E\left[1(T_i \geq 2) \mid X_1\beta_1, X_2\beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right) + \\ & \theta[\log(1 - G_2)]' \left(1(T_i = 1) - E\left[1(T_i = 1) \mid X_1\beta_1, X_2\beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right]\right). \end{aligned}$$

Since

$$E\left[1(T_i = 0) \mid X_1\beta_1, X_2\beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right] \neq 1(T_i = 0) \text{ if and only if } \alpha_2 = -1,$$

and

$$E\left[1(T_i = 2) \mid X_1\beta_1, X_2\beta_2, \delta_{T_i}(1) + \delta_{T_i}(2) + \delta_{T_i}(2)\alpha_2\right] \neq 1(T_i = 2) \text{ if and only if } \alpha_2 = -1,$$

we therefore have that β_1, β_2 and α_2 may be root-N estimable if and only if $\alpha_2 = -1$.

Of course, when $\alpha_2 \neq -1$, $\widetilde{\beta}$ is still root-N estimable under conditions on X similar to the case when $\overline{T} = 1$. ■

Lemma 3 (switching regression): For the model defined by equations (3a) and (3b), let the conditions for Corollary 4 be satisfied with $G_1(y), G_{21}(y), G_{20}(y) = \frac{\exp(y)}{1+\exp(y)}$. For simplicity we assume that X_t is independent across time or first-order Markov process. We have that:

$$\begin{aligned} & \log\left(\wp(\beta, \alpha, D_i | X_i, \Theta_i = \theta)\right) \\ = \;& \theta\Big[D_1 + D_1D_2(\alpha_{21} - \alpha_{20}) + D_2\alpha_{20}\Big] \\ & + \Big[D_1X_1\beta_1 + D_1D_2X_2(\beta_{21} - \beta_{20}) + D_2X_2\beta_{20}\Big] \\ & - \log\left(1 + \exp(X_1\beta_1 + \theta)\right) \\ & - D_1 \log\left(1 + \exp(X_2\beta_{21} + \alpha_{21}\theta)\right) \\ & - (1 - D_1)\log\left(1 + \exp(X_2\beta_{20} + \alpha_{20}\theta)\right) \end{aligned}$$

and

$$\frac{\partial \log \wp_o(\theta)}{\partial \beta_1} = X_1[D_1 - G_1(X_1\beta_1 + \theta)]$$

$$\frac{\partial \log \wp_o(\theta)}{\partial \beta_{21}} = X_2D_1[D_2 - G_{21}(X_2\beta_{21} + \alpha_{21}\theta)]$$

$$\frac{\partial \log \wp_o(\theta)}{\partial \beta_{20}} = X_2(1-D_1)[D_2 - G_{20}(X_2\beta_{20} + \alpha_{20}\theta)]$$

$$\frac{\partial \log \wp_o(\theta)}{\partial \alpha_{21}} = \theta D_1[D_2 - G_{21}(X_2\beta_{21} + \alpha_{21}\theta)]$$

$$\frac{\partial \log \wp_o(\theta)}{\partial \alpha_{20}} = \theta(1-D_1)[D_2 - G_{20}(X_2\beta_{20} + \alpha_{20}\theta)]$$

and the sufficient statistics is: $\{[D_1 + D_1D_2(\alpha_{21} - \alpha_{20}) + D_2\alpha_{20}], X_1\beta_1, X_2\beta_{21}, X_2\beta_{20}\}$. Let $\mathcal{T} = \{[D_1 + D_1D_2(\alpha_{21} - \alpha_{20}) + D_2\alpha_{20}], X_1\beta_1, X_2\beta_{21}, X_2\beta_{20}\}$.

Then the efficient scores for $(\beta_1, \beta_{21}, \beta_{20}, \alpha_{21}, \alpha_{20})$ are

$$\begin{aligned}
\mathcal{S}_{\beta_{1,o}}(D_i, X_i) &= (D_1X_1 - E[D_1X_1|\mathcal{T}]) \\
&\quad -\frac{\int G_1\wp_o(D_i|X_i,\theta)dH_o(\theta)}{\int \wp_o(D_i|X_i,\theta)dH_o(\theta)}\left\{X_1 - E[X_1|X_1\beta_{1,o}]\right\}
\end{aligned}$$

$$\begin{aligned}
\mathcal{S}_{\beta_{21,o}}(D_i, X_i) &= (D_1D_2X_2 - E[D_1D_2X_2|\mathcal{T}]) \\
&\quad -\frac{\int G_21\wp_o(D_i|X_i,\theta)dH_o(\theta)}{\int \wp_o(D_i|X_i,\theta)dH_o(\theta)}\left\{X_2 - E[X_2|X_2\beta_{21,o}]\right\}
\end{aligned}$$

Appendix C: Estimation Proofs

Denote $M_{[]}(\epsilon, \mathcal{F}_N)$ as the minimum number of pairs of functions which are $\epsilon-$apart in Hellinger distance needed to bracket any functions in $\mathcal{F}_N$, and $M(\epsilon, \mathcal{F}_N, |\cdot|)$ as the minimum number of balls with $\epsilon-$radius in $|\cdot|$ distance needed to cover any functions in $\mathcal{F}_N$.

Before we prove Theorem 9, we present a more general convergence rate result.

Theorem 11: Suppose conditions for Theorem 5 and Corollary 4 are satisfied. Let $\mathcal{B}$ and $\mathcal{A}$ be finite-dimensional compact sets, and let $\mathcal{H}$ be the space of probability distributions with known bounded support $[\underline{\theta}, \overline{\theta}]$ such that $\frac{\int \wp(\beta,\alpha,D|X,\theta)dH(\theta)}{\int \wp(\beta_o,\alpha_o,D|X,\theta)dH_o(\theta)}$ is bounded below and above by constants. Then: $\left\|(\widehat{\beta}_N, \widehat{\alpha}_N, \widehat{H}_N) - (\beta_o, \alpha_o, H_o)\right\| = O_p(N^{-1/3})$.

Proof. **(Theorem 11).** We prove this by verifying that conditions C1-C3 for theorem 1 in Shen and Wong (1994) are satisfied. First, condition C1 is satisfied since

$$KL((\beta_o, \alpha_o, H_o), (\beta, \alpha, H)) \geq 2 \left\| (\beta, \alpha, H) - (\beta_o, \alpha_o, H_o) \right\|^2 .$$

Second, the boundedness of the likelihood ratio assumption implies condition C2 of Shen and Wong since

$$\begin{aligned} & Var_o \left[l(\beta_o, \alpha_o, H_o; D_i, X_i) - l(\beta, \alpha, H; D_i, X_i) \right] \\ \leq \; & E_o \left[\log \left(\frac{dP_{\beta,\alpha,h}}{dP_o} \right) \right]^2 = 4E_o \left[\log \left(1 + [\sqrt{\frac{dP_{\beta,\alpha,h}}{dP_o}} - 1] \right) \right]^2 \\ \leq \; & 4E_o \left[\sqrt{\frac{dP_{\beta,\alpha,h}}{dP_o}} - 1 \right]^2 \leq 8 \left\| (\beta, \alpha, H) - (\beta_o, \alpha_o, H_o) \right\|^2 \end{aligned}$$

For condition C3, we need to compute the metric entropy for

$$\mathcal{F} = \left\{ l(\beta, \alpha, H; D_i, X_i) - l(\beta_o, \alpha_o, H_o; D_i, X_i) : (\beta, \alpha, H) \in \mathcal{B} \times \mathcal{A} \times \mathcal{H} \right\} .$$

Denote $B_\delta(*) = \{(\beta, \alpha, H) \in \mathcal{B} \times \mathcal{A} \times \mathcal{H} : |\beta - \beta_*| + |\alpha - \alpha_*| + |H - H_*|_{\sup} \leq \delta\}$. Notice that the boundedness of likelihood ratio also implies

$$\begin{aligned} & E_o \sup_{B_\delta(*)} \left[l(\beta_*, \alpha_*, H_*; D_i, X_i) - l(\beta, \alpha, H; D_i, X_i) \right]^2 \\ \leq \; & 4E_o \sup_{B_\delta(*)} \left[\sqrt{dP_{\beta,\alpha,h}} - \sqrt{dP_{\beta_*,\alpha_*,h_*}} \right]^2 \\ \leq \; & const. \int \sup_{B_\delta(*)} \frac{\left[\left(\int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta, \alpha, D|X, \theta) dH(\theta) - \int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta_*, \alpha_*, D|X, \theta) dH_*(\theta) \right) f_X \right]^2}{\left[\sqrt{dP_{\beta,\alpha,h}} + \sqrt{dP_{\beta_*,\alpha_*,h_*}} \right]^2} dP_o \\ \leq \; & const \int \sup_{B_\delta(*)} \left[\left(\int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta, \alpha, D|X, \theta) dH(\theta) - \int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta_*, \alpha_*, D|X, \theta) dH_*(\theta) \right) f_X \right]^2 d\mu. \end{aligned}$$

Notice that

$$\int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta, \alpha, D|X, \theta) dH(\theta) = - \int_{\underline{\theta}}^{\overline{\theta}} H(\theta) \frac{\partial \wp(\beta, \alpha, D|X, \theta)}{\partial \theta} d\theta + \wp(\beta, \alpha, D|X, \overline{\theta}),$$

and thus

$$\int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta, \alpha, D|X, \theta) dH(\theta) - \int_{\underline{\theta}}^{\overline{\theta}} \wp(\beta_*, \alpha_*, D|X, \theta) dH_*(\theta)$$

$$\begin{aligned} &= [\wp(\beta,\alpha,D|X,\overline{\theta})-\wp(\beta_*,\alpha_*,D|X,\overline{\theta})] \\ &\quad +\int_{\underline{\theta}}^{\overline{\theta}}[H_*(\theta)-H(\theta)]\frac{\partial\wp(\beta,\alpha,D|X,\theta)}{\partial\theta}d\theta \\ &\quad +\int_{\underline{\theta}}^{\overline{\theta}}H_*(\theta)[\frac{\partial\wp(\beta_*,\alpha_*,D|X,\theta)}{\partial\theta}-\frac{\partial\wp(\beta,\alpha,D|X,\theta)}{\partial\theta}]d\theta \end{aligned}$$

where

$$\begin{aligned} &\frac{\partial\wp(\beta,\alpha,D|X,\theta)}{\partial\theta} \\ &= \wp(\beta,\alpha,D|X,\theta)\sum_{t=1}^{\overline{T}}\alpha_t\left[D_t\frac{[G_t(X_t\beta_t+\alpha_t\theta)]'}{G_t(X_t\beta_t+\alpha_t\theta)}-(1-D_t)\frac{[G_t(X_t\beta_t+\alpha_t\theta)]'}{1-G_t(X_t\beta_t+\alpha_t\theta)}\right] \end{aligned}$$

Since $\mathcal{B}$ and $\mathcal{A}$ are finite-dimentional compact sets, the corresponding sup-norm metric entropy satisfies $\log(M(u,\mathcal{B},|\cdot|))+\log(M(u,\mathcal{A},|\cdot|))\leq C_1\log(\frac{1}{u})$. Since $\mathcal{H}$ consists of monotone increasing uniform bounded functions, by Birman and Solomjak (1967), $\log(M(u,\mathcal{H},|\cdot|_{\sup}))\leq\frac{C_2}{u}$. Therefore, the bracketing Hellinger metric entropy $\log(M_{[]}(u,\mathcal{F}))$ is bounded by:

$$\begin{aligned} \log(M_{[]}(u,\mathcal{F})) &\leq \log(M(u,\mathcal{B},|\cdot|))+\log(M(u,\mathcal{A},|\cdot|))+\log(M(u,\mathcal{H},|\cdot|_{\sup})) \\ &\leq C\log(\frac{1}{u})+\frac{C'}{u} \end{aligned}$$

Now the convergence rate is simply $\left\|(\widehat{\beta}_N,\widehat{\alpha}_N,\widehat{H}_N)-(\beta_o,\alpha_o,H_o)\right\|=O_p(\varepsilon_N)$, where

$$\varepsilon_N=\inf\left\{\varepsilon:\frac{1}{\varepsilon^2}\int_{\varepsilon^2}^{\varepsilon}\sqrt{\log(M_{[]}(u,\mathcal{F}))}du\leq const.\sqrt{N}\right\}$$

which is satisfied by $\varepsilon_N=const.N^{-1/3}$. ■

Proof. (**Theorem 9):** When all $G_t(y)=\frac{\exp(y)}{1+\exp(y)}$, the boundedness of the likelihood ratio is automatically satisfied as long as $\mathcal{B}$ and $\mathcal{A}$ are finite-dimensional compact sets, and $\mathcal{H}$ is the space of probability distributions with bounded known supports. Hence the above convergence rate holds. ■