Estimation and Inference in Panel Structure Models*

Yixiao Sun
Department of Economics
University of California, San Diego

This version: August 30, 2005

Abstract

This paper proposes and implements a tractable approach to detect group structure in panel data. The mechanism works by means of a panel structure model, which assumes that individuals form a number of homogeneous groups in a heterogeneous population. Within each group, the (linear) regression coefficients are the same, while they may be different across different groups. The econometrician is not presumed to know the group structure. Instead, a multinomial logistic regression is used to infer which individuals belong to which groups.

The model is estimated via maximum likelihood. We prove the consistency and asymptotic normality of a global MLE under the mild assumption that the time dimension is larger than the number of regressors in the linear regression. We propose a likelihood ratio test to test the null of one group against the alternative of multiple groups. Simulation studies show that the MLE performs quite well and the likelihood ratio test has good size and power properties in finite samples.

JEL Classification: C13; C23; C51

Keywords: dynamic panel data model; group structure; logistic regression; nonregular test; parameter heterogeneity

*This paper is a revision of an earlier paper titled "Asymptotic Theory for Panel Strucuture Models" (2001). I am especially grateful to Donald Andrews, Peter Phillips, Chris Udry for their insightful comments. I would like to thank Michael Keane and Zhijie Xiao who have contributed to this paper with constructive suggestions. All errors are mine alone. *Email:* yisun@ucsd.edu. *Address:* Yixiao Sun, Department of Economics, #0508, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0508, USA.

1 Introduction

Panel data techniques have become standard machinery for empirical analysts in many disciplines, including economics, accounting, finance, and marketing. In part, this is because panel data allow us to construct and test more complicated models that can not be identified using only cross section or time series data. In part also, panel data give more variability and more degrees of freedom, which leads to more precise estimation. Since panel data usually cover individuals coming from different backgrounds and living in different environments, it is of paramount importance to control unobserved heterogeneity in panel data modeling.

One way to deal with the unobserved heterogeneity is to assume complete homogeneity in which regression parameters are the same across all individuals. Under this assumption, individual observations can be viewed as random draws from a common population. By pooling observations across all individuals, one can draw more precise inferences on population characteristics. However, a large body of applied work has tested and rejected parameter homogeneity. For example, Baltagi and Griffin (1997) rejected the hypothesis that the gasoline demand elasticities across the OECD countries were equal. Haque, Pesaran and Sharma (2000) rejected the hypothesis that the savings behaviors across different countries were equal.

The other way to deal with the unobserved heterogeneity is to assume complete heterogeneity in which regression parameters may be completely different for different individuals. Under this assumption, the number of parameters goes to infinity as the number of individuals approaches infinity. This leads to the incidental parameter problem in that maximum likelihood estimators of population-specific parameters may be inconsistent for a fixed time dimension (c.f. Neyman and Scott (1948), Lancaster (2000)). When the time dimension is small, it is also very difficult to estimate individual-specific parameters accurately. For example, Baltagi and Griffin (1997) found that completely heterogeneous models led to very imprecise parameter estimates, which in some cases had the wrong sign. In addition, the assumption of complete heterogeneity may render the underlying empirical analysis meaningless. In an empirical investigation of international growth and convergence, Lee, Pesaran and Smith (1997, 1998) allowed complete heterogeneity in not only the speed of convergence but also the steady state growth rate. However, as pointed out by Islam (1998), the extension to allow varying growth rates runs the risk of robbing the concept of convergence of any real economic meaning.

Instead of assuming complete heterogeneity or homogeneity, we set up a new model of an intermediate form, in which individuals form a number of homogeneous groups in a heterogeneous population. Within each group, the regression coefficients are the same, while they may be different across different groups. The new model does not assume that the group structure is known *a priori*. Instead, it employs a multinomial logistic regression to model the membership probabilities. The model thus consists of a set of linear response patterns relating the dependent variable to explanatory variables and a logistic regression that classifies these response patterns. From an empirical Bayes perspective, the logistic regression provides prior probabilities that individuals belong to a particular group, whereas the linear response patterns provide information to update the priors in order to obtain posterior probabilities. Based on the posterior probabilities, an individual is assigned to the group of which it is most likely to be a member. The underlying structure can thus be recovered from the posterior probabilities.

The model aims to detect a group structure using panel data, hence the name "panel structure model." It can be regarded as a bridge between a model with homogeneous coefficients

and one with completely heterogeneous coefficients. To some extent, the panel structure model avoids the shortcomings of both models while retaining their advantages. First, compared to a model with completely homogeneous coefficients, the panel structure model is less likely to be mis-specified. On the other hand, the presence of a group structure allows us to pool individuals in the same group, so that we can estimate group-specific effects more precisely. Second, compared to a model with completely heterogeneous coefficients, the panel structure model avoids the incidental parameter problem, as the number of parameters in the model is fixed. On the other hand, it controls parameter heterogeneity to a large degree.

The panel structure model has a wide variety of potential applications, as a group structure exists in almost every aspect of economic activity. A group structure may be also embedded in some abstract space. Two examples are groups of individuals defined by their location in some socioeconomic "space" (c.f. Akerlof (1997)) and groups of firms defined according to some measure of economic distance (c.f. Conley and Dupor (2003)). Direct applications of the panel structure model include the following sets of problems:

(a) **History Dependence and Multiple Equilibria.** A wealth of literature is concerned with the occurrence of multiple steady states and their history dependence in economic modeling (e.g. Arthur (1999), Deissenberg et al. (2001)). Depending on historical conditions, the system converges towards two or more distinct steady states. The panel structure model provides a natural setting to analyze this problem. Specifically, a logistic regression can be used to capture the history dependence while a linear dynamic model can be employed to characterize the dynamics associated with a particular steady state. For example, a class of growth models produces multiple steady states in income per capita or its growth rate. Cross-country growth behavior in these models typically exhibits convergence clubs, in which countries associated with the same steady state obey a common linear model. Sun (2001) used the panel structure model given here to investigate these convergence clubs and unveiled some interesting features of club composition.

(b) Empirical Sample Splitting. Many economic applications perform sample splitting in order to investigate whether a linear relationship is stable across sub-samples. Typically, the sample splitting is based on some pre-specified variable and according to some pre-specified thresholds. Quite often, the variable and the corresponding thresholds are chosen based on some subjective judgement. The so formed sub-samples are thus quite arbitrary. To remove the arbitrariness, a panel structure approach may be employed. In this approach, multiple variables may be included in the logistic regression, which, together with the linear model, provides a coherent way to split the sample and investigate the economic problem at hand. For example, to test whether financing constraints affect investment decisions, existing studies divide a sample of firms into several groups based on an *ad hoc* measure of financing constraints such as dividend-income ratio (e.g. Fazzari, Hubbard and Petersen (1988)). Similarly, to test whether liquidity constraints affect consumption decisions, previous studies split a sample on the basis of some *ad hoc* variable such as wealth (e.g. Zeldes (1989)). In both cases, different studies tend to use different measures of financial constraints. As a consequence, they reached quite different conclusions. The use of a panel structure approach may shed some new light on these controversies, as multiple measures can be used together to infer whether a firm or consumer is financially constrained.

To uncover the underlying group structure, the first step is to compute the group probabilities conditional on the data we saw. The group probabilities evidently depend on model

parameters, including the membership parameters in the logistic regression and the regression parameters in the linear regressions. Since group memberships are not observable, it is not straightforward to estimate these parameters. The purpose of this paper is to develop and investigate the maximum likelihood (ML) estimator. The ML approach allows the use of an MEM (Modified Expectation and Maximization) algorithm for maximization and provides an asymptotically efficient estimator. We establish the consistency and asymptotic normality of the ML estimator under the assumption that the number of individuals goes to infinity for a fixed time dimension. For many panel data sets, especially survey data, the cross-sectional dimension is usually large relative to the time dimension so that the consideration of large N and small T asymptotics is natural. Asymptotics along alternative directions are left for future research.

We find that there always exists a local maximizer of the likelihood function (called the local MLE), which is consistent and asymptotically normal, even if the likelihood function is unbounded. However, several local maximizers can exist for a given sample. The major difficulty lies in pinning down the correct one. Interestingly, for the panel structure model, a global maximizer is consistent when the time dimension is larger than the number of regressors in the linear regression. For many panel data sets, this mild condition holds even though the time dimension may not be large. The present paper therefore provides a novel way to overcome the problem of an unbounded likelihood function.

The unboundedness of the likelihood function in a normal mixture model was first demonstrated by Kiefer and Wolfowitz (1956); for an extended discussion, see Titterington, Smith and Makov (1985). Due to the unboundedness, a global MLE is not defined. Various approaches have been proposed to overcome this problem. For example, Hathaway (1985) proposed a constrained MLE based on maximization of the likelihood over an appropriately chosen subset of the parameter space. Geweke and Keane (2000) used a Bayesian approach and imposed Gamma priors for the precision parameters (the inverse of the variance parameters) to ensure the existence of a posterior distribution. While successful, these approaches may not be appealing as they impose *ad hoc* restrictions or priors on the parameter space to circumvent the unboundedness problem. Instead, by utilizing the time dimension variation that is present in panel data, this paper shows that the problem is resolved without imposing any restrictions on the parameter space and without the use of prior distributions. It has long been recognized by econometricians that the availability of panel data provides opportunities to identify parameters and overcome estimation problems that may be present when only cross section or time series data are available. The panel structure model provides a further instance of this phenomenon.

Another contribution of the paper is that we develop a test for homogeneity (i.e. one group) against heterogeneity (more than one group). It is well known in both the econometric and statistical literature that such a testing problem is nonstandard. For a simple mixture model with constant membership probabilities, we may have the problem that a parameter is on the boundary of the parameter space or the problem that a parameter is not identified under the null hypothesis (e.g. see Cho and White (2003) for a recent discussion on iid mixtures). Because of these problems, the expected Hessian matrix is singular under the null, which prevents us from using the empirical process approach of Andrews and Ploberger (1994, 1995) and Hansen (1996) to derive the asymptotic distribution of the likelihood ratio statistic. We show that the existence of a covariate in the logistic regression part of the panel structure model ensures the nonsingularity of the Hessian matrix and the validity of quadratic

approximation of the likelihood function. As a consequence, we can use the empirical process approach to establish the asymptotic theory of the likelihood ratio test.

The rest of the paper is organized as follows. We describe the model and investigate its identification in Section 2. Section 3 proposes the maximum likelihood estimation and describes the MEM algorithm in detail. Section 4 establishes the consistency and asymptotic normality of a local MLE. Conditions that ensure the existence and consistency of a global MLE are also presented. Section 5 examines the likelihood-ratio type of test for homogeneity and establishes the limit of the sup-LR statistic under both the null and the local alternatives. Section 6 presents two simulation studies. The first study shows that the ML estimator works well while the second study shows that the homogeneity test has good size and power properties. Section 7 concludes.

Throughout the paper, C is a generic constant that may be different across different lines. "$\equiv$" denotes definitional equivalence. $\|a\|$ is the Euclidean norm of vector a. 1_k is the $k \times 1$ vector of ones.

2 Panel Structure Model and Identification

The panel structure model consists of a set of linear regressions and a logistic regression that classifies these linear regressions. In this section, we define the panel structure model and investigate its identification.

2.1 Panel Structure Model

Let y_{it} be the dependent variable for individual i, measured at time t, $i = 1, 2, ..., N$, $t = 1, 2, ..., T$. We consider a model of the form[1]

$$\begin{aligned} y_{i0} &= \varphi_{(i)} + \phi_{(i)}\mu_i + e_i, \\ y_{it} &= \tilde{\alpha}_{(i)} + \beta_{(i)} y_{it-1} + x_{it}'\gamma_{(i)} + \tau_t'\delta_{(i)} + z_{it}'\eta + \mu_i + \varepsilon_{it}, \end{aligned} \tag{2.1}$$

where x_{it} and z_{it} are $k_1 \times 1$ and $k_2 \times 1$ explanatory variables, respectively, $\tau_t = (\tau_t^1, \tau_t^2, ..., \tau_t^{k_3})'$ is a $k_3 \times 1$ vector of deterministic time trends, e_i is iid $N(0, \sigma^2_{e,(i)})$, μ_i is iid $N(0, \sigma^2_{\mu,(i)})$, and $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ is a normal vector with mean zero and diagonal variance matrix $\sigma^2_{\varepsilon,(i)} I_T$. We assume that ε_i, μ_i and e_i are mutually independent from each other for all i and that ε_i and e_i are independent of x_{it} and z_{it} for all t.

The individual specific effect μ_i may be correlated with x_{it} and z_{it}. An appealing approach in the literature is to take explicit account of the linear dependence between μ_i and (x_{it}, z_{it}) by letting $\mu_i = \bar{x}_i' a_{(i)} + \bar{z}_i' b_{(i)} + \nu_i$ (Mundlak (1978)) and the model becomes

$$\begin{aligned} y_{i0} &= \varphi_{(i)} + \bar{x}_i' a_{(i)}\phi_{(i)} + \bar{z}_i' b_{(i)}\phi_{(i)} + \nu_i\phi_{(i)} + e_i, \\ y_{it} &= \tilde{\alpha}_{(i)} + \beta_{(i)} y_{it-1} + x_{it}'\gamma_{(i)} + \tau_t'\delta_{(i)} + z_{it}'\eta + \bar{x}_i' a_{(i)} + \bar{z}_i' b_{(i)} + \nu_i + \varepsilon_{it}, \end{aligned} \tag{2.2}$$

where now ν_i is assumed to be independent of $(x_{it},\ z_{it})$ for all t. If we include the time invariant regressors $\bar{x}_i$ and $\bar{z}_i$ in x_{it} or z_{it} and add the regressors $\bar{x}_i$ and $\bar{z}_i$ to the first equation

[1] We index the parameters by '(i)' instead of 'i' in order to distinguish, for example, $\varphi_{(i)}$ from φ_g defined below.

in (2.1), the estimation of (2.2) is formally equivalent to the estimation of (2.1). So for ease of exposition, we assume that μ_i is independent of (x_{it}, z_{it}).

Model (2.1) assumes that y_{it} follows a dynamic linear model, where some regression coefficients are population-specific whereas other coefficients are individual-specific. If all regression coefficients are the same across different individuals, the model reduces to the standard dynamic panel data model with unrestricted initial conditions (e.g. Arellano (2003, page 96), Hsiao (2003, page 75)).

The search for a group structure leads us to a model that contains G groups. The regression coefficients $\varphi_{(i)}, \phi_{(i)}, \tilde{\alpha}_{(i)}, \beta_{(i)}, \gamma_{(i)}, \delta_{(i)}$ and the variance parameters $\sigma_{e,(i)}$, $\sigma_{\mu,(i)}$, $\sigma_{\varepsilon,(i)}$ take different values depending on the group membership. Let c_{ig} be the indicator of individual i's membership of group g, i.e. $c_{ig} = 1$ if individual i belongs to group g. We assume that

$$\begin{aligned} \varphi_{(i)} &= \varphi' c_i, \phi_{(i)} = \phi' c_i, \\ \tilde{\alpha}_{(i)} &= \tilde{\alpha}' c_i, \beta_{(i)} = \beta' c_i, \gamma'_{(i)} = \gamma' c_i, \delta_{(i)} = \delta' c_i, \\ \sigma_{e,(i)} &= \sigma'_e c_i,\ \sigma_{\mu,(i)} = \sigma'_\mu c_i \text{ and } \sigma_{\varepsilon,(i)} = \sigma'_\varepsilon c_i \end{aligned} \tag{2.3}$$

where $c_i = (c_{i1}, ..., c_{iG})'$, $\varphi = (\varphi_1, \varphi_2, ..., \varphi_G)'$, $\phi = (\phi_1, \phi_2, ..., \phi_G)'$, $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2, ..., \tilde{\alpha}_G)'$, $\beta = (\beta_1, \beta_2, ..., \beta_G)', \gamma = (\gamma_1, \gamma_2, ..., \gamma_G)'$, $\delta = (\delta_1, \delta_2, ..., \delta_G)'$, $\sigma_e = (\sigma_{e,1}, \sigma_{e,2}, ..., \sigma_{e,G})'$, $\sigma_\mu = (\sigma_{\mu,1}, \sigma_{\mu,2}, ..., \sigma_{\mu,G})'$ and $\sigma_\varepsilon = (\sigma_{\varepsilon,1}, \sigma_{\varepsilon,2}, ..., \sigma_{\varepsilon,G})'$ are vectors of group parameters.

We model the distribution of c_i with a polychotomous regression, which is allowed to depend on covariates specific to individual i. Specifically, c_i is multinomially distributed with G categories. The probability π_{ig} that individual i belongs to group g is a function of $w_i = (w_{i1}, w_{i2}, ..., w_{ik_4})'$ in a multinomial logistic regression:

$$\pi_{ig}(\xi) = P(c_{ig} = 1 | w_i) = \frac{\exp(w_i' \xi_g)}{\sum_{j=1}^{G} \exp(w_i' \xi_j)}, \tag{2.4}$$

for $g = 1, 2, ..., G$. Here $\xi' = (\xi'_1, ..., \xi'_G)$ and ξ_g is a column vector containing the membership parameters for group g.

The first element of w_i is assumed to be one, which implies that the membership probabilities $\{\pi_{ig}\}_{i=1}^N$ are constant across all individuals if no covariate is included. Put another way, if we can not infer any membership information from any covariate, we assume that membership probabilities are the same for all individuals. To ensure no crossing between groups over time, we assume all the covariates are time invariant.

The multinomial logistic model is one of the most often used multinomial choice models as it yields convenient functional forms for choice probabilities. However, the model suffers from the well-known IIA (Independence from Irrelevant Alternatives) assumption. To overcome this problem, we may use the mixed logit model or the multinomial probit model and reply on Monte Carlo integration techniques to compute the membership probabilities (e.g. Train (2002)). We will leave the extension along this line to future research.

Equations (2.1), (2.3) and (2.4) combine to define the panel structure model. The underlying data generating process consists of a set of G subprocesses, which is identified with the set of unit vectors $\mathcal{E} = \{\mathcal{E}_g : g = 1, 2, ..., G\}$. First, a data point $(x'_{i1}, ..., x'_{iT}, z'_{i1}, ..., z'_{iT}, w'_i)'$ is chosen from some probability distribution λ_2 on $(\Omega_2, \mathcal{F}_2) \subseteq \mathbb{R}^{(k_1+k_2)T+k_4}$, where the probability space $(\Omega_2, \mathcal{F}_2, \lambda_2)$ will be specified below. Second, given the data point $(x'_{i1}, ..., x'_{iT}, z'_{i1}, ..., z'_{iT}, w'_i)'$, a subprocess $c_i \in \mathcal{E}$ is selected from a multinomial distribution with probabilities $\{\pi_{ig}(\xi)\}_{g=1}^G$. Finally, given that $c_i = \mathcal{E}_g$, for some g, a $(T + 2) \times 1$ vector $(e_i, \mu_i, \varepsilon_{i1}, ..., \varepsilon_{iT})$ is generated

from the multivariate normal $N(0, diag(\sigma^2_{e,g}, \sigma^2_{\mu,g}, \sigma^2_{\varepsilon,g}, ..., \sigma^2_{\varepsilon,g}))$ and the time series $\{y_{it}\}_{t=1}^T$ is generated according to

$$\begin{cases} y_{i0} = \varphi_g + \phi_g \mu_i + e_i \\ y_{it} = \tilde{\alpha}_g + \beta_g y_{it-1} + x'_{it}\gamma_g + \tau'_t \delta_g + z'_{it}\eta + \mu_i + \varepsilon_{it}, \end{cases} \tag{2.5}$$

$t = 1, 2, ..., T$. The data generating process implicitly defines the population, which is the entire collection of the data points $(x'_{i1}, ..., x'_{iT}, z'_{i1}, ..., z'_{iT}, w'_i, y'_{i0}, y'_{i1}, ..., y'_{iT})'$. A sample of size N is a subset of N individuals from the population where the subset is chosen in such a way that every individual has the same chance of being selected as any other. Although the panel structure model has a Bayesian interpretation, we adopt classical inference in this paper.

The panel structure model includes two sets of variables: one for the linear regressions and the other for the logistic regression. While the former variables affect the dependent variable through marginal responses, the latter variables exert their effects through the relationship between the dependent variable and the former variables. In some sense, the latter variables are more important as they determine which linear specification best describes the relationship between the dependent variable and the former variables. In many empirical applications, qualitative variables, such as the degree of openness and the level of democracy, affect the dependent variable. Quite often, empirical analysts employ some proxies for these variables and use these proxies in a linear regression. A potentially more fruitful exercise is to use these proxies in the logistic regression in order to identify groups of individuals that obey a common and parsimonious relationship.

When the parameters in the second equation of (2.5) are the same for all individuals, the panel structure model becomes the standard dynamic panel data model except that it initializes the dynamic process differently. The problem of initialization is important because the interpretation and consistency properties of the MLE and GLS estimator depend on how the process is initialized. Existing approaches that treat y_{i0} as random all assume that φ_g, ϕ_g, $\sigma_{e,g}$ are the same for all individuals. This assumption is certainly restrictive and the panel structure model can be used to relax it.

Although the panel structure model is specified in a dynamic form, it incorporates a static formation as a special case (i.e. when $\beta_g = 0$ for all g). In this case, the initial condition disappears and the model becomes

$$y_{it} = \tilde{\alpha}_g + x'_{it}\gamma_g + \tau'_t \delta_g + z'_{it}\eta + \mu_i + \varepsilon_{it}. \tag{2.6}$$

We will focus on the dynamic panel data model but all our results hold for the static model.

In the panel structure model, explanatory variables in the linear regressions are divided into two sets. The division is based on whether the marginal response is membership dependent. When all marginal responses are membership invariant, the panel structure model reduces to a model with homogeneous coefficients. On the other hand, when all marginal responses are membership dependent and the number of groups is equal to the cross-sectional sample size, the panel structure model becomes a model with completely heterogeneous coefficients. The inclusion of two sets of explanatory variables thus gives us more flexibility in controlling for heterogeneity.

Since the panel structure model bears at least superficial similarity to the threshold regression model (e.g. see Hansen (2000)), a few words are in order on the precise relationship. Put in our context, a panel threshold model assumes that the regression coefficients $\tilde{\alpha}_{(i)}, \beta_{(i)}, \gamma_{(i)}, \delta_{(i)}$

and the variance parameters $\sigma_{e,(i)}, \sigma_{\mu,(i)}, \sigma_{\varepsilon,(i)}$ take different values depending on the magnitude of some threshold variable w_i. When there is only one threshold, i.e. $G = 2$, the panel threshold model includes the following specification:

$$\pi_{i1} = 1\{w_i \leq s\}, \pi_{i2} = 1\{w_i > s\} \tag{2.7}$$

for some parameter s, where $1\{\cdot\}$ is the indicator function. Comparing this to (2.4), we see that the panel threshold model can be framed as a degenerate panel structure model, with group probabilities equal to zero or one. In addition, a non-degenerate panel structure model can be easily made to approximate a panel threshold model. All that is required is to note that the logistic function can be made to approximate the indicator function, i.e. $(1 + \exp[(w_i - s)\xi])^{-1} \to 1\{w_i \leq s\}$ as $\xi \to \infty$.

2.2 Identification

For a panel structure model to successfully characterize the group structure in a panel data set, it is necessary that the model be identified. Before exploring the issue of identification we first introduce some notation. Denote $v_i' = (v_{i1}', ..., v_{it}', ..., v_{iT}')$, $v_{it}' = (x_{it}', z_{it}')$ and $y_i = (y_{i1}, ..., y_{iT})'$. Let $(v', w')'$ be a random vector with probability distribution λ_2 on $(\Omega_2, \mathcal{F}_2)$, where $\Omega_2 = \left(\otimes_{t=1}^T \Omega_2^t\right) \otimes \Omega_2^w$, $\Omega_2^t \subseteq \mathbb{R}^{k_1+k_2}$, $\Omega_2^w \in \mathbb{R}^{k_4}$ and $\mathcal{F}_2 = \left(\otimes_{t=1}^T \mathcal{F}_2^t\right) \otimes \mathcal{F}_2^w$. Let y_0 be a random variable with probability distribution $\mathcal{P}_0$ on $(\Omega_0, \mathcal{F}_0)$, conditional on v and w. Let y be a random vector with probability distribution $\mathcal{P}$ on $(\Omega_1, \mathcal{F}_1)$, conditional on y_0, v, and w. With these notations, we are ready to derive the density of (y_i, y_{i0}) conditional on v_i and w_i and investigate the identification of the model.

First, conditional on the group membership g, the probability density $m^0(y_{i0}; \varphi_g, \omega_g)$ of y_{i0} with respect to the Lesbegue measure λ_0 is

$$m^0(y_{i0}; \varphi_g, \omega_g) = (2\pi)^{-1/2} \left|\omega_g^2\right|^{-1/2} \exp\left(-\frac{(y_{i0} - \varphi_g)^2}{2\omega_g^2}\right), \tag{2.8}$$

where $\omega_g^2 = \phi_g^2 \sigma_{\mu,g}^2 + \sigma_{e,g}^2$. Second, conditional on y_{i0} and group membership g, we have

$$\mu_i \sim N(\rho_g(y_{i0} - \varphi_g), \sigma_{\mu,g}^2 - \rho_g^2 \omega_g^2) \text{ where } \rho_g = \phi_g \sigma_{\mu,g}^2 / \omega_g^2. \tag{2.9}$$

Therefore, the conditional distribution of $\mu_i + \varepsilon_i = \mu_i + (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iT})'$ is multivariate normal with mean $\rho_g(y_{i0} - \varphi_g)$ and variance Σ_g :

$$\begin{aligned} \Sigma_g &= \left(\sigma_{\mu,g}^2 - \rho_g^2 \omega_g^2\right) J_T + \sigma_{\varepsilon,g}^2 I_T \\ &= \sigma_{\varepsilon,g}^2 (I_T - J_T/T) + \left[T\left(\sigma_{\mu,g}^2 - \rho_g^2 \omega_g^2\right) + \sigma_{\varepsilon,g}^2\right] J_T/T \\ &\equiv \sigma_{\varepsilon,g}^2 (I_T - J_T/T) + \sigma_{o,g}^2 J_T/T, \end{aligned} \tag{2.10}$$

where I_T is the $T \times T$ identity matrix and J_T is the $T \times T$ matrix with unity in every element. It follows that the probability density $m(y_i; \theta_g, \sigma_g, \eta)$ of y_i conditional on y_{i0}, v_i, w_i and group membership g is given by

$$m(y_i; \theta_g, \sigma_g, \eta) = (2\pi)^{-T/2} |\Sigma_g|^{-1/2} \exp\left(-\frac{1}{2} u_{i,g}' \Sigma_g^{-1} u_{i,g}\right), \tag{2.11}$$

where $\theta_g' = (\alpha_g', \beta_g', \gamma_g', \delta_g', \rho_g'), \alpha_g = \tilde{\alpha}_g - \rho_g \varphi_g$, $\sigma_g' = \left(\sigma_{o,g}', \sigma_{\varepsilon,g}'\right)$ and

$$u_{i,g} = y_i - \alpha_g - \beta_g y_{i,-1} - x_i' \gamma_g - \tau' \delta_g - z_i' \eta - \rho_g y_{i0}. \tag{2.12}$$

As a consequence, the density of (y_i, y_{i0}) conditional on v_i and w_i is

$$f(y_i, y_{i0}; \psi) = \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g), \tag{2.13}$$

where $\psi = (\xi', \eta', \theta', \sigma', \varphi', \omega')'$ and

$$\pi_{ig}(\xi) = \frac{\exp(w_i' \xi_g)}{\sum_{j=1}^{G} \exp(w_i' \xi_j)}. \tag{2.14}$$

The dominating measure for the conditional density is the Lesbegue measure $\lambda_0 \times \lambda_1$ on $\Omega_0 \otimes \Omega_1$. Since λ_2 is the probability measure on Ω_2 induced by the random variables v_i and w_i. (2.13) is also the joint density of (y_i, y_{i0}, v_i, w_i) with respect to the measure $\lambda_0 \times \lambda_1 \times \lambda_2$.

Equations (2.11), (2.13) and (2.14) provide an unrestricted parameterization of the group structure. Note that there is a one-to-one correspondence between $(\alpha_g, \beta_g', \gamma_g', \delta_g', \rho_g, \sigma_g, \varphi_g, \omega_g)$ and $(\tilde{\alpha}_g, \beta_g', \gamma_g', \delta_g', \sigma_{eg}, \sigma_{\mu,g}, \sigma_{\varepsilon,g}, \varphi_g, \phi_g)$, we will focus on $(\alpha_g, \beta_g', \gamma_g', \delta_g', \rho_g, \sigma_g, \varphi_g, \omega_g)$ and η hereafter. For this parameterization, the density $f(y_i, y_{i0}; \psi)$ is invariant under all permutations of the group labels:

$$(\xi_g', \theta_g', \sigma_g', \varphi_g', \omega_g')' \to (\xi_{per(g)}', \theta_{per(g)}', \sigma_{per(g)}', \varphi_{per(g)}', \omega_{per(g)}')' \tag{2.15}$$

for $g = 1, 2, ..., G$, where $(per(1), ..., per(G)) = per(1, ..., G)$ is any permutation of $(1, ..., G)$. The density $f(y_i, y_{i0}; \psi)$ is also invariant under the translation $\xi \to \xi + \xi^0$ for any constant vector ξ^0.

To uniquely identify the parameters, we impose restrictions to break the invariance. Specifically, we order the regression parameters to break the permutation invariance, and initialize the membership parameters to break the translation invariance. We assume throughout that (i) the parameters $\{(\theta_g', \sigma_g', \varphi_g', \omega_g')'\}_{g=1}^{G}$ are ordered lexicographically, i.e. $(\theta_g', \sigma_g', \varphi_g', \omega_g')' \prec (\theta_{g'}', \sigma_{g'}', \varphi_{g'}', \omega_{g'}')'$ if (a) $\theta_{g1} < \theta_{g'1}$ or (b) $\theta_{g1} = \theta_{g'1}$ but $\theta_{g2} < \theta_{g'2}$, and so on; (ii) the membership parameters are initialized by setting $\xi_G = 0$. Through ordering and initializing, we achieve the uniqueness of parameterization.

Now we drop the subscript i for clarity of exposition and introduce the notion of identification:

Definition: *The panel structure model is identified, if for any two parameter combinations $\psi^{(1)}$ and $\psi^{(2)}$, $f(y; \psi^{(1)}) = f(y; \psi^{(2)})$ for almost all $(y', y_0', (v', w'))' \in \Omega_1 \otimes \Omega_0 \otimes \Omega_2$ if and only if $\psi^{(1)} = \psi^{(2)}$.*

Apparently, a necessary condition for the identification of f is: $m(y; \theta_g^{(1)}, \sigma_g^{(1)}, \eta^{(1)}) = m(y; \theta_g^{(2)}, \sigma_g^{(2)}, \eta^{(2)})$ for almost all y only if $\theta_g^{(1)} = \theta_g^{(2)}$, $\sigma_g^{(1)} = \sigma_g^{(2)}$ and $\eta^{(1)} = \eta^{(2)}$. We first investigate whether this condition is satisfied. Let

$$M(\beta_g) = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \beta_g & 1 & 0 & \dots & 0 \\ \beta_g^2 & \beta_g & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \beta_g^{T-1} & \beta_g^{T-2} & \beta_g^{T-3} & \dots & 1 \end{pmatrix}, \tag{2.16}$$

$$V(\theta_g, \sigma_g) = M(\beta_g)\Sigma(\sigma_g)M'(\beta_g),\ \Sigma(\sigma_g) \equiv \Sigma_g, \tag{2.17}$$

and

$$\begin{aligned} A(\theta_g, \eta) &= M(\beta_g)x\gamma_g + M(\beta_g)z\eta + M(\beta_g)\tau\delta_g \\ &\quad + M(\beta_g)1_T\alpha_g + M(\beta_g)1_T\rho_g y_0 + y_0(\beta_g, \beta_g^2, ..., \beta_g^T)', \end{aligned} \tag{2.18}$$

where $x = (x_{\cdot 1}, x_{\cdot 2}, ..., x_{\cdot T})'$, $z = (z_{\cdot 1}, z_{\cdot 2}, ..., z_{\cdot T})'$, and $\tau = (\tau_1, \tau_2, ..., \tau_T)'$. Then $m(y; \theta_g, \sigma_g, \eta)$ can be written as

$$m(y; \theta_g, \sigma_g, \eta) = (2\pi)^{-T/2} |V(\theta_g)|^{-1/2} \exp\left(-\frac{1}{2}[y - A(\theta_g, \eta)]'V^{-1}(\theta_g, \sigma_g)[y - A(\theta_g, \eta)]\right), \tag{2.19}$$

which is a multivariate normal density. For any two vectors $(\theta_g^{(1)\prime}, \sigma_g^{(1)\prime}, \eta^{(1)\prime})'$ and $(\theta_g^{(2)\prime}, \sigma_g^{(2)\prime}, \eta^{(2)\prime})'$, $m(y; \theta_g^{(1)}, \sigma_g^{(1)}, \eta^{(1)}) = m(y; \theta_g^{(2)}, \sigma_g^{(2)}, \eta^{(2)})$ with probability one if and only if $A(\theta_g^{(1)}, \eta^{(1)}) = A(\theta_g^{(2)}, \eta^{(2)})$ and $V(\theta_g^{(1)}, \sigma_g^{(1)}) = V(\theta_g^{(2)}, \sigma_g^{(2)})$ hold with probability one. The following lemma shows that $A(\theta_g^{(1)}, \eta^{(1)}) = A(\theta_g^{(2)}, \eta^{(2)})$ and $V(\theta_g^{(1)}, \sigma_g^{(1)}) = V(\theta_g^{(2)}, \sigma_g^{(2)})$ hold with probability one if and only if $\theta_g^{(1)} = \theta_g^{(2)}$, $\sigma_g^{(1)} = \sigma_g^{(2)}$ and $\eta^{(1)} = \eta^{(2)}$. The following quantities arise in the Lemma and the subsequent theorem:

$$D_v^t(\zeta) = \{v_{\cdot t} : v_{\cdot t}'\zeta = 0\};\ D_w(\zeta) = \{w : w'\zeta = 0\}; \tag{2.20}$$

$$\lambda_2^t = \text{marginal probability distribution of } v_{\cdot t} \text{ on } (\Omega_2^t, \mathcal{F}_2^t); \tag{2.21}$$

$$\lambda_2^w = \text{marginal probability distribution of } w \text{ on } (\Omega_2^w, \mathcal{F}_2^w). \tag{2.22}$$

Lemma 1 *Assume that*
(i) For some integer $t_0 \in [1, T]$, $\lambda_2^{t_0}(D_v^{t_0}(\zeta)) < 1$ *for any* $\zeta \neq 0$, $\zeta \in \mathbb{R}^{k_1+k_2}$.
(ii) The probability distribution of y_0 *is not degenerate.*
(iii) The matrix $(\tau, 1_T)$ *has full column rank.*
If $m(y; \theta_g^{(1)}, \sigma_g^{(1)}, \eta^{(1)}) = m(y; \theta_g^{(2)}, \sigma_g^{(2)}, \eta^{(2)})$ *with probability one, then* $\theta_g^{(1)} = \theta_g^{(2)}$, $\sigma_g^{(1)} = \sigma_g^{(2)}$ *and* $\eta^{(1)} = \eta^{(2)}$.

An intermediate corollary of Lemma 1 is that if $m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)}) = m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)})$ with probability one, then $\varphi_g^{(1)} = \varphi_g^{(2)}$ and $\omega_g^{(1)} = \omega_g^{(2)}$. To see this, we note that $m^0(y_0; \varphi_g, \omega_g)$ can be viewed as a special case of $m(y; \theta_g, \sigma_g, \eta)$ with $\beta_g = \gamma_g = \delta_g = \rho_g = \sigma_{o,g} = \eta = 0$.

Lemma 1 assists in proving the following theorem, which establishes the identification of the panel structure model.

Theorem 2 *Assume*
(i) For some integer $t_0 \in [1, T]$, $\lambda_2^{t_0}(D_v^{t_0}(\zeta)) < 1$ *for any* $\zeta \neq 0$, $\zeta \in \mathbb{R}^{k_1+k_2}$*;*
(ii) The probability distribution of y_0 *is not degenerate;*
(iii) The matrix $(\tau, 1_T)$ *has full column rank;*
(iv) $\lambda_2^w(D_w(\zeta)) < 1$, *for any* $\zeta \neq 0$, $\zeta \in \mathbb{R}^{k_4}$;
Then the panel structure model is identified.

By assuming that $\lambda_2^{t_0}(D_v^{t_0}(\zeta)) < 1$ for any nonzero ζ, the theorem implicitly assumes an appropriate level of variability of $(x_{\cdot t_0}', z_{\cdot t_0}')'$. When $(x_{\cdot t_0}', z_{\cdot t_0}')'$ contains only continuous random variables, a sufficient condition is that $\lambda_2^{t_0}$ has a positive density with respect to

the Lebesgue measure on $\Omega_2^{t_0}$. When $(x'_{\cdot t_0}, z'_{\cdot t_0})'$ contains only discrete random variables, a sufficient condition is that the support of $(x'_{\cdot t_0}, z'_{\cdot t_0})'$ spans $\mathbb{R}^{k_1+k_2}$. When $(x'_{\cdot t_0}, z'_{\cdot t_0})'$ contains both continuous and discrete random variables, say $(x'_{\cdot t_0}, z'_{\cdot t_0}) = ((v^c_{t_0})', (v^d_{t_0})')$, where $v^c_{t_0}$ and $v^d_{t_0}$ contain only continuous and discrete variables respectively, sufficient conditions are: (i) $v^c_{t_0}$ has a positive density with respect to the underlying Lebesgue measure; (ii) the support of $v^d_{t_0}$ spans the space $\mathbb{R}^{k_d}$, where k_d is the number of discrete variables. The sufficient conditions are easily seen from the proof of the theorem. Similar sufficient conditions can be derived for w.

3 ML Estimation and Classification

Let $y_i = (y_{i1}, y_{i2}, ..., y_{iT})'$, $x_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, $z_i = (z_{i1}, z_{i2}, ..., z_{iT})'$, and $\tau = (\tau_1, ..., \tau_T)'$. Then conditioning on $\{x_i\}_{i=1}^N, \{z_i\}_{i=1}^N, \{w_i\}_{i=1}^N$, the log-likelihood of the panel structure model can be expressed as

$$L(\psi|y, y_0) = \sum_{i=1}^{N} \log \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g). \tag{3.1}$$

To estimate the model parameters, we maximize the log-likelihood function:

$$\hat{\psi} = (\hat{\xi}', \hat{\eta}', \hat{\theta}', \hat{\sigma}', \hat{\varphi}', \hat{\omega}')' = \arg\max_{\psi \in \Psi} L(\psi|y, y_0). \tag{3.2}$$

When $\hat{\psi}$ globally maximizes the likelihood function over the parameter space Ψ, we call it a global MLE. In contrast, when $\hat{\psi}$ maximizes the likelihood function over a closed subset of the parameter space, we call it a local MLE.

Given the estimate $\hat{\psi}$, we can assign the group membership as follows: individual i is assigned to group g if

$$\pi_{ig}(\hat{\xi}) m(y_i; \hat{\theta}_g, \hat{\sigma}_g, \hat{\eta}) m^0(y_{i0}; \hat{\varphi}_g, \hat{\omega}_g) = \max_j \pi_{ij}(\hat{\xi}) m(y_i; \hat{\theta}_j, \hat{\sigma}_j, \hat{\eta}) m^0(y_{i0}; \hat{\varphi}_j, \hat{\omega}_j). \tag{3.3}$$

In other words, we assign a given individual to the group of which it is most likely to be a member based on the posterior probabilities. Such an assignment rule is consistent with loss minimization for the 0-1 loss function. More specifically, let $Loss(g_1, g_2)$ be the price paid for assigning a group g_1 individual to group g_2. For a given assignment rule $\mathcal{A}(y_i, y_{i0}, v_i, w_i)$, the expected loss is

$$E\left(Loss\right) = E \sum_{g=1}^{G} Loss(g, \mathcal{A}(y_i, y_{i0}, v_i, w_i)) P\left(g|y_i, y_{i0}, v_i, w_i\right), \tag{3.4}$$

where $P\left(g|y_i, y_{i0}, v_i, w_i\right)$ is the probability of belonging to group g conditional on (y_i, y_{i0}, v_i, w_i) and the expectation is taken with respect to the joint distribution of (y_i, y_{i0}, v_i, w_i). With the 0-1 loss function where all mis-assignments are charged a single unit, the expected loss becomes

$$E\left(Loss\right) = 1 - E \sum_{g=1}^{G} 1\left\{\mathcal{A}(y_i, y_{i0}, v_i, w_i) = g\right\} P\left(g|y_i, y_{i0}, v_i, w_i\right). \tag{3.5}$$

Minimizing the above expected loss gives

$$\mathcal{A}(y_i, y_{i0}, v_i, w_i) = g_0 \text{ if } P\left(g_0|y_i, y_{i0}, v_i, w_i\right) = \max_{g=1,\ldots,G} P\left(g|y_i, y_{i0}, v_i, w_i\right). \tag{3.6}$$

Note that $P\left(g|y_i, y_{i0}, v_i, w_i\right)$ is proportional to $\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)$. So the assignment rule (3.3) minimizes the expected loss for the 0-1 loss conditioning on the parameter estimates.

To search for maximizers of the likelihood function, the so-called Expectation Maximization (EM) algorithm can be used (Dempster, Laird and Rubin (1977)). The EM algorithm is a general technique for maximum likelihood estimation in a wide variety of situations best described as the incomplete data problem. The recent monograph by McLachlan and Krishman (1996) provides an excellent introduction to the EM algorithm.

An application of the EM algorithm generally begins with the observation that the optimization of the likelihood function would be simplified if a set of missing variables or hidden variables were known. In our context, if the group membership c_i is observable, then the log-likelihood for $\{y_{i0,}y_i\}$ and $\{c_i\}$ becomes

$$L(\psi|y, y_0, c) = \sum_{i=1}^{N}\sum_{g=1}^{G} c_{ig}\left\{\log \pi_{ig}(\xi) + \log m(y_i;\theta_g,\sigma_g,\eta) + \log m^0(y_{i0};\varphi_g,\omega_g)\right\}. \tag{3.7}$$

The use of the indicator variable c_{ig} has allowed the logarithm to be brought inside the summation sign, substantially simplifying the maximization problem. But c_{ig} is not observable. Instead of maximizing $L(\psi|y, y_0, c)$ itself, we maximize the expectation of $L(\psi|y, y_0, c)$, where the expectation is taken with respect to all the unobserved c_{ig}. In the expectation step, the conditional expectation of $L(\psi|y, y_0, c)$ given $\{y_i, y_{i0}, v_i, w_i\}$ is calculated. In the maximization step, the so obtained expected log-likelihood is maximized with respect to ψ, which provides an updated estimate of ψ. Finally, we keep iterating between the E-step and the M-step until convergence is attained. Dempster, Laird and Rubin (1977) showed that the likelihood function $L(\psi|y, y_0)$ increases along the iterated parameter values. We modify the EM algorithm for the panel structure model and show that the modified EM (MEM) algorithm converges to a local maximum of the likelihood function. The effectiveness of the MEM algorithm is demonstrated by the simulation experiments in Section 6.

Let $\psi^{(k)} = (\xi^{(k)\prime}, \eta^{(k)\prime}, \theta^{(k)\prime}, \sigma^{(k)\prime}, \varphi^{(k)\prime}, \omega^{(k)\prime})'$ be the current estimate of ψ, and $\psi^{(k+1)} = (\xi^{(k+1)\prime}, \eta^{(k+1)\prime}, \theta^{(k+1)\prime}, \varphi^{(k+1)\prime}, \omega^{(k+1)\prime})'$ stand for the updated estimate. We follow the steps below:

The E-step: The conditional expectation is given by

$$\begin{aligned} Q(\psi|\psi^{(k)}) &= E\left(L(\psi|y, y_0, c)|\psi^{(k)}\right) \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)}\left\{\log \pi_{ig}(\xi) + \log m(y_i;\theta_g,\sigma_g,\eta) + \log m^0(y_{i0};\varphi_g,\omega_g)\right\}, \end{aligned} \tag{3.8}$$

where $p_{ig}^{(k)} = p_{ig}(\psi^{(k)})$ and

$$\begin{aligned} p_{ig}\left(\psi\right) &= E(c_{ig}|y_i, y_{i0}, v_i, w_i; \psi) \\ &= \frac{\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)}{\sum_{j=1}^{G}\pi_{ij}(\xi)m(y_i;\theta_j,\sigma_j,\eta)m^0(y_{i0};\varphi_j,\omega_j)}. \end{aligned} \tag{3.9}$$

The modified M-step: To get the updated estimate $\psi^{(k+1)}$, we maximize $Q(\psi|\psi^{(k)})$ with respect to ψ. By inspection, parameters θ_g, σ_g and η affect Q only through the term $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m(y_i;\theta_g,\sigma_g,\eta)$; parameters φ_g and ω_g affect Q only through the term $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m^0(y_{i0};\varphi_g,\omega_g)$; and parameter ξ affects Q only through the term $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log \pi_{ig}(\xi)$. Therefore, we can maximize these three terms of $Q(\psi|\psi^{(k)})$ separately.

(i) We first maximize $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log \pi_{ig}(\xi)$ with respect to ξ. This objective function may be seen as a log-likelihood in a multinomial logistic regression with fractional observations $\{p_{ig}^{(k)}\}$. It is also similar to the log-likelihood function in a multinomial logistic regression for grouped data. The first order conditions are easily shown to be

$$\sum_{i=1}^{N}\left(p_i^{(k)} - \pi_i(\xi)\right)\otimes w_i = 0, \tag{3.10}$$

where $p_i^{(k)} = (p_{i1}^{(k)}, p_{i2}^{(k)}, ..., p_{iG}^{(k)})'$ and $\pi_i(\xi) = (\pi_{i1}(\xi), \pi_{i2}(\xi)..., \pi_{iG}(\xi))'$. Intuitively, $\xi^{(k+1)}$ should be chosen so that $\pi_{ig}(\xi^{(k+1)})$ is as close to $p_{ig}^{(k)}$ as possible. Since the first element of w_i is 1, we have

$$\frac{1}{N}\sum_{i=1}^{N}\pi_{ig}(\xi) = \frac{1}{N}\sum_{i=1}^{N} p_{ig}^{(k)} \text{ for } g = 1, 2, ..., G. \tag{3.11}$$

In other words, the predicted shares are the same as the 'observed' shares. Some calculations show that the Hessian of $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log \pi_{ig}(\xi)$ is given by $H_\xi(\psi) = -\sum_{i=1}^{N} cov_i(\xi)\otimes w_i w_i'$ where

$$cov_i \equiv cov_i(\xi) = \begin{pmatrix} \pi_{i1}(\xi) - \pi_{i1}^2(\xi) & -\pi_{i2}(\xi)\pi_{i1}(\xi) & ... & -\pi_{iG}(\xi)\pi_{i1}(\xi) \\ -\pi_{i1}(\xi)\pi_{i2}(\xi) & \pi_{i2}(\xi) - \pi_{i2}^2(\xi) & ... & ... \\ ... & ... & ... & ... \\ -\pi_{i1}(\xi)\pi_{iG}(\xi) & ... & ... & \pi_{iG}(\xi) - \pi_{iG}^2(\xi) \end{pmatrix}. \tag{3.12}$$

Note that cov_i is the covariance matrix of the multinomially distributed vector $(c_{i1}, c_{i2}, ..., c_{iG})$ with parameters $(\pi_{i1}, \pi_{i2}, ..., \pi_{iG})$, cov_i is positive semi-definite and $H_\xi(\psi)$ is negative semi-definite. When ξ_G is normalized to be zero, we need to remove the last k_4 rows and k_4 columns of $H_\xi(\psi)$ to get the Hessian $\tilde{H}_\xi(\psi)$ for the remaining parameters. It is easy to show that $\tilde{H}_\xi(\psi)$ is negative definite and the maximand $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log \pi_{ig}(\xi)$ is globally concave. We can find the optimal ξ by the Newton-Raphson algorithm using $\tilde{\xi}^{(k)}$ as the starting point:

$$\tilde{\xi}^{(k+1)} = \tilde{\xi}^{(k)} - \tilde{H}_\xi^{-1}\left(\psi^{(k)}\right)\left(\sum_{i=1}^{N}\left(\tilde{p}_i^{(k)} - \tilde{\pi}_i(\xi^{(k)})\right)\otimes w_i\right), \tag{3.13}$$

where $\tilde{\xi}' = (\tilde{\xi}_1', \tilde{\xi}_2', ..., \tilde{\xi}_{G-1}')$, $\tilde{p}_i^{(k)} = (p_{i1}^{(k)}, p_{i2}^{(k)}, ..., p_{iG-1}^{(k)})'$ and $\tilde{\pi}_i(\xi) = (\pi_{i1}(\xi), \pi_{i2}(\xi)..., \pi_{iG-1}(\xi))'$. For notational convenience, we have assumed that Newton-Raphson algorithm stops after one iteration. In practice, we may have to use the above iterative formula a few times to attain convergence.

(ii) We next maximize $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m(y_i;\theta_g,\sigma_g,\eta)$ with respect to θ_g, σ_g and η. This is a nonlinear optimization problem, which has no closed form solution. To accommodate the

iterative nature of our estimating strategy, we propose to maximize $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m(y_i;\theta_g,\sigma_g,\eta)$ with respect only to the regression parameters for a fixed Σ_g. This is an easier task as it is equivalent to solving the minimization problem:

$$\min_{\theta_g,\eta} \sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} u_{i,g}' \left(\Sigma_g^{(k)}\right)^{-1} u_{i,g}, \tag{3.14}$$

where $u_{i,g} = y_i - \Gamma_i\theta_g - z_i\eta$, $\Gamma_i = (1_T, y_{i,-1}, x_i, \tau, 1_T y_{i0})$ and $y_{i,-1} = (y_{i0}, y_{i1}, ..., y_{i,T-1})$. Using the well-known result that

$$\left(\Sigma_g^{(k)}\right)^{-1/2} = \left(\sigma_{o,g}^{(k)}\right)^{-1} J_T/T + \left(\sigma_{\varepsilon,g}^{(k)}\right)^{-1} (I_T - J_T/T), \tag{3.15}$$

we have $(\Sigma_g^{(k)})^{-1/2} u_{i,g} = (\sigma_{\varepsilon,g}^{(k)})^{-1} \tilde{u}_{i,g}^{(k)}$, where

$$\tilde{u}_{i,g}^{(k)} = \left(y_i - \varrho_g^{(k)}\bar{y}_i\right) - \left(\Gamma_i - \varrho_g^{(k)}\bar{\Gamma}_i\right)\theta_g - \left(z_i - \varrho_g^{(k)}\bar{z}_i\right)'\eta \tag{3.16}$$

and $\varrho_g^{(k)} = 1 - \sigma_{\varepsilon,g}^{(k)}/\sigma_{o,g}^{(k)}$. Therefore, the minimization problem in (3.14) reduces to

$$\min_{\theta_g,\eta} \sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \tilde{u}_{i,g}^{(k)\prime} \tilde{u}_{i,g}^{(k)} / \left(\sigma_{\varepsilon,g}^{(k)}\right)^2. \tag{3.17}$$

Let $\tilde{y}_i^{(k)} = y_i - \varrho_g^{(k)}\bar{y}_i$, $\tilde{\Gamma}_{ig}^{(k)} = \Gamma_i - \varrho_g^{(k)}\bar{\Gamma}_i$, and $\tilde{z}_i^{(k)} = z_i - \varrho_g^{(k)}\bar{z}_i$. Then some algebraic manipulations show that the solution to (3.17) is:

$$\eta^{(k+1)} = \left\{\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \left(z_{ig}^o\right)' \left(\sigma_{\varepsilon,g}^{(k)}\right)^{-2} z_{ig}^o\right\}^{-1} \left\{\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \left(z_{ig}^o\right)' \left(\sigma_{\varepsilon,g}^{(k)}\right)^{-2} y_{ig}^o\right\}, \tag{3.18}$$

$$\left(\theta_g^{(k+1)}\right)' = \left\{\sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' \left(\tilde{\Gamma}_{ig}^{(k)}\right)\right\}^{-1} \left\{\sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' y_i^o\right\}, \tag{3.19}$$

where $y_i^o = \tilde{y}_i^{(k)} - \tilde{z}_i^{(k)}\eta^{(k+1)}$ and

$$\begin{aligned} z_{ig}^o &= \tilde{z}_i^{(k)} - \tilde{\Gamma}_i^{(k)} \left\{\sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' \left(\tilde{\Gamma}_{ig}^{(k)}\right)\right\}^{-1} \sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' \tilde{z}_i^{(k)}, \\ y_{ig}^o &= \tilde{y}_i^{(k)} - \tilde{\Gamma}_i^{(k)} \left\{\sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' \left(\tilde{\Gamma}_{ig}^{(k)}\right)\right\}^{-1} \sum_{i=1}^{N} p_{ig}^{(k)} \left(\tilde{\Gamma}_{ig}^{(k)}\right)' \tilde{y}_i^{(k)}. \end{aligned} \tag{3.20}$$

Here y_i^o, y_{ig}^o, and z_{ig}^o should have an additional superscript (k), which is omitted for notational simplicity.

It remains to update the variance parameters $\sigma_{\varepsilon,g}^2$ and $\sigma_{o,g}^2$. Let $u_{i,g}^{(k)} = y_i - \Gamma_i^{(k)}\theta_g^{(k)} - z_i\eta^{(k)}$. Plugging the regression parameters in (3.18) and (3.19) into $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m(y_i;\theta_g,\sigma_g,\eta)$

yields:

$$\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \left\{ -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_g| - \frac{1}{2}\left(u_{i,g}^{(k)}\right)' \Sigma_g^{-1} u_{i,g}^{(k)} \right\}$$
$$= \sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \left\{ -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log\sigma_{\varepsilon,g}^{2(T-1)} - \frac{1}{2}\log(\sigma_{o,g}^2) \right.$$
$$\left. -\frac{1}{2T\sigma_{o,g}^2}\left(u_{i,g}^{(k+1)}\right)' J_T u_{i,g}^{(k+1)} - \frac{1}{2\sigma_{\varepsilon,g}^2}\left(u_{i,g}^{(k)}\right)' (I_T - J_T/T)\, u_{i,g}^{(k)} \right\} \tag{3.21}$$

where we have used $|\Sigma_g| = \sigma_{\varepsilon,g}^{2(T-1)}\sigma_{o,g}^2$ and

$$\Sigma_g^{-1} = \sigma_{o,g}^{-2} J_T/T + \sigma_{\varepsilon,g}^{-2}(I_T - J_T/T). \tag{3.22}$$

Maximizing (3.21) produces the updated maximum likelihood estimators:

$$(\sigma_{\varepsilon,g}^2)^{(k+1)} = \frac{\sum_{i=1}^{N} p_{ig}^{(k)} \left(u_{i,g}^{(k)}\right)' (I_T - J_T/T)\, u_{i,g}^{(k)}}{(T-1)\sum_{i=1}^{N} p_{ig}^{(k)}}, \tag{3.23}$$

$$(\sigma_{o,g}^2)^{(k+1)} = \frac{\sum_{i=1}^{N} p_{ig}^{(k)} \left(u_{i,g}^{(k)}\right)' J_T u_{i,g}^{(k)}}{T\sum_{i=1}^{N} p_{ig}^{(k)}}. \tag{3.24}$$

(iii) Finally, we maximize $\sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}^{(k)} \log m^0(y_{i0};\varphi_g,\omega_g)$ to update $\varphi_g^{(k)}$ and $\omega_g^{(k)}$. It is easy to see that the solution is

$$\varphi_g^{(k+1)} = \left(\sum_{i=1}^{N} p_{ig}^{(k)} y_{i0}\right) / \left(\sum_{i=1}^{N} p_{ig}^{(k)}\right), \tag{3.25}$$

$$(\omega_g^2)^{(k+1)} = \left(\sum_{i=1}^{N} p_{ig}^{(k)} (y_{i0} - \varphi_g^{(k)})^2\right) / \left(\sum_{i=1}^{N} p_{ig}^{(k)}\right). \tag{3.26}$$

The MEM algorithm reduces the maximization problem into a sequence of logit regressions and weighted least squares regressions. The computational cost of the MLE for a panel structure model is thus comparable to that of the MLE for a dynamic panel data model. The idea of dividing the model parameters into different blocks and updating one block at a time while holding other blocks constant can be used to attack other missing data problems.

We proceed to investigate the convergence of the MEM algorithm. The following theorem establishes the relationship between the MEM algorithm and a gradient based optimization algorithm:

Theorem 3 *The updating scheme in equations (3.13), (3.18), (3.19), and (3.23)–(3.26) satisfies*

$$\psi^{(k+1)} - \psi^{(k)} = -H_{mem}^{-1}(\psi^{(k)}) \frac{\partial L(\psi|y,y_0)}{\partial\psi}|_{\psi=\psi^{(k)}}$$

for some matrix $H_{mem}(\psi^{(k)})$. In addition, if the model is identified and N is large enough, then $H_{mem}(\psi^{(k)})$ is negative definite with probability one.

To ensure and speed up the convergence of the MEM algorithm, we may multiply the step size by a positive number, say κ, to get

$$\psi^{(k+1)} - \psi^{(k)} = -\kappa H_{mem}^{-1}(\psi^k)\frac{\partial L(\psi^{(k)}|y, y_0)}{\partial \psi}. \tag{3.27}$$

As in the Newton-Raphson algorithm or any other gradient-based algorithms, we may do a grid search over $\kappa \in [0, 1]$ to pick up the optimal κ at each iteration.

Equations (3.23), (3.24) and (3.26) use $\theta_g^{(k)}$, $\eta^{(k)}$ and $\varphi_g^{(k)}$ to update σ_g and ω_g. If we use $\theta_g^{(k+1)}$, $\eta^{(k+1)}$ and $\varphi_g^{(k+1)}$ instead, we obtain

$$(\sigma_{\varepsilon,g}^2)^{(k+1)} = \frac{\sum_{i=1}^N p_{ig}^{(k)} \left(u_{i,g}^{(k+1)}\right)' (I_T - J_T/T)\, u_{i,g}^{(k+1)}}{(T-1)\sum_{i=1}^N p_{ig}^{(k)}}, \tag{3.28}$$

$$(\sigma_{o,g}^2)^{(k+1)} = \frac{\sum_{i=1}^N p_{ig}^{(k)} \left(u_{i,g}^{(k+1)}\right)' J_T u_{i,g}^{(k+1)}}{T\sum_{i=1}^N p_{ig}^{(k)}}, \tag{3.29}$$

$$(\omega_g^2)^{(k+1)} = \frac{\sum_{i=1}^N p_{ig}^{(k)} (y_{i0} - \varphi_g^{(k+1)})^2}{\sum_{i=1}^N p_{ig}^{(k)}}. \tag{3.30}$$

In this case, Theorem 3 does not hold for the above updating formulae. However, for these updating formulae, we have $Q(\psi^{(k+1)}|\psi^{(k)}) \geq Q(\psi^{(k}|\psi^{(k)})$. This is because

$$\begin{aligned} Q(\psi^{(k+1)}|\psi^{(k)}) &= Q(\xi^{(k+1)}, \eta^{(k+1)}, \theta^{(k+1)}, \sigma^{(k+1)}, \varphi^{(k+1)}, \omega^{(k+1)}|\psi^{(k)}) \\ &\geq Q(\xi^{(k+1)}, \eta^{(k+1)}, \theta^{(k+1)}, \sigma^{(k)}, \varphi^{(k+1)}, \omega^{(k)}|\psi^{(k)}) \\ &\geq Q(\xi^{(k)}, \eta^{(k)}, \theta^{(k)}, \varphi^{(k)}, \sigma^{(k)}, \omega^{(k)}|\psi^{(k)}) = Q(\psi^{(k}|\psi^{(k)}), \end{aligned} \tag{3.31}$$

where the two inequalities follow by definition. It now follows from the general EM theory of Dempster, Laird and Rubin (1977) that $L(\psi^{(k+1)}|y, y_0) \geq L(\psi^{(k)}|y, y_0)$. In our particular situation,

$$Q(\psi|\psi^{(k)}) - L(\psi|y, y_0) = \sum_{i=1}^N \sum_{g=1}^G p_{ig}(\psi^{(k)}) \log p_{ig}(\psi). \tag{3.32}$$

Hence

$$\begin{aligned} Q(\psi^{(k+1)}|\psi^{(k)}) - L(\psi^{(k+1)}|y, y_0) &= \sum_{i=1}^N \sum_{g=1}^G p_{ig}(\psi^{(k)}) \log p_{ig}(\psi^{(k+1}) \\ \leq \sum_{i=1}^N \sum_{g=1}^G p_{ig}(\psi^{(k)}) \log p_{ig}(\psi^{(k}) &= Q(\psi^{(k)}|\psi^{(k)}) - L(\psi^{(k)}|y, y_0), \end{aligned} \tag{3.33}$$

which implies that $L(\psi^{(k+1)}|y, y_0) - L(\psi^{(k)}|y, y_0) \geq Q(\psi^{(k+1)}|\psi^{(k)}) - Q(\psi^{(k)}|\psi^{(k)}) \geq 0$. Therefore, the MEM algorithm will converge to a local maximum of the likelihood function, no matter which updating scheme is used to update the variance parameters.

The MEM algorithm calls for an initial estimate of ψ. To initialize, we set $\xi_g^{(0)} = 0$, so that $\pi_{ig} = 1/G$ for $g = 1, 2, ..., G$, and randomly assign individuals to G groups. That is, for each individual i, we randomly generate an integer between 1 and G. If this random integer equals g, then we assign individual i to group g. With this assignment, we can obtain the initial estimate $\psi^{(0)}$. The MEM algorithm can then start from this initial estimate.

4 Consistency and Asymptotic Normality of the ML Estimator

To establish the consistency and asymptotic normality of the ML estimator, we need the following assumptions.

Assumption 1 $(y'_{i0}, x'_{it}, z'_{it}, w'_{it})$ are cross-sectionally independent for all t, and $E(y'_{i0}, x'_{it}, z'_{it}, w'_i)'(y'_{i0}, x'_{is}, z'_{is}, w'_{is}) < \infty$, for all t and s.

Assumption 2 ε_i is cross-sectionally independent, and ε_i is independent of $(y'_{i0}, x'_{it}, z'_{it}, w'_{it})$ for all i and t.

Assumption 3 $(e_i, \mu_i, \varepsilon_i)$ is a normal vector with mean zero and scalar variance matrix $diag(\sigma^2_{e,g}, \sigma^2_{\mu,g}, \sigma^2_{\varepsilon,g}, ..., \sigma^2_{\varepsilon,g})$ given that individual i belongs to group g.

Assumption 4 $Ey^4_{i0} < \infty, E\left\|x_{it}\right\|^4 < \infty, E\left\|z_{it}\right\|^4 < \infty, E\left\|w_i\right\|^4 < \infty$ for all i and t.

Assumption 5 The true parameter $\psi_0 = (\xi'_0, \eta'_0, \theta'_0, \sigma'_0, \varphi'_0, \omega'_0)'$ is an interior point of a compact space Ψ.

Assumptions 1 and 2 impose cross-sectional independence, an assumption that may be restrictive for some economic applications. However, because of the lack of natural ordering, there is no completely satisfactory and general way of modelling cross-sectional dependence, although some important progresses have been made, see, for example, Conley (1999), Phillips and Sul (2003) and Andrews (2003). In this paper, we follow the large panel data literature and maintain the assumption of cross sectional independence.

The distributional assumption in Assumption 3 validates the use of the ML estimator. Under this assumption, y_i follows a mixture distribution with multivariate normal components. In view of the fact that any continuous distribution can be arbitrarily well approximated by large enough normal mixtures, the normality assumption is not as restrictive as it seems. Nevertheless, it is worthwhile to relax the distributional assumption and only impose some moment conditions, even if this may invalidate the convenient MEM algorithm. Assumption 3 also imposes temporal independence of $\{\varepsilon_{it}\}_{t=1}^T$, which facilitates the presentation. Our results should be easily extended to allow for general forms of weak temporal dependence at the cost of notational complications.

Assumption 4 imposes some moment conditions, which is used only in proving the asymptotic normality of the MLE. Assumption 4 is stronger than necessary and is in a form which makes a brief proof possible. Assumption 5 is standard in the nonlinear estimation literature. It is necessary for the MLE to be asymptotically normal.

4.1 Consistency

Before we investigate the consistency of the ML estimator, we present a standard lemma that provides sufficient conditions for the existence of a consistent sequence of maximizers in a stochastic maximization problem. Next, we apply the lemma to the maximization of the likelihood function in (3.1).

Let $Q_N(\psi) = 1/N \sum_{i=1}^N q(U_i; \psi)$, for $N \geq 1$ be a sequence of maximands for estimation of the parameter $\psi \in K \subset \mathbb{R}^k$, where K is a compact parameter space. Let ψ_0 be the true value of the parameter.

Lemma 4 *(a) Suppose*
(i) $\{q(U_i; \psi) : i \geq 1\}$ *are independently and identically distributed.*

(ii) $q(U;\psi)$ is continuous at each $\psi \in K$ with probability one.
(iii) There is a function $d(U)$ with $|q(U;\psi)| \leq d(U)$ for all $\psi \in K$ and $Ed(U) < \infty$.
Then $Q_N(\psi)$ converges to $Eq(U;\psi)$ uniformly over $\psi \in K$.

(b) In addition,
(iv) for any $\psi \in K$, $\psi \neq \psi_0$, $Eq(U;\psi) < Eq(U;\psi_0)$.
Then $\hat{\psi} = \arg\max_K Q_N(\psi)$ is consistent as $N \to \infty$.

Let $q(U_i;\psi) = \log f(y_i, y_{i0};\psi)$ and $K = \Psi$. To establish the consistency of an MLE, a typical argument would verify all the conditions in Lemma 4. However, the domination condition (iii) is not satisfied if the variance parameters $\{\omega_g\}_{g=1}^G$ of $\psi \in \Psi$ are not bounded below from zero.

To illustrate this point, consider $G = 2$ and $T = 2$. In this case, $q(U_i, \psi)$ equals

$$\log \left[\pi_{i1}(\xi) m(y_i; \theta_1, \sigma_1, \eta) \frac{1}{\sqrt{2\pi\omega_1^2}} \exp\left(-\frac{1}{2} \frac{(y_{i0} - \varphi_1)^2}{\omega_1^2} \right) \right.$$
$$\left. + \pi_{i2}(\xi) m(y_i; \theta_2, \sigma_2, \eta) \frac{1}{\sqrt{2\pi\omega_2^2}} \exp\left(-\frac{1}{2} \frac{(y_{i0} - \varphi_2)^2}{\omega_2^2} \right) \right]. \tag{4.1}$$

If we set $\psi = \tilde{\psi}$ such that $\tilde{\omega}_1 = 0$, $\tilde{\omega}_2 > 0$, $\tilde{\sigma}_{\varepsilon,g}^2 > 0$, $\tilde{\sigma}_{o,g}^2 > 0$, $\pi_{ig}(\tilde{\xi}) > 0$ for all $i = 1, ..., N$, $g = 1, ..., G$, and $\tilde{\varphi}_1 = y_{i^*,0}$ for some i^*, then $q(U_{i^*}, \tilde{\psi}) = \infty$. As a consequence, $E \sup_{\psi \in K} |q(U_{i^*}, \psi)| \geq Eq(U_{i^*}, \tilde{\psi}) = \infty$, and condition (iii) can not be true. In addition, for $i \neq i^*$ such that $y_{i0} \neq \tilde{\varphi}_1$,

$$q(U_i, \tilde{\psi}) = \log \left[\pi_{i2}(\tilde{\xi}) m(y_i; \tilde{\theta}_2, \tilde{\sigma}_2, \tilde{\eta}) \frac{1}{\sqrt{2\pi\omega_2^2}} \exp\left(-\frac{1}{2} (y_{i0} - \varphi_2)^2 \right) \right] > -\infty. \tag{4.2}$$

Therefore, $Q_N(\tilde{\psi}) = \infty$.

Because of the unboundedness, a global MLE always fails to exist. This is a major difficulty in using ML to estimate a normal mixture model. The difficulty can be overcome by imposing some restriction on the parameter space (Redner (1981), Hathaway (1985)). For example, we may assume that $\omega_g \geq c_0, g = 1, ..., G$ for some positive constant c_0. But c_0 is not known *a priori*, such a restriction is thus quite arbitrary. Fortunately, even with the unconstrained parameter space Ψ, we can still prove the consistency of a local MLE.

The idea of the proof is to show that all the conditions in Lemma 4 are satisfied if we consider the parameters in a small neighborhood of the true value. Specifically, let K be the closed ball centered on the true parameter with radius r. We choose r to be small so that for all possible parameter values in K we have $\omega \in [\omega_{\min}, \omega_{\max}]$, $\sigma_{\varepsilon,g} \in [\sigma_{\varepsilon,\min}, \sigma_{\varepsilon,\max}]$, $\sigma_{o,g} \in [\sigma_{o,\min}, \sigma_{o,\max}]$, $0 < \omega_{\min} < \omega_{\max} < \infty$, $0 < \sigma_{\varepsilon,\min} < \sigma_{\varepsilon,\max} < \infty$, and $0 < \sigma_{o,\min} < \sigma_{o,\max} < \infty$ for $g = 1, 2, ..., G$. For the so chosen compact set K, conditions (i) and (ii) are trivially satisfied. Conditions (iii) and (iv) are established in the following lemma.

Lemma 5 *Let Assumptions 1, 2, 3, and 5 hold. If the panel structure model is identified, then*
(a) $E \sup_{\psi \in K} |\log f(y_i, y_{i0};\psi)| < \infty$.
(b) for any $\psi \in K$, $\psi \neq \psi_0$, $E \log f(y_i, y_{i0};\psi) < E \log f(y_i, y_{i0};\psi_0)$.
(c) there exists a consistent local MLE.

The consistency result in Lemma 5 seems to contradict the well-known inconsistency of the ML estimator for a dynamic panel model with fixed effects. In the latter case, the number of parameters goes to infinity, which leads to the incidental parameter problem. In contrast, the number of the parameters in the panel structure model above is fixed, and so the incidental parameter problem is avoided from the beginning. However, for the panel structure model, even if the number of the groups goes to infinity so that the number of parameters goes to infinity, we conjecture that there exists a consistent local ML estimator, as long as the number of individuals grows faster than the number of groups. This is because infinitely many cross-sectional observations are available for estimating a finite number of parameters specific to a particular group.

Lemma 5 only shows the existence of a consistent local maximizer, but does not propose a method to isolate a consistent one among possibly multiple and local maximizers. If a consistent estimate is available, then the local maximizer closest to it is also consistent. Therefore, the problem of selecting a consistent local MLE amounts to searching for a consistent estimate. We now propose such a consistent estimate of $\psi_c = (\xi', \eta', \theta', \sigma')'$. Let

$$f_c(y_i; \psi_c) = \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) \tag{4.3}$$

be the density of y_i conditional on y_{i0}, x_i, z_i and w_i. Define the conditional maximum likelihood estimator as:

$$\hat{\psi}_c = \arg \max_{\psi_c \in \Psi_c} \sum_{i=1}^{N} \log f_c(y_i; \psi_c), \tag{4.4}$$

where Ψ_c is the coordinate projection of the parameter space Ψ, i.e.

$$\Psi_c = \left\{ (\xi', \eta', \theta', \sigma')' : (\xi', \eta', \theta', \sigma', \varphi', \omega')' \in \Psi \text{ for some } \varphi, \omega \right\}. \tag{4.5}$$

The following theorem shows that when the time dimension is large enough, the conditional MLE is consistent.

Theorem 6 *Let Assumptions 1, 2, 3, and 5 hold. If the panel structure model is identified and $T > k_1 + k_2 + k_3 + 3$, then $\hat{\psi}_c$ is consistent.*

To prove the theorem, we first show that: (i) the maximum value of the conditional likelihood function over Ψ_c is achieved at $\psi \in \Psi_c^0$ with probability one when N is large enough, where Ψ_c^0 is the set of parameter values in Ψ_c satisfying $\sum_{g=1}^{G} \sigma_{\varepsilon,g}\sigma_{o,g} \geq r$ for some positive constant r. In other words,

$$\hat{\psi}_c = \arg \max_{\psi \in \Psi_c^0} \sum_{i=1}^{N} \log f_c(y_i; \psi_c), \tag{4.6}$$

with probability one when N is large enough. So we can take Ψ_c^0 as the parameter space in an attempt to establish the consistency. For any vector $\psi_c \in \Psi_c^0$, there exists at least one g_0 such that $\sigma_{\varepsilon,g_0}\sigma_{o,g_0} > r/G > 0$. This ensures that $\log f_c(y_i, \psi_c)$ is continuous for $\psi_c \in \Psi_c^0$. We next observe that: (ii) $y_i - \Gamma_i\theta_g - z_i\eta = 0 > 0$ for any θ_g and η, with probability one for all i. This holds because $T > k_1 + k_2 + k_3 + 3$ and ε_i is assumed to be (conditionally) normal. The

essential point is that θ_g and η can not be chosen such that some time series is perfectly fitted. Time series information thus ensures that $\log f_c(y_i, \psi_c)$ is bounded and uniformly integrable over $\psi_c \in \Psi_c$. These two steps help verify the conditions in Lemma 4. The details of the proof are provided in the appendix.

Given the consistent estimator $\hat{\psi}_c$, we can construct consistent estimators of φ and ω using the full maximum likelihood approach. We maximize $L(\psi|y, y_0)$ and choose the maximizer $\hat{\psi}^* = \left(\hat{\xi}^{*\prime}, \hat{\eta}^{*\prime}, \hat{\theta}^{*\prime}, \hat{\sigma}^{*\prime}, \hat{\varphi}^{*\prime}, \hat{\omega}^{*\prime}\right)'$ such that the subvector $(\hat{\xi}^{*\prime}, \hat{\eta}^{*\prime}, \hat{\theta}^{*\prime}, \hat{\sigma}^{*\prime})'$ is closest to $\hat{\psi}_c$. This method avoids the unbounded problem because in large samples no element of $\hat{\omega}^*$ converges to zero. The claim can be proved by contradiction. If an element of $\hat{\omega}^*$ converges to zero, then there is a group of cross sectional units whose initial observations are the same and membership probabilities of belonging to the group that they form converge to one. But this group is formed by pure coincidence (i.e. their initial observations y_0 happen to be the same). As a result, $\hat{\xi}^*$ can not be consistent and $\hat{\psi}^*$ can not be the local maximum with subvector $(\hat{\xi}^{*\prime}, \hat{\eta}^{*\prime}, \hat{\theta}^{*\prime}, \hat{\sigma}^{*\prime})'$ that is closest to $\hat{\psi}_c$. This argument shows that the so-chosen $\hat{\psi}^*$ must be a consistent local maximum rather than the inconsistent global maximum.

In the simulation study below, we first obtain $\hat{\psi}_c$ using the MEM algorithm for the conditional likelihood function and construct initial estimates of φ and ω using (3.25) and (3.26). We then use these estimates as the starting point to implement the MEM algorithm for the full likelihood function. These steps ensure that the local maximum of the full likelihood function we obtained is the one that is closest to $\hat{\psi}_c$.

It is important to point out that the first observations $\{y_{i0}\}$ are used only to improve the efficiency of the MLE. The efficiency improvement comes at the cost of additional computing time. If the time dimension is reasonably large, then efficiency improvement is expected to be small, in which case the conditional MLE is recommended for practical use.

4.2 Asymptotic Normality

In this subsection, we show that the consistent maximizer is asymptotically normal. Typically, the asymptotic normality of an ML estimator calls for domination and integrability conditions, which hold for the panel structure model under Assumption 4.

Before presenting the asymptotic normality result, we introduce some notation. Let

$$s_i(\psi) = \frac{\partial}{\partial \psi} \log f(y_i, y_{i0}; \psi) \text{ and } h_i(\psi) = \frac{\partial^2}{\partial \psi \partial \psi'} \log f(y_i, y_{i0}; \psi). \tag{4.7}$$

Denote

$$S(\psi) = \frac{1}{N} \sum_{i=1}^{N} s_i(\psi) \text{ and } H(\psi) = \frac{1}{N} \sum_{i=1}^{N} h_i(\psi). \tag{4.8}$$

Let $\Psi_0 = \{\psi : ||\psi - \psi_0|| \leq \epsilon\}$, where ϵ is chosen to be small so that $\omega \in [\omega_{\min}, \omega_{\max}]$, $\sigma_{\varepsilon,g} \in [\sigma_{\varepsilon,\min}, \sigma_{\varepsilon,\max}]$, $\sigma_{o,g} \in [\sigma_{o,\min}, \sigma_{o,\max}]$, $0 < \omega_{\min} < \omega_{\max} < \infty$, $0 < \sigma_{\varepsilon,\min} < \sigma_{\varepsilon,\max} < \infty$, and $0 < \sigma_{o,\min} < \sigma_{o,\max} < \infty$ for $g = 1, 2, ..., G$ for all possible parameter values in Ψ_0. With the consistency result, it suffices to take Ψ_0 as the parameter space in attempting to establish the asymptotic normality.

Theorem 7 *Let Assumptions 1-5 hold. Then*

(a) $\int \sup_{\psi \in \Psi_0} \left| \frac{\partial}{\partial \psi_r} f(y_i, y_{i0}; \psi) \right| d\lambda_0 d\lambda_1 d\lambda_2 < \infty$, *for* $r = 1, 2, ...,$

(b) $\int \sup_{\psi\in\Psi_0} \left|\frac{\partial^2}{\partial\psi_r\partial\psi_s} f(y_i, y_{i0};\psi)\right| d\lambda_0 d\lambda_1 d\lambda_2 < \infty$, *for* $r, s = 1, 2, ...,$

(c) $E\sup_{\psi\in\Psi_0} \left|\frac{\partial^2}{\partial\psi_r\partial\psi_s} \log f(y_i, y_{i0};\psi)\right| < \infty$, *for* $r, s = 1, 2, ...,$

(d) $\sqrt{N}S(\psi_0) \Rightarrow N(0, \mathcal{I})$ *where* $\mathcal{I}$ *is the Fisher information matrix*

$$\mathcal{I} = Es_i(\psi_0)s_i'(\psi_0). \tag{4.9}$$

(e) If the panel structure model is identified, the solution $\hat{\psi} = \arg\max_{\psi\in\Psi_0} \sum_{i=1}^N \log f(y_i, y_{i0};\psi)$ *is asymptotically normally distributed and*

$$\sqrt{N}(\hat{\psi} - \psi_0) \Rightarrow N(0, \mathcal{I}^{-1}). \tag{4.10}$$

The proof of the theorem is given in the appendix.

Part (c) of Theorem 7 ensures that a uniform law of large numbers holds for $H(\psi) = 1/N\sum_{i=1}^N h_i(\psi)$. Together with the consistency of $\hat{\psi}$, this result implies that $-H(\hat{\psi})$ is a consistent estimate of the information matrix $\mathcal{I}$. However, for panel structure models with many covariates, the calculation of the Hessian matrix $h_i(\psi)$ is algebraically tedious. Hence we use the outer product of the score function to approximate the information matrix. From equation (4.9), the information matrix can be estimated by

$$\frac{1}{N}\sum_{i=1}^N s_i(\hat{\psi})s_i'(\hat{\psi}). \tag{4.11}$$

This is validated by the uniform convergence of $\sum_{i=1}^N s_i(\psi)s_i'(\psi)/N$ and the consistency of $\hat{\psi}$. The uniform convergence follows easily from Lemma 4 by noting that

$$E\sup_{\psi\in K}\left|\frac{\partial f(y_i, y_{i0};\psi)}{\partial\psi_r}\frac{\partial f(y_i, y_{i0};\psi)}{\partial\psi_s'}\frac{1}{f^2}\right| < \infty, \tag{4.12}$$

which is implied by Part (c) of Theorem 7.

The consistency and asymptotic normality results are established under the assumption that the model is correctly specified. The results may be extended to a mis-specified model in the direction of White (1994), in which case the estimator converges to a value which makes the model closest to the true model in the sense of the Kullback-Leibler distance.

5 A Homogeneity Test

The previous sections assume that the population is heterogeneous due to the presence of a group structure. A practical question is: does the group structure really exist or is G really greater than 1? One way to address this question is to consider testing the null $H_0 : G = 1$ against the alternative $H_1 : G = G_1$ for some $G_1 > 1$. It has long been pointed out in the statistical literature that such a problem is not regular for simple mixture models with constant mixing probabilities. In this section, we provide a novel way to establish the asymptotic distribution of the likelihood ratio statistic.

To motivate our method, consider a simple mixture model in which the density of Y is $\pi f(y, \alpha_1) + (1-\pi) f(y, \alpha_2)$ for $\alpha_1, \alpha_2 \in A \subseteq \mathbb{R}$ and $\pi \in [0, 1]$. Our objective is to test $H_0 : Y \sim f(y, \alpha_0)$ for some unknown $\alpha_0 \in A$ against $H_1 : Y \sim \pi f(y, \alpha_1) + (1-\pi) f(y, \alpha_2)$

for some unknown $\alpha_1 \neq \alpha_2$ and $\pi \neq 0, 1$. Given the maintained hypothesis, the null can be formally described as

$$H_0 : \alpha_1 = \alpha_2, \text{ or } \pi = 0, \text{ or } \pi = 1. \tag{5.13}$$

With this formulation, we can see that several standard assumptions are violated, leading to the nonstandard limiting distribution of the likelihood ratio statistic. First, when $\pi = 0$ or 1, we have the problem that the true parameter is on the boundary of the parameter space; see Andrews (2001). In addition, α_1 is not identified when $\pi = 0$ while α_2 is not identified when $\pi = 1$. Second, when $\alpha_1 = \alpha_2$, π is not identified. We have the problem that a nuisance parameter is present only under the alternative; see Davies (1977, 1987). In this case, one may attempt to use the empirical process approach of Andrews and Ploberger (1994, 1995) and Hansen (1996) to derive the limiting distribution of the LR statistic. However, when $\alpha_1 = \alpha_2 = \alpha_0$, the expected Hessian matrix under the null hypothesis is

$$-\begin{pmatrix} \pi^2 & \pi(1-\pi) \\ \pi(1-\pi) & (1-\pi)^2 \end{pmatrix} E\left(\frac{\partial \log f(Y,\alpha)}{\partial \alpha}\right)^2 |_{\alpha=\alpha_0} \tag{5.14}$$

for any given π, where the expectation is taken under the null that $Y \sim f(y, \alpha_0)$. Obviously, the expected Hessian matrix is singular, which prevents us from using a quadratic approximation of the log-likelihood function for a given π. Indeed, Andrews and Ploberger (1994, 1995) and Hansen (1996) pointed out that their techniques are not directly applicable to simple mixture models.

These two problems are potentially present for the panel mixture model. However, when the membership probabilities depend on some covariates W, the boundary problem disappears and the expected Hessian matrix is no longer singular. This is an important observation that enables us to derive the asymptotic theory for the LR test. To illustrate the basic idea, consider a mixture model in which the density of Y conditional on a covariate $W = w$ is given by

$$\pi(w, \xi) f(y, \alpha_1) + (1 - \pi(w, \xi)) f(y, \alpha_2), \tag{5.15}$$

where

$$\pi(w, \xi) = \exp(w\xi)/(1 + \exp(w\xi)). \tag{5.16}$$

This model is different from the simple mixture model with constant membership probabilities in the following aspects. First, when the parameter ξ lies in a compact space, the membership probabilities are always greater than 0 and less than 1. Because of this, the null hypothesis of homogeneity holds only under the condition that $\alpha_1 = \alpha_2$. Note that the compactness of the parameter space is a mild condition that is typically assumed in nonlinear models. Second, the membership probabilities are random variables. The expected Hessian becomes

$$-\begin{pmatrix} E\pi(W,\xi)^2 \left(\frac{\partial \log f(Y,\alpha_0)}{\partial \alpha}\right)^2 & E\pi(W,\xi)(1-\pi(W,\xi)) \left(\frac{\partial \log f(Y,\alpha_0)}{\partial \alpha}\right)^2 \\ E\pi(W,\xi)(1-\pi(W,\xi)) \left(\frac{\partial \log f(Y,\alpha_0)}{\partial \alpha}\right)^2 & E(1-\pi(W,\xi))^2 \left(\frac{\partial \log f(Y,\alpha_0)}{\partial \alpha}\right)^2 \end{pmatrix}, \tag{5.17}$$

which is not singular under certain conditions. For example, when $\pi(W, \xi)$ is independent of $\partial \log f(Y, \alpha_0)/\partial \alpha$, the expected Hessian is proportional to

$$\begin{pmatrix} E\pi(W,\xi)^2 & E[\pi(W,\xi)(1-\pi(W,\xi))] \\ E[\pi(W,\xi)(1-\pi(W,\xi))] & E[1-\pi(W,\xi)]^2 \end{pmatrix}, \tag{5.18}$$

which is in general non-singular because

$$\left\{E\left[\pi(W,\xi)\left(1-\pi(W,\xi)\right)\right]\right\}^2 < E\pi(W,\xi)^2 E\left[1-\pi(W,\xi)\right]^2. \tag{5.19}$$

Only when π is a constant such that π is proportional to $(1-\pi)$, the expected Hessian matrix is singular.

We now proceed to examine the LR test for $H_0 : G = 1$ against the alternative $H_1 : G > 1$. Recall that

$$f(y_i, y_{i0}; \psi) = \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g). \tag{5.20}$$

For notational convenience, let

$$m(y_i, y_{i0}; \zeta_g, \eta) = m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g), \tag{5.21}$$

where $\zeta_g = \left(\theta_g', \sigma_g', \varphi_g', \omega_g'\right)'$. Then the unrestricted log-likelihood is

$$L_{ur}(\xi, \vartheta) = \sum_{i=1}^{N} \log \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i, y_{i0}; \zeta_g, \eta), \tag{5.22}$$

where $\vartheta' = (\zeta_1', ..., \zeta_G', \eta')$. Let $\hat{\vartheta}_{ur} := \hat{\vartheta}_{ur}(\xi)$ be the unrestricted maximum likelihood estimator for any given ξ and $\vartheta_0' = (\zeta_0', \zeta_0', ..., \zeta_0', \eta_0')$ be the true parameter value under the null hypothesis. In the appendix, we prove the quadratic approximation:

$$\begin{aligned} &2(L_{ur}(\xi, \hat{\vartheta}_{ur}) - L_{ur}(\xi, \vartheta_0)) \\ &= \left(\frac{1}{\sqrt{N}} \frac{\partial L_{ur}(\xi, \vartheta_0)}{\partial \vartheta}\right)' \mathcal{I}_{ur}^{-1}(\xi, \vartheta_0) \frac{1}{\sqrt{N}} \frac{\partial L_{ur}(\xi, \vartheta_0)}{\partial \vartheta} + o_p(1), \end{aligned} \tag{5.23}$$

uniformly over $\xi \in \Xi$, where

$$\mathcal{I}_{ur}(\xi, \vartheta) = -E \frac{1}{N} \frac{\partial^2_{ur} L(\xi, \vartheta)}{\partial \vartheta \partial \vartheta'}. \tag{5.24}$$

Here $\partial L_{ur}(\xi, \vartheta_0)/\partial \vartheta$ is understood to be $\partial L_{ur}(\xi, \vartheta)/\partial \vartheta$ evaluated at $\vartheta = \vartheta_0$. We use the same convention hereafter.

The restricted log-likelihood is

$$L_r(\xi, \vartheta_r) = \sum_{i=1}^{N} \log m(y_i, y_{i0}; \zeta_r, \eta_r) \text{ where } \zeta_r = \left(\theta_1', \sigma_1', \varphi_1', \omega_1'\right)'. \tag{5.25}$$

and $\vartheta_r = (\zeta_r', \zeta_r', ..., \zeta_r', \eta_r')'$. Let $\hat{\vartheta}_r = (\hat{\zeta}_r', \hat{\zeta}_r', ..., \hat{\zeta}_r', \hat{\eta}_r')'$ be the restricted MLE. Following the same argument, we have the quadratic approximation:

$$\begin{aligned} &2(L_r(\xi, \hat{\vartheta}_r) - L_r(\xi, \vartheta_0)) \\ &= \left(\frac{1}{\sqrt{N}} \frac{\partial L_r(\xi, \vartheta_0)}{\partial (\zeta_r', \eta_r')'}\right)' \mathcal{I}_r^{-1}(\xi, \zeta_0, \eta_0) \left(\frac{1}{\sqrt{N}} \frac{\partial L_r(\xi, \vartheta_0)}{\partial (\zeta_r', \eta_r')'}\right) + o_p(1) \end{aligned} \tag{5.26}$$

uniformly over $\xi \in \Xi$, where

$$\mathcal{I}_r(\xi, \zeta_r, \eta_r) = -E \frac{1}{N} \frac{\partial^2 L_r(\xi, \vartheta_r)}{\partial (\zeta_r', \eta_r')' \partial (\zeta_r', \eta_r')}. \tag{5.27}$$

In view of $L_{ur}(\xi,\vartheta_0)=L_r(\xi,\vartheta_0)$, we get

$$\begin{aligned}
&2(L_{ur}(\xi,\hat{\vartheta}_{ur})-L_r(\xi,\hat{\vartheta}_r))\\
&=\left(\frac{1}{\sqrt{N}}\frac{\partial L_{ur}(\xi,\vartheta_0)}{\partial\vartheta}\right)'\mathcal{I}_{ur}^{-1}(\xi,\vartheta_0)\frac{1}{\sqrt{N}}\frac{\partial L_{ur}(\xi,\vartheta_0)}{\partial\vartheta}\\
&-\left(\frac{1}{\sqrt{N}}\frac{\partial L_r(\xi,\vartheta_0)}{\partial\left(\zeta_r',\eta_r'\right)'}\right)'\mathcal{I}_r^{-1}(\xi,\zeta_0,\eta_0)\left(\frac{1}{\sqrt{N}}\frac{\partial L_r(\xi,\vartheta_0)}{\partial\left(\zeta_r',\eta_r'\right)'}\right)+o_p(1),
\end{aligned}\tag{5.28}$$

uniformly over $\xi\in\Xi$. Let

$$R=\begin{pmatrix}1_G'\otimes I_{k_1+k_2+7} & 0\\ 0 & I_{k_2}\end{pmatrix}\tag{5.29}$$

be a $[(k_1+k_3+7)+k_2]$ by $[G(k_1+k_3+7)+k_2]$ matrix, then it is easy to show that

$$\frac{\partial L_r(\xi,\vartheta_0)}{\partial\left(\zeta_r',\eta_r'\right)'}=R\frac{\partial L_{ur}(\xi,\vartheta_0)}{\partial\vartheta}.\tag{5.30}$$

Therefore,

$$-2(L_r(\xi,\hat{\vartheta}_r)-L_{ur}(\xi,\hat{\vartheta}_{ur}))=NS_N'(\xi)\,Q(\xi)S_N(\xi)+o_p(1),\tag{5.31}$$

uniformly over $\xi\in\Xi$, where

$$\begin{aligned}
Q(\xi)&=\mathcal{I}_{ur}^{-1/2}(\xi,\vartheta_0)\left[I-\mathcal{I}_{ur}^{1/2}(\xi,\vartheta_0)R'\left(R\mathcal{I}_{ur}(\xi,\vartheta_0)R'\right)^{-1}R\mathcal{I}_{ur}^{1/2}(\xi,\vartheta_0)\right]\mathcal{I}_{ur}^{-1/2}(\xi,\vartheta_0),\\
S_N(\xi)&=\frac{1}{N}\frac{\partial L_{ur}(\xi,\vartheta_0)}{\partial\vartheta}=\frac{1}{N}\sum_{i=1}^{N}s_i(\xi,\vartheta_0),
\end{aligned}\tag{5.32}$$

and

$$s_i(\xi,\vartheta_0)=\begin{pmatrix}\pi_i(\xi)\otimes\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)/\partial\zeta\\ \partial\log m(y_i,y_{i0};\zeta_0,\eta_0)/\partial\eta\end{pmatrix}.\tag{5.33}$$

Let

$$\mathbb{I}=\begin{pmatrix}\mathbb{I}_{\zeta\zeta} & \mathbb{I}_{\zeta\eta}\\ \mathbb{I}_{\zeta\eta}' & \mathbb{I}_{\eta\eta}\end{pmatrix}:=E\frac{\partial\log m(y_i,y_{i0};\zeta,\eta)}{\partial(\zeta',\eta')'}\left(\frac{\partial\log m(y_i,y_{i0};\zeta,\eta)}{\partial(\zeta',\eta')'}\right)'\Big|_{\zeta=\zeta_0,\eta=\eta_0}.\tag{5.34}$$

Then we have the following theorem.

Theorem 8 *Let Assumptions 1-5 hold. Define* $LR(\xi):=-2\left[L_r(\xi,\hat{\vartheta}_r)-L_{ur}(\xi,\hat{\vartheta}_{ur})\right]$. *If the panel structure model is identified and the support of* $\tilde{\pi}_i=(\pi_{i,1},...,\pi_{i,G-1})'$ *spans* $\mathbb{R}^{G-1}$, *then under the null hypothesis of homogeneity,*

$$LR(\xi)\Rightarrow W(\xi)'Q(\xi)W(\xi)\text{ for }\xi\in\Xi,\tag{5.35}$$

where $W(\xi)$ *is a Gaussian process with mean 0 and covariance kernel*

$$C(\xi_1,\xi_2)=cov\left(W(\xi_1),W(\xi_2)\right)=\begin{pmatrix}E\pi(\xi_1)\pi'(\xi_2)\otimes\mathbb{I}_{\zeta\zeta} & E\pi(\xi_1)\otimes\mathbb{I}_{\zeta\eta}\\ E\pi'(\xi_2)\otimes\mathbb{I}_{\zeta\eta}' & \mathbb{I}_{\eta\eta}\end{pmatrix},$$

and

$$\sup_{\xi\in\Xi}LR(\xi)\Rightarrow\sup_{\xi\in\Xi}W(\xi)'Q(\xi)W(\xi).\tag{5.36}$$

Theorem 8 gives the limiting distribution of the empirical process $LR(\xi)$. For each ξ, the marginal distribution of the limiting process is a Chi-square distribution. Hence, $W(\xi)'Q(\xi)W(\xi)$ is called a Chi-square process. In this paper, we use $\sup LR(\xi)$, a functional of the $LR(\xi)$ process as our test statistic. We can also use other functionals of $LR(\xi)$, such as an average exponential form of $LR(\xi)$ (Andrews and Ploberger (1994, 1995)), as the test statistics. The extension along this line is left for future research.

To investigate the power of the sup-LR test, we consider a sequence of local alternatives parameterized by

$$\xi_N = \xi_0, \eta_N = \eta_0,\ \zeta_{Ng} = \zeta_0 + \frac{d_g}{\sqrt{N}},\ g = 1, 2, 3, ..., G \tag{5.37}$$

for some vectors $\{d_g\}_{g=1}^G$. We have the following theorem, which establishes the consistency of the sup-LR test. The condition that $d \in D_A$ in the theorem is a technical condition that facilitates the proof.

Theorem 9 *Let Assumptions 1-5 hold. If the panel structure model is identified and the support of* $\tilde{\pi}_i = (\pi_{i,1}, ..., \pi_{i,G-1})'$ *spans* $\mathbb{R}^{G-1}$, *then under the local alternatives,*

$$\lim_{\|d\|\to\infty, d\in D_A} \lim_{N\to\infty} P\left\{\sup_{\xi\in\Xi} LR(\xi) > CV\right\} = 1 \tag{5.38}$$

for any finite constant CV *where*

$$d = (d_1', d_2', d_3', ..., d_G', \mathbf{0}_{k_3}')',$$

$\mathbf{0}_{k_3}$ *is the* $k_3 \times 1$ *vector of zeros and*

$$D_A = \{d : \textit{none of the elements of } \mathcal{I}(\xi_0, \vartheta_0)d \textit{ is zero}\}.$$

6 Monte Carlo Simulations

In this section, we conduct two sets of Monte Carlo experiments. In the first set, we investigate the finite sample properties of the ML estimator. In the second set, we investigate the finite sample performance of the homogeneity test.

6.1 ML Estimation and Classification

To evaluate the finite sample properties of the MLE, we consider a simple dynamic panel data model:

$$\begin{aligned} y_{i0} &= \varphi + \phi\mu_i + e_i \\ y_{it} &= \alpha + \beta y_{i,t-1} + \mu_i + \varepsilon_{it}, \end{aligned} \tag{6.1}$$

where μ_i is iid $N(0, \sigma_\mu^2)$, ε_{it} is iid $N(0, \sigma_\varepsilon^2)$ and e_i is iid $N(0, \sigma_e^2)$ and μ_i, e_i and ε_{it} are mutually independent.

For simplicity, we consider the case of two groups. For the first group, the model parameters are chosen such that $\{y_{it}\}$ is stationary. More specifically, the parameters for the first group satisfy

$$\phi = \frac{1}{1-\beta},\ \sigma_e^2 = \frac{1}{1-\beta^2}\sigma_\varepsilon^2 \text{ and } \varphi = \frac{\alpha}{1-\beta}. \tag{6.2}$$

Given this, the free parameters in the first group are α, β, σ_μ^2, and σ_ε^2. Without the loss of generality, we normalize σ_ε^2 to be 1. As a result, we only need to choose α, β and σ_μ^2 to generate the data for the first group. We set $\alpha = 0$ and consider all possible combinations of $\beta = 0.25, 0.5$ and $\sigma_\mu^2 = 0.5, 1, 1.5$. For the second group, all parameters except β are the same as those in the first group. When $\beta = 0.25$ for the first group, we consider $\beta = 0.5, 0.75$ for the second group. When $\beta = 0.5$ for the first group, we consider $\beta = 0.75, 1$ for the second group.

We consider the case of two covariates in the logistic regression. We let $w_{i1} = 1$ and w_{i2} be standard normal. We set $\xi_1 = (0, 3)$ and normalize $\xi_2 = (0, 0)$. For the so-chosen ξ, the two groups are of the same size on average. We have also considered other values for ξ_1 so that the two groups are of different sizes but the qualitative results are similar.

The experiments are carried out for a wide range of (N, T) combinations. For each (N, T) combination, there are 12 different data generating processes. We perform 1000 replications. A single experiment consists of the following steps:

(a) Let $w_{i1} = 1$ and draw w_{i2} from the standard normal.

(b) Calculate

$$\pi_{ig} = \frac{\exp\left(\xi_{g1} + w_{i2}\xi_{g2}\right)}{\sum_{k=1}^{G} \exp\left(\xi_{k1} + w_{i2}\xi_{k2}\right)}, g = 1, 2, \tag{6.3}$$

and generate a uniformly distributed variable $\nu_i \in [0, 1]$. If $\nu_i \leq \pi_{i1}$, assign individual i to group 1. Otherwise, assign individual i to group 2.

(c) Let individual i be assigned to group g in step (b). Construct $\{y_{it}\}$ according to

$$\begin{aligned} y_{i0} &= \varphi_g + \phi_g \mu_i + \sigma_{e,g} e_i \\ y_{it} &= \alpha_g + \beta_g y_{i,t-1} + \sigma_{\mu,g} \mu_i + \sigma_{\varepsilon,g} \varepsilon_{it}, \end{aligned} \tag{6.4}$$

where $(\mu_i, e_i, \varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iT})$ is iid $N(0, I_{T+2})$.

(d) Assuming that all model parameters are different across the two groups, compute the MLE using the algorithm outlined in Section 3.

(e) Classify cross section units according to (3.3) and compute the percentage of correct classifications.

For each of the regression parameters including $\alpha_g, \beta_g, \varphi_g$ and ϕ_g, we report the bias and standard deviation of the MLE. The bias and standard deviation for an estimator, say $\hat{\beta}_g$, are computed as follows:

$$bias(\hat{\beta}_g) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\beta}_g^b - \beta_g^0\right), \tag{6.5}$$

$$sd(\hat{\beta}_g) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\beta}_g^b - \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_g^b\right)^2, \tag{6.6}$$

where $\hat{\beta}_g^b$ is the estimate for the b-th replication and β_g^0 is the true value. These numbers are of obvious interest as we want to see whether the MLE is substantially biased and whether

the ML approach provides an estimator with reasonable precision. We also construct a 95% confidence interval (CI) based on the asymptotic distribution, and report the true coverage probability of the CI. For example, the CI for dynamic parameter β_1 is

$$\left(\hat{\beta}_1 - 1.96\hat{s}_{\beta_1}, \hat{\beta}_1 + 1.96\hat{s}_{\beta_1}\right), \tag{6.7}$$

where $\hat{\beta}_1$ is the ML estimate and $\hat{s}_{\beta_1}$ is its asymptotic standard error. The coverage probability is the percentage of the number of times that the true value belongs to the CI. If the asymptotic theory is reasonably reliable, we should find the coverage probability to be close to 95%. Finally, we report the average and the standard deviation of the percentage of correct classifications, in an attempt to see to what extent the model can correctly classify the cross sectional units.

We focus on the estimation of the β_1 and β_2. Tables 1 and 2 report the biases and standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the coverage probabilities of 95% CI's constructed as above. We report the cases $N = 100, 200$ and omit the case $N = 400$ to save space. The tables show that the asymptotic results are reflected in the finite sample scenarios. For all the cases considered, the biases of the ML estimates are quite small, and the empirical coverage probabilities are in general close to 95%, indicating the reliability of the asymptotic approximation. As we expect, for a given σ_μ^2, the absolute bias and standard error decrease as N or T increases. For a given (N, T) combination, the absolute bias and standard error tend to decrease as σ_μ^2 increases, although this pattern is not very clear. All else being equal, the absolute bias and standard error become smaller as β_2 moves away from β_1. This is not surprising, as the larger the difference between β_1 and β_2 is, the easier it is to detect the group structure.

Tables 3 and 4 report the percentage of correct classifications for all data generating processes and for different (N, T) combinations. For each given values of σ_μ^2 and (β_1, β_2), we first use the estimated parameters to calculate the percentage of correct classification and then use the true parameters to carry out the same calculation. In the table, the former is denoted as $\hat{P}$ and the latter is denoted as P^*. Since the true parameters are not known in practice, P^* is not feasible in empirical applications and is used here only as a benchmark. As it is clear from the tables, when $(\beta_1, \beta_2) = (0.25, 0.50), (0.50, 0.75)$, the average of P^* is around 86%. Therefore, even if the true parameters are known, there is still some uncertainty about group memberships. Note that the average of $\hat{P}$ is always lower than that of P^*, classification is thus not a trivial problem for the data generating processes we considered.

We now discuss the results on $\hat{P}$. First, for a given N, the average of $\hat{P}$ increases as T increases. This is expected, as the longer a time series is, the more membership information it reveals. Second, for a given T, the average of $\hat{P}$ also increases as N increases. The reason for this is that a larger N leads to more precise estimates and more reliable classification. However, a larger N also means more individuals to be classified. $\hat{P}$ is thus expected to become closer to P^* as N goes to infinity. This observation is corroborated by the results reported in Tables 3 and 4. Third, $\hat{P}$ increases as σ_μ^2 increases. This pattern is consistent with the modest decrease of both absolute bias and standard error as σ_μ^2 increases. We may conclude that the larger the individual effects (compared to the idiosyncratic effects) are, the easier it is to detect the group memberships. Finally, $\hat{P}$ increases substantially as β_2 moves away from β_1. This is well expected as it is easier to classify the cross sectional units when their difference is larger.

The main results of the first set of Monte Carlo experiments can be summarized as follows. The ML estimator performs quite well. The asymptotic results are quite reliable, at least for

the (N, T) combinations considered here. If the data generating process is correctly specified, the panel structure model provides a reasonably reliable classification.

6.2 Homogeneity Test

In this subsection, we investigate the finite sample performance of the homogeneity test. The data generating process is the same as in the previous subsection. For each of $\sigma_\mu^2 = 0.5, 1, 1.5$, we consider the following combinations of β_1 and β_2 :

β_1	0	0	0	0	0	0	0	0	0	0	0
β_2	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50

For all the combinations, we set $\xi_1 = (0, 3)$ and normalize $\xi_2 = (0, 0)$. When $(\beta_1, \beta_2) = (0, 0)$, the null hypothesis of homogeneity holds, in which case the percentage of rejections gives the empirical size of the test. For other combinations, the alternative of two groups holds, in which case, the percentage of rejections gives us the finite sample power of the test.

To carry out the sup-LR test, we need to maximize the log-likelihood function

$$L(\psi|y, y_0) = \sum_{i=1}^{N} \log \sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g) \tag{6.8}$$

for each ξ in the parameter space Ξ. For a given ξ, the maximum value of $L(\psi|y, y_0)$ gives us $L_r(\xi, \hat{\vartheta}_r)$ when $G = 1$ and gives us $L_{ur}(\xi, \hat{\vartheta}_{ur})$ when $G = 2$. By definition, $LR(\xi) = -2\left[L_r(\xi, \hat{\vartheta}_r) - L_{ur}(\xi, \hat{\vartheta}_{ur})\right]$. Since Ξ is a continuous parameter space, calculation of $LR(\xi)$ is excessively costly. To overcome this problem, we replace Ξ by a discrete approximation $\Xi_A := \{(\xi_1, \xi_2) \mid \xi_2 = (0, 0),\ \xi_1 = (0, \xi_{12}) \text{ for } \xi_{12} = (0, 0.5, 1, 1.5, ..., 10)\}$. The test statistic we use is then $\sup_{\xi \in \Xi_A} LR(\xi)$. The reason for choosing Ξ_A is that the probability density function (pdf) of $\pi_{i1}(\xi)$ has a wide range of shapes. It is easy to see that when w_i is standard normal, the pdf of $\pi_{i1}(\xi)$ is given by

$$pdf(x) = \frac{1}{x - x^2} \frac{1}{\sqrt{2\pi}\xi_{12}} \exp\left(-\frac{\left(\log\left(x^{-1} - 1\right)\right)^2}{2\xi_{12}^2}\right). \tag{6.9}$$

Figure 1 graphs the pdf for $\xi_{12} = 0.5, 1.0, ..., 2.5$, illustrating the richness of the possible distributions when $\xi \in \Xi_A$. It should be pointed out that replacing Ξ by Ξ_A has implications on the power of the test; see Hansen (1996) for a detailed discussion.

Theorem 8 establishes the asymptotic distribution of the sup-LR statistic. However, the limiting distribution is nonstandard and nonpivotal. To obtain the critical values for the sup-LR test, we use the resampling technique proposed by Hansen (1996). Although a bootstrapping method is also an option, it requires repeated maximizations of the likelihood function. Following Hansen (1996), we outline the main steps in generating the statistic that has the same asymptotic distribution as the sup-LR statistic:

(i) Generate $\{e_{i,j},\ i = 1, 2, ..., N,\ j = 1, 2, ..., J\}$ iid $N(0, 1)$ random variables.

(ii) Compute

$$\hat{Q}_N(\xi) := \hat{\mathcal{I}}_{ur}^{-1/2}(\xi, \hat{\vartheta}_r)\left[I - \hat{\mathcal{I}}_{ur}^{1/2}(\xi, \hat{\vartheta}_r)R'\left(R\hat{\mathcal{I}}_{ur}(\xi, \hat{\vartheta}_r)R'\right)^{-1} R\hat{\mathcal{I}}_{ur}^{1/2}(\xi, \hat{\vartheta}_r)\right]\hat{\mathcal{I}}_{ur}^{-1/2}(\xi, \hat{\vartheta}_r), \tag{6.10}$$

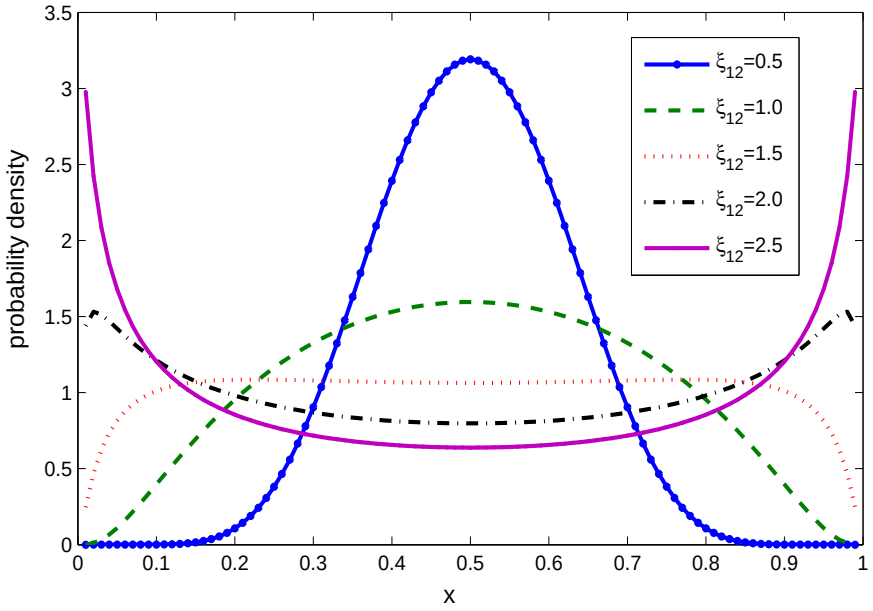

Figure 1: Probability Density of $\pi_{i1}(\xi)$ for Different Values of ξ

where $\hat{\vartheta}_r = \left(\hat{\zeta}_r', \hat{\zeta}_r', \hat{\eta}_r'\right)'$ is the ML estimate of ϑ under the null hypothesis and

$$\hat{\mathcal{I}}_{ur}(\xi, \hat{\vartheta}_r) = -\frac{1}{N}\frac{\partial^2 L_{ur}(\xi, \vartheta)}{\partial\vartheta\partial\vartheta'}|_{\vartheta=\hat{\vartheta}_r}. \tag{6.11}$$

To reduce the computational cost, we estimate $\mathcal{I}_{ur}(\xi, \vartheta_0)$ using the outer-product form:

$$\hat{\mathcal{I}}_{ur}(\xi, \hat{\vartheta}_r) = \frac{1}{N}\sum_{i=1}^{N} s_i(\xi, \hat{\vartheta}_r)s_i'(\xi, \hat{\vartheta}_r). \tag{6.12}$$

(iii) For each $j = 1, 2, ..., J$, compute

$$W_N^j(\xi) := \frac{1}{\sqrt{N}}\sum_{i=1}^{N} s_i(\xi, \hat{\vartheta}_r)e_{i,j}, \tag{6.13}$$

where $s_i(\xi, \vartheta)$ is the individual score function defined in (5.33). It is important to note that $W_N^j(\xi)$ converges weakly to $W(\xi)$ as $N \to \infty$. This result can be proved using the conditional central limit theorem; see Pollard (1990, Theorem 10.2) and Hansen (1996, Theorem 2).

(iv) For each $j = 1, 2, ..., J$, compute

$$\widehat{LR}^j(\xi) := W_N^j(\xi)'\hat{Q}_N(\xi)W_N^j(\xi) \tag{6.14}$$

and find the maximum value of $\widehat{LR}^j(\xi)$: $\widehat{\sup LR}^j := \max_{\xi\in\Xi_A} \widehat{LR}^j(\xi)$.

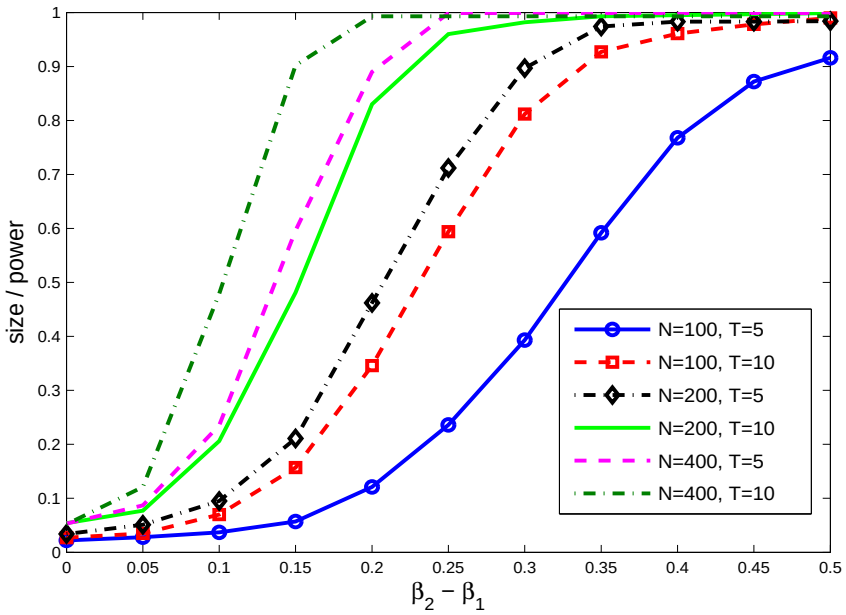

Figure 2: Size and Power of the Homogeneity Test

It can be shown that the empirical distribution of $\{\widehat{\sup LR}^j\}_{j=1}^J$ converges to the asymptotic distribution of $\sup_{\xi\in\Xi_A} LR(\xi)$, as $J\to\infty$. In this simulation study, we set $J=1000$. The 95% empirical quantile of $\{\widehat{\sup LR}^j\}_{j=1}^J$ is used as the critical value for the sup-LR test with nominal size 5%.

Figure 2 reports the percentage of rejections against the difference $\beta_2-\beta_1$ when $\sigma_\mu^2=1$. The figures for other values of σ_μ^2 are similar. When $\beta_2-\beta_1=0$, figure 2 presents the empirical size of the test. We find that for all the cases considered, the empirical size is smaller than the normal size. However, the difference between the empirical size and nominal size is smaller than 2% in the worst scenario. The size distortion becomes smaller as either N or T increases. When $N=200, T=10$, there is virtually no size distortion. Given that the standard deviation of the empirical size is 0.7% ($\sqrt{0.05\times 0.95/1000}$), we can conclude that the asymptotic distribution in Theorem 8 provides an excellent approximation to the finite sample distribution. When $\beta_2-\beta_1\neq 0$, figure 2 presents the power of the test. As expected, power is increasing in $\beta_2-\beta_1$, N and T. When the difference between β_2 and β_1 is larger than 0.2, N is larger than 200 (400) and T is larger than 10 (5), the power of the test is larger than 80%. The simulated power curve corroborates the asymptotic result given in Theorem 9. We can conclude that the homogeneity test has good size and power properties in finite samples.

7 Conclusion and Discussion

This paper has developed a framework for detecting group structure in panel data. The mechanism is via a panel structure model, which assumes that individuals form a number of homogeneous groups in a heterogeneous population. Within each group, individuals follow the same behavioral equations while these equations may be different across different groups. The

econometrician is not presumed to know the group structure. Instead, some covariates are used in a multinomial logistic regression to infer which individuals belong to which groups. Simulation experiments show that the asymptotic results are reflected in finite sample performance and that the proposed classification methodology is quite reliable.

The present study can be extended in several ways, and we briefly discuss some possibilities as follows:

First, we may relax the assumption that covariates in the logistic regression are time invariant. With this relaxation, individuals may switch memberships over time. A Markov Chain may be constructed to describe the membership dynamics. Knowledge of membership dynamics may deepen our understanding of the relationship between the dependent variable and the explanatory variables and may be useful in prediction.

Second, the model can be extended to incorporate nonstationary processes with consequent effects on the asymptotic theory and model interpretation. In panel cointegration modeling, for instance, cointegrating vectors are often assumed to be the same across all individuals. This may be restrictive. It may well be the case that the cointegrating vectors are the same only within groups of individuals. For example, in testing the PPP hypothesis, countries with large mutual trade flows are logically connected and we may expect those countries to form a natural group and share the same cointegrating vector. The PPP hypothesis may then hold within such country groupings but not for countries across groups.

Finally, the paper considers testing the null of one group against the alternative of multiple groups. The idea of using covariate variation in the logistic regression to facilitate the development of asymptotic theory can be extended to test G groups against $G+1$ groups for any $G \geq 1$. If no covariate is available in the logistic regression in empirical applications, we can artificially introduce one, but the resulting test may not be the most powerful one. In this case, we may use a locally conic parameterization of the panel structure model and employ the approach of Dacunha-Castelle and Gassiat (1999) to derive the asymptotic distribution of the sup-LR statistic.

Table 1: Bias and Standard Error of the Maximum Likelihood Estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ and Coverage Probability of 95% Confidence Interval for $N = 100$

	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
True Value	0.25	0.50	0.25	0.75	0.50	0.75	0.50	1.00
				$T = 5$				
$\sigma_\mu^2 = 0.5$								
Bias	−0.07	−0.03	−0.01	−0.04	−0.06	−0.03	−0.00	−0.00
Standard Error	0.15	0.14	0.11	0.16	0.15	0.13	0.11	0.09
Coverage	0.93	0.93	0.94	0.90	0.94	0.96	0.97	0.98
$\sigma_\mu^2 = 1.0$								
Bias	−0.04	0.01	0.00	−0.00	−0.03	−0.00	0.00	−0.00
Standard Error	0.15	0.13	0.12	0.13	0.14	0.12	0.12	0.07
Coverage	0.95	0.97	0.94	0.96	0.96	0.97	0.97	0.97
$\sigma_\mu^2 = 1.5$								
Bias	−0.03	0.02	0.00	0.00	−0.02	0.01	0.01	0.00
Standard Error	0.14	0.14	0.12	0.11	0.14	0.11	0.12	0.05
Coverage	0.94	0.97	0.95	0.97	0.96	0.98	0.97	0.97
				$T = 10$				
$\sigma_\mu^2 = 0.5$								
Bias	−0.01	0.01	−0.00	0.00	−0.01	0.00	−0.00	0.00
Standard Error	0.10	0.09	0.07	0.06	0.08	0.07	0.06	0.02
Coverage	0.90	0.93	0.94	0.98	0.91	0.96	0.97	0.98
$\sigma_\mu^2 = 1.0$								
Bias	−0.01	0.01	−0.00	0.00	−0.01	0.00	−0.00	0.00
Standard Error	0.09	0.08	0.07	0.05	0.08	0.06	0.06	0.02
Coverage	0.91	0.92	0.94	0.98	0.92	0.97	0.96	0.98
$\sigma_\mu^2 = 1.5$								
Bias	−0.01	0.01	−0.00	0.00	−0.01	0.00	−0.00	0.00
Standard Error	0.09	0.08	0.06	0.04	0.08	0.05	0.06	0.01
Coverage	0.91	0.93	0.94	0.98	0.93	0.97	0.96	0.98

Table 2: Bias and Standard Error of the Maximum Likelihood Estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ and Coverage Probability of 95% Confidence Interval for $N = 200$

	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
True Value	0.25	0.50	0.25	0.75	0.50	0.75	0.50	1.00
				$T = 5$				
$\sigma_\mu^2 = 0.5$								
Bias	−0.02	0.02	−0.00	0.01	−0.01	0.01	0.00	0.00
Standard Error	0.10	0.11	0.08	0.10	0.10	0.11	0.09	0.05
Coverage	0.94	0.96	0.95	0.95	0.96	0.96	0.97	0.96
$\sigma_\mu^2 = 1.0$								
Bias	−0.01	0.02	−0.00	0.01	−0.00	0.01	0.00	0.00
Standard Error	0.10	0.12	0.08	0.08	0.10	0.09	0.09	0.03
Coverage	0.94	0.96	0.96	0.96	0.95	0.96	0.97	0.96
$\sigma_\mu^2 = 1.5$								
Bias	−0.01	0.02	−0.00	0.00	0.00	0.01	0.00	0.00
Standard Error	0.09	0.11	0.08	0.06	0.09	0.08	0.09	0.03
Coverage	0.94	0.96	0.95	0.96	0.95	0.96	0.97	0.96
				$T = 10$				
$\sigma_\mu^2 = 0.5$								
Bias	−0.00	0.01	0.00	−0.00	0.00	0.00	0.00	−0.00
Standard Error	0.06	0.06	0.04	0.04	0.05	0.05	0.04	0.02
Coverage	0.91	0.91	0.95	0.97	0.93	0.96	0.96	0.97
$\sigma_\mu^2 = 1.0$								
Bias	−0.00	0.00	0.00	−0.00	0.00	−0.00	0.00	−0.00
Standard Error	0.06	0.08	0.04	0.04	0.05	0.06	0.04	0.01
Coverage	0.91	0.92	0.96	0.96	0.94	0.96	0.97	0.97
$\sigma_\mu^2 = 1.5$								
Bias	−0.00	0.00	0.00	−0.00	0.00	−0.00	0.00	−0.00
Standard Error	0.06	0.05	0.04	0.03	0.05	0.04	0.04	0.01
Coverage	0.91	0.92	0.96	0.96	0.94	0.96	0.97	0.96

Table 3: Average and Standard Deviation of the Percentage of Correct Classifications for $N = 100$

(β_1, β_2)		$(0.25, 0.50)$		$(0.25, 0.75)$		$(0.50, 0.75)$		$(0.50, 1.00)$	
		P^*	$\hat{P}$	P^*	$\hat{P}$	P^*	$\hat{P}$	P^*	$\hat{P}$
					$T = 5$				
$\sigma_\mu^2 = 0.5$	Mean	84.89	64.06	88.06	83.30	85.64	68.04	90.17	88.08
	S.D.	3.55	16.72	3.21	8.07	3.48	19.54	2.89	5.17
$\sigma_\mu^2 = 1.0$	Mean	85.15	65.52	88.74	85.09	86.25	72.21	91.40	89.83
	S.D.	3.52	17.12	3.14	6.71	3.42	18.79	2.78	5.15
$\sigma_\mu^2 = 1.5$	Mean	85.33	66.33	89.28	86.37	86.82	74.76	92.29	91.16
	S.D.	3.48	17.63	3.08	5.63	3.37	18.74	2.58	4.05
					$T = 10$				
$\sigma_\mu^2 = 0.5$	Mean	85.87	74.13	90.89	89.33	87.08	79.93	93.67	92.92
	S.D.	3.38	12.58	2.91	3.73	3.26	10.19	2.43	2.57
$\sigma_\mu^2 = 1.0$	Mean	86.14	74.99	91.48	90.14	87.83	82.56	94.74	94.09
	S.D.	3.37	12.76	2.84	3.44	3.23	8.49	2.26	2.43
$\sigma_\mu^2 = 1.5$	Mean	86.41	76.33	91.98	90.75	88.41	84.33	95.38	94.89
	S.D.	3.34	12.00	2.76	3.51	3.16	7.32	2.09	2.24

Table 4: Average and Standard Deviation of the Percentage of Correct Classifications for $N = 200$

	(β_1, β_2)	$(0.25, 0.50)$		$(0.25, 0.75)$		$(0.50, 0.75)$		$(0.50, 1.00)$	
		P^*	$\hat{P}$	P^*	$\hat{P}$	P^*	$\hat{P}$	P^*	$\hat{P}$
					$T = 5$				
$\sigma_\mu^2 = 0.5$	Mean	84.91	72.96	88.06	86.44	85.77	77.31	90.35	89.59
	S.D.	2.52	13.86	2.22	3.23	2.44	15.25	2.11	2.34
$\sigma_\mu^2 = 1.0$	Mean	85.11	74.29	88.76	87.41	86.41	80.15	91.58	91.07
	S.D.	2.50	13.90	2.17	2.91	2.41	13.78	2.00	2.51
$\sigma_\mu^2 = 1.5$	Mean	85.35	75.44	89.36	88.25	86.97	82.25	92.44	92.06
	S.D.	2.50	14.00	2.17	2.71	2.38	12.16	1.90	2.00
					$T = 10$				
$\sigma_\mu^2 = 0.5$	Mean	85.81	80.11	90.78	89.98	86.94	84.20	93.61	93.23
	S.D.	2.42	7.75	2.01	2.27	2.30	4.32	1.78	1.84
$\sigma_\mu^2 = 1.0$	Mean	86.07	81.15	91.37	90.70	87.68	85.61	94.66	94.35
	S.D.	2.37	7.06	1.94	2.17	2.23	3.64	1.65	1.68
$\sigma_\mu^2 = 1.5$	Mean	86.29	81.88	91.86	91.32	88.28	86.59	95.30	95.05
	S.D.	2.34	6.83	1.88	2.07	2.18	3.51	1.54	1.57

Note: $\hat{P}$ indicates the classification based on the estimated parameters while P^* indicates the classification based on the true parameters

8 Appendix

A.1 Proof of Lemma 1

Since $V(\theta_g^{(1)}) = V(\theta_g^{(2)})$ if and only if $(V(\theta_g^{(1)}))^{1/2} = (V(\theta_g^{(2)}))^{1/2}$. We first calculate $(V(\theta_g))^{1/2}$. Note that $\Sigma_g = \sigma_{\varepsilon,g}^2 (I_T - J_T/T) + \sigma_{o,g}^2 J_T/T$, J_T/T and $(I_T - J_T/T)$ are idempotent symmetric matrices, we have

$$\Sigma_g^{1/2} = \sigma_{\varepsilon,g} (I_T - J_T/T) + \sigma_{o,g} J_T/T = \sigma_{\varepsilon,g} I_T + (\sigma_{o,g} - \sigma_{\varepsilon,g}) J_T/T. \tag{A.1}$$

It follows that $(V(\theta_g))^{1/2} = M(\beta_g) [\sigma_{\varepsilon,g} I_T + (\sigma_{o,g} - \sigma_{\varepsilon,g}) J_T/T]$. Some elementary manipulations show that the equality $(V(\theta_g^{(1)}))^{1/2} = (V(\theta_g^{(2)}))^{1/2}$ implies that

$$\beta_g^{(1)} = \beta_g^{(2)}, \sigma_{\varepsilon,g}^{(1)} = \sigma_{\varepsilon,g}^{(2)}, \text{ and } \sigma_{o,g}^{(1)} = \sigma_{o,g}^{(2)}. \tag{A.2}$$

In view of equation (A.2) and the assumption $A(\theta_g^{(1)}, \eta^{(1)}) = A(\theta_g^{(2)}, \eta^{(2)})$ with probability one, we have

$$\begin{aligned} &M(\beta_g^{(1)}) x \gamma_g^{(1)} - M(\beta_g^{(2)}) x \gamma_g^{(2)} + M(\beta_g^{(1)}) z \eta^{(1)} - M(\beta_g^{(2)}) z \eta^{(2)} \\ &+ M(\beta_g^{(1)}) 1_T \rho_g^{(1)} y_0 - M(\beta_g^{(2)}) 1_T \rho_g^{(2)} y_0 \\ &= M(\beta_g^{(2)}) \tau \delta_g^{(2)} - M(\beta_g^{(1)}) \tau \delta_g^{(1)} + M(\beta_g^{(2)}) 1_T \alpha_g^{(2)} - M(\beta_g^{(1)}) 1_T \alpha_g^{(1)}, \end{aligned} \tag{A.3}$$

with probability one. But (A.2) implies $M(\beta_g^{(1)}) = M(\beta_g^{(2)})$. Note that $M(\beta_g^{(1)})$ and $M(\beta_g^{(2)})$ are invertible, we can write (A.3) as

$$\begin{aligned} &x \left(\gamma_g^{(1)} - \gamma_g^{(2)}\right) + z \left(\eta^{(1)} - \eta^{(2)}\right) + 1_T \left(\rho_g^{(1)} - \rho_g^{(2)}\right) y_0 \\ &= \tau \left(\delta_g^{(2)} - \delta_g^{(1)}\right) + 1_T \left(\alpha_g^{(2)} - \alpha_g^{(1)}\right). \end{aligned} \tag{A.4}$$

In particular,

$$\begin{aligned} &x'_{\cdot t_0} \left(\gamma_g^{(1)} - \gamma_g^{(2)}\right) + z'_{\cdot t_0} \left(\eta^{(1)} - \eta^{(2)}\right) + \left(\rho_g^{(1)} - \rho_g^{(2)}\right) y_0 \\ &= \tau_{t_0} \left(\delta_g^{(2)} - \delta_g^{(1)}\right) + \left(\alpha_g^{(2)} - \alpha_g^{(1)}\right). \end{aligned} \tag{A.5}$$

But (A.5) holds with probability one only if

$$\gamma_g^{(1)} = \gamma_g^{(2)}, \eta^{(1)} = \eta^{(2)} \text{ and } \rho_g^{(1)} = \rho_g^{(2)} \tag{A.6}$$

by conditions (i) and (ii). Combining (A.4) and (A.6) yields

$$\tau \left(\delta_g^{(2)} - \delta_g^{(1)}\right) + 1_T \left(\alpha_g^{(2)} - \alpha_g^{(1)}\right) = 0. \tag{A.7}$$

Since the matrix $(\tau, 1_T)$ has full column rank by condition (iii), (A.7) implies

$$\delta_g^{(1)} = \delta_g^{(2)} \text{ and } \alpha_g^{(1)} = \alpha_g^{(2)}. \tag{A.8}$$

The lemma is thus proved. □

A.2 Proof of Theorem 2

If $f(y, y_0; \psi^{(1)}) = f(y, y_0; \psi^{(2)})$ for almost all $(y', y_0', (v', w'))'$, then

$$\begin{aligned}&\sum_{g=1}^{G} \pi_g(\xi^{(1)}) m(y; \theta_g^{(1)}, \sigma_g^{(1)}, \eta^{(1)}) m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)}))\\ &= \sum_{g=1}^{G} \pi_g(\xi^{(2)}) m(y; \theta_g^{(2)}, \sigma_g^{(2)}, \eta^{(2)}) m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)})) \end{aligned} \tag{A.9}$$

for almost all $(y', y_0', (v', w'))'$. The equation can be put in the form

$$\sum_{g=1}^{G^*} \mathcal{K}(\xi_g^*, \varphi_g^*, \omega_g^*) m(y; \theta_g^*, \sigma_g^*, \eta^*) = 0, \tag{A.10}$$

where $G \leq G^* \leq 2G$, $\left(\theta_g^{*\prime}, \sigma_g^{*\prime}, \eta^{*\prime}\right)'$, $g = 1, 2, \ldots, G^*$ are mutually distinct vectors from the collection of $2G$ vectors $\{(\theta_g^{(1)\prime}, \sigma_g^{(1)\prime}, \eta^{(1)\prime})', (\theta_g^{(2)\prime}, \sigma_g^{(2)\prime}, \eta^{(2)\prime})'\}_{g=1}^{G}$. The function $\mathcal{K}(\xi_g^*, \varphi_g^*, \omega_g^*)$ is obtained from taking the difference of the coefficients of $m(y; \theta_g^*, \sigma_g^*, \eta^*)$ on the two sides of (A.9). Therefore $\mathcal{K}(\xi_g^*, \varphi_g^*, \omega_g^*)$ has the form $\pi_g(\xi^{(1)}) m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)})) - \pi_g(\xi^{(2)}) m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)}))$ if $m(y; \theta_g^*, \sigma_g^*, \eta^*)$ appears on both sides of (A.9), or has the form $\pm \pi_g(\xi) m^0(y_{i0}; \varphi_g, \omega_g))$ for $(\xi, \varphi_g, \omega_g) = (\xi^{(1)}, \varphi_g^{(1)}, \omega_g^{(1)})$ or $(\xi^{(2)}, \varphi_g^{(2)}, \omega_g^{(2)})$ if $m(y; \theta_g^*, \sigma_g^*, \eta^*)$ appears only on one side of (A.9). Since $\{(\theta_g^{*\prime}, \sigma_g^{*\prime}, \eta^{*\prime})'\}_{g=1}^{G^*}$ are mutually distinct, it follows from Lemma 1 that there exists a set $\mathcal{Y} \in \mathcal{F}_1$ with $\lambda_1(\mathcal{Y}) > 0$ such that the vectors $\left(A(\theta_g^*, \eta^*)', [Vec(V(\theta_g^*))]'\right)_{g=1}^{G^*}$ are mutually distinct for all $y \in \mathcal{Y}$. In view of the identification of finite normal mixtures, we deduce that $\{m(y; \theta_g^*, \sigma_g^*, \eta^*)\}_{g=1}^{G^*}$ are linearly independent functions of y on $\mathcal{Y} \subseteq \Omega_1$ (see Teicher (1963), Yakowitz and Spargins (1967)). That is to say, if $\sum_{g=1}^{G^*} a_g m(y; \theta_g^*, \sigma_g^*, \eta^*) = 0$ with positive probability, then $a_g = 0$, for $1 \leq g \leq G^*$. As a consequence, $\mathcal{K}(\xi_g^*, \varphi_g^*, \omega_g^*) = 0$, $g = 1, 2, ..., G^*$ for all (y', y_0', w') such that $y \in \mathcal{Y}$. Therefore, when $y \in \mathcal{Y}$, each distinct factor $m(y; \theta_g^*, \sigma_g^*, \eta^*)$ must appear on both sides of (A.9). Hence

$$\theta_g^{(1)} = \theta_g^{(2)}, \sigma_g^{(1)} = \sigma_g^{(2)} \text{ for all } g,\ \eta^{(1)} = \eta^{(2)},\ G^* = G. \tag{A.11}$$

Next, since $\mathcal{K}(\xi_g^*, \varphi_g^*, \omega_g^*) = 0$, $g = 1, 2, ..., G^*$ for (y', y_0', w') such that $y \in \mathcal{Y}$, we have

$$\pi_g(\xi^{(1)}) m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)})) - \pi_g(\xi^{(2)}) m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)})) = 0 \tag{A.12}$$

for (y', y_0', w') such that $y \in \mathcal{Y}$. Note that the condition $y \in \mathcal{Y}$ does not restrict the support of (y_0', w'), equation (A.12) thus holds for almost all y_0 and w.

We now consider two cases. First, if $(\varphi_g^{(1)}, \omega_g^{(1)}) = (\varphi_g^{(2)}, \omega_g^{(2)})$, then

$$\frac{\exp(w' \xi_g^{(1)})}{\sum_{g=1}^{G} \exp(w' \xi_g^{(1)})} = \frac{\exp(w' \xi_g^{(2)})}{\sum_{g=1}^{G} \exp(w' \xi_g^{(2)})},\ g = 1, 2, ..., G \tag{A.13}$$

for almost all w. In particular, when $g = G$, (A.13) becomes

$$\sum_{g=1}^{G} \exp(w' \xi_g^{(1)}) = \sum_{g=1}^{G} \exp(w' \xi_g^{(2)}) \tag{A.14}$$

for almost all w. This is because $\xi_G^{(1)}$ and $\xi_G^{(2)}$ are initialized parameters so that $\xi_G^{(1)} = \xi_G^{(2)} = 0$. Combining (A.13) and (A.14) leads to $w'\xi_g^{(1)} = w'\xi_g^{(2)}$ for almost all w. In view of condition (iv), this implies $\xi_g^{(1)} = \xi_g^{(2)}$.

Second, if $(\varphi_g^{(1)}, \omega_g^{(1)}) \neq (\varphi_g^{(2)}, \omega_g^{(2)})$, then conditional on w, $m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)}))$ and $m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)}))$ are linearly independent. As a result, we must have $\pi_g(\xi^{(1)}) = \pi_g(\xi^{(2)})$, which implies $\xi_g^{(1)} = \xi_g^{(2)}$. Given that $\xi_g^{(1)} = \xi_g^{(2)}$, we have $m^0(y_0; \varphi_g^{(1)}, \omega_g^{(1)})) = m^0(y_0; \varphi_g^{(2)}, \omega_g^{(2)}))$ for almost all y_0. This contradicts with the assumption that $(\varphi_g^{(1)}, \omega_g^{(1)}) \neq (\varphi_g^{(2)}, \omega_g^{(2)})$. Therefore we can only have $(\varphi_g^{(1)}, \omega_g^{(1)}) = (\varphi_g^{(2)}, \omega_g^{(2)})$. This completes the proof of the theorem. □

A.3 Proof of Theorem 3

Let $L \equiv L(\psi|y, y_0)$. We will prove that

$$\tilde{\xi}^{(k+1)} - \tilde{\xi}^{(k)} = -\tilde{H}_\xi^{-1}\left(\psi^{(k)}\right) \frac{\partial L}{\partial \tilde{\xi}}|_{\psi=\psi^{(k)}} \tag{A.15}$$

$$\begin{pmatrix} \theta^{(k+1)} - \theta^{(k)}) \\ \eta^{(k+1)} - \eta^{(k)}) \end{pmatrix} = -H_{\theta\eta}^{-1}(\psi^{(k)}) \begin{pmatrix} \frac{\partial L}{\partial \theta} \\ \frac{\partial L}{\partial \eta} \end{pmatrix}_{\psi=\psi^{(k)}} \tag{A.16}$$

$$\begin{pmatrix} \left(\sigma_{\varepsilon,g}^2\right)^{(k+1)} - \left(\sigma_{\varepsilon,g}^2\right)^{(k)} \\ \left(\sigma_{o,g}^2\right)^{(k+1)} - \left(\sigma_{o,g}^2\right)^{(k)} \end{pmatrix} = -H_\sigma^{-1}(\psi^{(k)}) \begin{pmatrix} \frac{\partial L}{\partial \sigma_{\varepsilon,g}^2} \\ \frac{\partial L}{\partial \sigma_{o,g}^2} \end{pmatrix}_{\psi=\psi^{(k)}} \tag{A.17}$$

$$\begin{pmatrix} \varphi_g^{(k+1)} - \varphi_g^{(k)} \\ \left(\omega_g^2\right)^{(k+1)} - \left(\omega_g^2\right)^{(k)} \end{pmatrix} = -H_{\varphi\omega}^{-1}(\psi^{(k)}) \begin{pmatrix} \frac{\partial L}{\partial \varphi_g} \\ \frac{\partial L}{\partial \omega_g^2} \end{pmatrix}_{\psi=\psi^{(k)}} \tag{A.18}$$

where $\tilde{H}_\xi(\psi^{(k)})$ is defined in (3.13),

$$H_{\theta\eta}(\psi^{(k)}) = -\begin{pmatrix} a_1(\psi^{(k)}) & 0 & ... & 0 & b_1(\psi^{(k)}) \\ 0 & a_2(\psi^{(k)}) & ... & 0 & b_2(\psi^{(k)}) \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & a_G(\psi^{(k)}) & b_G(\psi^{(k)}) \\ b_1'(\psi^{(k)}) & b_2'(\psi^{(k)}) & ... & b_G'(\psi^{(k)}) & c(\psi^{(k)}) \end{pmatrix} \tag{A.19}$$

$$H_\sigma(\psi^{(k)}) = -\begin{pmatrix} \frac{(T-1)\sum_{i=1}^N p_{ig}(\psi^{(k)})}{2\left(\sigma_{\varepsilon,g}^4\right)^{(k)}} & 0 \\ 0 & \frac{\sum_{i=1}^N p_{ig}(\psi^{(k)})}{2\left(\sigma_{og}^4\right)^{(k)}} \end{pmatrix} \tag{A.20}$$

$$H_{\varphi\omega}(\psi^{(k)}) = -\begin{pmatrix} \frac{\sum_{i=1}^N p_{ig}(\psi^{(k)})}{\left(\omega^{(k)}\right)^2} & 0 \\ 0 & \frac{\sum_{i=1}^N p_{ig}(\psi^{(k)})}{2\left(\omega_g^4\right)^{(k)}} \end{pmatrix} \tag{A.21}$$

and

$$a_g(\psi^{(k)}) = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} \Gamma_i \tag{A.22}$$

$$b_g(\psi^{(k)}) = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} z_i \tag{A.23}$$

$$c(\psi^{(k)}) = \sum_{i=1}^{N} \sum_{g=1}^{G} p_{ig}(\psi^{(k)}) z_i' \left(\Sigma_g^{(k)}\right)^{-1} z_i. \tag{A.24}$$

First, we calculate the gradient of the log-likelihood function. Some simple algebraic manipulations show that

$$\frac{\partial L}{\partial \xi_g} = \sum_{i=1}^{N} \left(p_{ig}(\psi) - \pi_{ig}(\xi)\right) w_i, \tag{A.25}$$

$$\frac{\partial L}{\partial \theta_g} = \sum_{i=1}^{N} p_{ig}(\psi) \Gamma_i' \Sigma_g^{-1} u_{i,g}, \tag{A.26}$$

$$\frac{\partial L}{\partial \eta} = \sum_{i=1}^{N} \sum_{g=1}^{G} p_{ig}(\psi) z_i' \Sigma_g^{-1} u_{i,g}, \tag{A.27}$$

$$\frac{\partial L}{\partial \sigma_{\varepsilon,g}^2} = -\frac{1}{2} \sum_{i=1}^{N} p_{ig}(\psi) \left\{ \frac{T-1}{\sigma_{\varepsilon,g}^2} - \frac{1}{\left(\sigma_{\varepsilon,g}^2\right)^2} \left(u_{i,g}\right)' \left(I_T - J_T/T\right) u_{i,g} \right\}, \tag{A.28}$$

$$\frac{\partial L}{\partial \sigma_{o,g}^2} = -\frac{1}{2} \sum_{i=1}^{N} p_{ig}(\psi) \left\{ \frac{1}{\sigma_{o,g}^2} - \frac{1}{\left(\sigma_{o,g}^2\right)^2} \left(u_{i,g}\right)' \left(\frac{J_T}{T}\right) u_{i,g} \right\}, \tag{A.29}$$

$$\frac{\partial L}{\partial \varphi_g} = \sum_{i=1}^{N} p_{ig}(\psi) \left(\frac{y_{i0} - \varphi_g}{\omega_g^2} \right), \tag{A.30}$$

and

$$\frac{\partial L}{\partial \omega_g^2} = -\frac{1}{2} \sum_{i=1}^{N} p_{ig}(\psi) \left(\frac{1}{\omega_g^2} - \frac{(y_{i0} - \varphi_g)^2}{\omega_g^4} \right). \tag{A.31}$$

Second, we prove each of the relationships in (A.15)–(A.18). (A.15) follows immediately from (3.13) and (A.25). To prove (A.16), we note that $\theta_g^{(k+1)}$ and $\eta^{(k+1)}$ satisfy

$$0 = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} (y_i - \Gamma_i' \theta_g^{(k+1)} - z_i' \eta^{(k+1)}). \tag{A.32}$$

In view of (A.26), we have

$$\frac{\partial L}{\partial \theta_g}|_{\psi=\psi^{(k)}} = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} (y_i - \Gamma_i' \theta_g^{(k)} - z_i' \eta^{(k)}). \tag{A.33}$$

It follows from (A.32) and (A.33) that

$$\frac{\partial L}{\partial \theta_g}|_{\psi=\psi^{(k)}} = \sum_{i=1}^{N} p_{ig}(\psi^{(k)})\Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} \left(\Gamma_i(\theta_g^{(k+1)} - \theta_g^{(k)}) + z_i(\eta^{(k+1)} - \eta^{(k)})\right). \tag{A.34}$$

Similarly, we can show that

$$\frac{\partial L}{\partial \eta}|_{\psi=\psi^{(k)}} = \sum_{i=1}^{N}\sum_{g=1}^{G} p_{ig}(\psi^{(k)})z_i' \left(\Sigma_g^{(k)}\right)^{-1} \left(\Gamma_i(\theta_g^{(k+1)} - \theta_g^{(k)}) + z_i(\eta^{(k+1)} - \eta^{(k)})\right). \tag{A.35}$$

Combining the above two equations yields (A.16).

To prove (A.17), we note that $\left(\sigma_{\varepsilon,g}^2\right)^{(k+1)}$ satisfies

$$0 = -\frac{1}{2}\sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \left\{(T-1)\left(\sigma_{\varepsilon,g}^2\right)^{(k+1)} - \left(u_{i,g}^{(k)}\right)' (I_T - J_T/T)\, u_{i,g}^{(k)}\right\}. \tag{A.36}$$

But

$$\left(\sigma_{\varepsilon,g}^4\right)^{(k)} \frac{\partial L}{\partial \sigma_{\varepsilon,g}^2}|_{\psi=\psi^{(k)}} = -\frac{1}{2}\sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \left\{(T-1)\left(\sigma_{\varepsilon,g}^2\right)^{(k)} - \left(u_{i,g}^{(k)}\right)' (I_T - J_T/T)\, u_{i,g}^{(k)}\right\}. \tag{A.37}$$

So

$$\left(\sigma_{\varepsilon,g}^4\right)^{(k)} \frac{\partial L}{\partial \sigma_{\varepsilon,g}^2}|_{\psi=\psi^{(k)}} = \frac{1}{2}\sum_{i=1}^{N} p_{ig}(\psi^{(k)})(T-1)\left\{\left(\sigma_{\varepsilon,g}^2\right)^{(k+1)} - \left(\sigma_{\varepsilon,g}^2\right)^{(k)}\right\}, \tag{A.38}$$

which implies that

$$\left(\sigma_{\varepsilon,g}^2\right)^{(k+1)} - \left(\sigma_{\varepsilon,g}^2\right)^{(k)} = \frac{2\left(\sigma_{\varepsilon,g}^4\right)^{(k)}}{(T-1)\sum_{i=1}^{N} p_{ig}(\psi^{(k)})} \frac{\partial L}{\partial \sigma_{\varepsilon,g}^2}|_{\psi=\psi^{(k)}}. \tag{A.39}$$

Similarly, we can show that

$$\left(\sigma_{o,g}^2\right)^{(k+1)} - \left(\sigma_{o,g}^2\right)^{(k)} = \frac{2\left(\sigma_{o,g}^4\right)^{(k)}}{\sum_{i=1}^{N} p_{ig}(\psi^{(k)})} \frac{\partial L}{\partial \sigma_{o,g}^2}|_{\psi=\psi^{(k)}}. \tag{A.40}$$

Combining (A.39) and (A.40) leads to the stated result.

To prove the first relationship in (A.18), we note that

$$\frac{\partial L}{\partial \varphi_g}|_{\psi=\psi^{(k)}} = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \left(\frac{y_{i0} - \varphi_g^{(k)}}{\left(\omega^{(k)}\right)^2}\right), \tag{A.41}$$

$$0 = \sum_{i=1}^{N} p_{ig}(\psi^{(k)}) \left(\frac{y_{i0} - \varphi_g^{(k+1)}}{\left(\omega^{(k)}\right)^2}\right). \tag{A.42}$$

We deduce immediately from the above two equations that

$$\varphi_g^{(k+1)} - \varphi_g^{(k)} = \left(\frac{\sum_{i=1}^{N} p_{ig}(\psi^{(k)})}{\left(\omega^{(k)}\right)^2}\right)^{-1} \frac{\partial L}{\partial \varphi_g}|_{\psi=\psi^{(k)}}, \tag{A.43}$$

as desired. Following the same step as the proof of (A.39), we can prove the second relationship in (A.18). Details are omitted.

Finally, we prove $H_\xi(\psi^{(k)}), H_{\theta\eta}(\psi^{(k)}), H_\sigma(\psi^{(k)})$ and $H_{\varphi\omega}(\psi^{(k)})$ are negative definite with probability one for large N. By inspection, it suffices to show that $H_{\theta\eta}(\psi^{(k)})$ is negative definite with probability one for large N. This is true because for any vector $(\theta', \eta')' = (\theta_1', ...\theta_G', \eta')'$ we have

$$(\theta', \eta')H_{\theta\eta}(\psi^{(k)})(\theta', \eta')' = -\sum_{g=1}^{G} \theta_g' a_g(\psi^{(k)})\theta_g - \eta' c\eta - \sum_{g=1}^{G} 2\theta_g' b_g(\psi^{(k)})\eta \leq 0. \tag{A.44}$$

The inequality holds because $\sum_{g=1}^{G} \theta_g' a_g(\psi^{(k)})\theta_g \geq 0$, $\eta' c\eta \geq 0$ and

$$\begin{aligned}
&\left|\sum_{g=1}^{G} 2\theta_g' b_g(\psi^{(k)})\eta\right| \leq \sum_{g=1}^{G}\sum_{i=1}^{N} p_{ig}(\psi^{(k)}) 2 \left|\theta_g'\Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} z_i\eta\right| \\
&\leq \sum_{g=1}^{G}\sum_{i=1}^{N} p_{ig}(\psi^{(k)})\theta_g'\Gamma_i' \left(\Sigma_g^{(k)}\right)^{-1} \Gamma_i\theta_g + \sum_{g=1}^{G}\sum_{i=1}^{N} p_{ig}(\psi^{(k)})\eta' z_i' \left(\Sigma_g^{(k)}\right)^{-1} z_i\eta \\
&= \sum_{g=1}^{G} \theta_g' a_g(\psi^{(k)})\theta_g + \eta' c\eta.
\end{aligned} \tag{A.45}$$

The above inequality also implies that $(\theta', \eta')H_{\theta\eta}(\psi^{(k)})(\theta', \eta')' = 0$ if and only if $\Gamma_i\theta_g = z_i\eta$ for all $i = 1, ..., N$. The latter holds with probability zero when the model is identified and N is large enough. Therefore, $H_{\theta\eta}(\psi^{(k)})$ is negative definite with probability one for large N.

This completes the proof of the theorem. □

A.4 Proof of Lemma 4

Part (a) of the Lemma is the same as Lemma 2.4 of Newey and McFadden (1994). Part (b) follows from standard textbook arguments. Details are omitted. □

A.5 Proof of Lemma 5

Proof of Part (a)

Since for all parameter values in K, $\sigma_{o,g}^2 \geq \sigma_{o,\min}^2 > 0$, $\sigma_{\varepsilon,g}^2 \geq \sigma_{\varepsilon,\min}^2 > 0$, $\omega_g^2 \geq \omega_{\min}^2 > 0$ for $g = 1, ..., G$, we have

$$\begin{aligned}
&\log\left(\sum_{g=1}^{G} \pi_{ig}(\xi) m(y_i; \theta_g, \sigma_g, \eta) m^0(y_{i0}; \varphi_g, \omega_g)\right) \\
&\leq \log\left(\sum_{g=1}^{G} \left(2\pi\sigma_{\varepsilon,g}^{2(T-1)}\sigma_{o,g}^2\right)^{-1/2} \left(2\pi\omega_g^2\right)^{-1/2}\right) \leq M,
\end{aligned} \tag{A.46}$$

for some constant $M > 0$. On the other hand,

$$\begin{aligned}
&\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)\right)\\
&\geq \log\left(\pi_{ig^*}(\xi)m(y_i;\theta_{g^*},\sigma_{g^*},\eta)m^0(y_{i0};\varphi_{g^*},\omega_{g^*})\right)
\end{aligned} \tag{A.47}$$

for any $1 \leq g^* \leq G$. Using (A.46) and (A.47), we have

$$\begin{aligned}
&\left|\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)\right)\right|\\
&\leq M + \left|\log\left(\pi_{ig^*}(\xi)\right)\right| + \left|\log m(y_i;\theta_{g^*},\sigma_{g^*},\eta)\right| + \left|\log m^0(y_{i0};\varphi_{g^*},\omega_{g^*})\right|.
\end{aligned} \tag{A.48}$$

Let $|A|$ signify the matrix consisting of the absolute values of the elements of A. Let B be a positive scalar such that $|\alpha_g| \leq B, |\beta_g| \leq B, |\gamma_g| \leq B1_{k_1}$, $|\eta| \leq B1_{k_2}$, $|\delta_g| \leq B1_{k_3}$ and $|\xi_g| \leq B1_{k_4}$ for all $1 \leq g \leq G$, where the inequalities hold element by element. We now bound each term in (A.48). Firstly,

$$\begin{aligned}
\left|\log\left(\pi_{ig^*}(\xi)\right)\right| &= \left|\log\left(\exp(w_i'\xi_{g^*})\right) - \log\left(\sum_{g=1}^{G}\exp(w_i'\xi_g)\right)\right|\\
&= \log\left(\sum_{g=1}^{G}\exp(w_i'\xi_g)\right) - w_i'\xi_{g^*} \leq \log\left(\sum_{g=1}^{G}\exp\left(w_i'\xi_g\right)\right) + \left|w_i'\xi_{g^*}\right|\\
&\leq \log\left(\sum_{g=1}^{G}\exp\left(\sum_{j=1}^{k_4}\left|w_{ij}\right|B\right)\right) + \sum_{j=1}^{k_4}\left|w_{ij}\right|B\\
&\leq \left(2B\sum_{j=1}^{k_4}\left|w_{ij}\right|\right) + \log G.
\end{aligned} \tag{A.49}$$

Secondly, since Σ_g^{-1} is positive definite and $|\Sigma_g|^{-1} = \sigma_{\varepsilon,g}^{-2(T-1)}\sigma_{o,g}^{-2} \leq \sigma_{\varepsilon,\min}^{-2(T-1)}\sigma_{o,\min}^{-2}$, the largest eigenvalue of Σ_g^{-1} is less than $\left(\sigma_{\varepsilon,\min}^{-2(T-1)/T}\sigma_{o,\min}^{-2/T}\right)$. Therefore

$$\begin{aligned}
&\left|\log m(y_i;\theta_{g^*},\sigma_{g^*},\eta)\right|\\
&\leq C + C(y_i^* - \Gamma_i^*\theta_{g^*} - z_i\eta)'\Sigma_g^{-1}(y_i^* - \Gamma_i^*\theta_{g^*} - z_i\eta)\\
&\leq C + C(y_i^* - \Gamma_i^*\theta_{g^*} - z_i\eta)'(y_i^* - \Gamma_i^*\theta_{g^*} - z_i\eta)'\\
&\leq C + C(y_i^{*\prime}y_i^* + \theta_{g^*}'\Gamma_i^{*\prime}\Gamma_i^*\theta_g + \eta' z_i z_i'\eta)\\
&\leq C + C(y_i^{*\prime}y_i^* + 1_{k_1+k_3+1}'\left|\Gamma_i^*\right|'\left|\Gamma_i^*\right|1_{k_1+k_3+1} + 1_{k_2}'\left|z_i\right|\left|z_i\right|'1_{k_2}),
\end{aligned} \tag{A.50}$$

where C is a generic constant. Similarly, $\left|\log m^0(y_{i0};\varphi_{g^*},\omega_{g^*})\right| \leq C + Cy_{i0}^2$. So

$$\sup_{\psi\in K}\left|\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)\right)\right| \tag{A.51}$$
$$\leq C\sum_{j=1}^{k_4}|w_{ij}| + C(y_i^{*\prime}y_i^* + 1'_{k_1+k_3+1}\left|\Gamma_i^*\right|'\left|\Gamma_i^*\right|1_{k_1+k_3+1} + 1'_{k_2}\left|z_i\right|\left|z_i\right|1_{k_2})1_{k_2} + Cy_{i0}^2 + C.$$

Under Assumption 4, the above upper bound has a finite expectation. We have thus proved $E\sup_{\psi\in K}|\log f(y_i,y_{i0};\psi)| < \infty$ as desired.

Proof of Part (b)

Note that $\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)m^0(y_{i0};\varphi_g,\omega_g)$ is the density of (y_i,y_{i0}) conditioning on (v_i,w_i). By the Kullback-Leibler information inequality, we have

$$E\left(\log f(y_i,y_{i0};\psi)\left|(v_i,w_i)\right.\right) < E\left(\log f(y_i,y_{i0};\psi_0)\left|(v_i,w_i)\right.\right) \tag{A.52}$$

for all $\psi\neq\psi_0$, $\psi\in K$. The inequality is strict because the model is identified on K. Taking expectations with respect to v_i and w_i on both sides of (A.52) yields

$$E\left(\log f(y_i,y_{i0};\psi)\right) < E\left(\log f(y_i,y_{i0};\psi_0)\right)$$

for all $\psi\neq\psi_0$, $\psi\in K$.

Proof of Part (c)

Part (c) holds because all the conditions in Lemma 4 are satisfied. □

A.6 Proof of Theorem 6

Let $\{\psi_{c,\ell}\}\in\Psi_c$ be a sequence of estimates satisfying $\lim_{\ell\to\infty}\sigma_{\varepsilon,g_0,\ell} = 0$ or ∞ for some g_0. Let $\{\psi^0_{c,\ell}\}\in\Psi_c$ be another sequence of estimates, which is obtained by changing the variance parameter $\sigma_{\varepsilon,g_0,\ell}$ of $\psi_{c,\ell}$ into c_0, for some small positive constant c_0. Since $T > k_1+k_2+k_3+3$, $u'_{i,g}(I_T-J_T/T)u_{i,g} > 0$ with probability one for all i and g. Therefore, with probability one, $\lim_{\ell\to\infty}m(y_i;\theta_{g_0,\ell},\sigma_{o,g_0,\ell},\sigma_{\varepsilon,g_0,\ell},\eta_\ell) = 0$ and $\lim_{\ell\to\infty}m(y_i;\theta_{g_0,\ell},\sigma_{o,g_0,\ell},c_0,\eta_\ell) > 0$, which implies $\lim_{\ell\to\infty}\sum_{i=1}^{N}\log f_c(y_i;\psi_{c,\ell}) < \lim_{\ell\to\infty}\sum_{i=1}^{N}\log f_c(y_i;\psi^0_{c,\ell})$. It follows that $\sum_{i=1}^{N}\log f_c(y_i;\psi_c)$, as a function of $\sigma^2_{\varepsilon,g_0,\ell}$, achieves its maximum value at some interior point of the interval $(0,\infty)$. Invoking the first order condition, we can show that the maximizer of $\sum_{i=1}^{N}\log f_c(y_i,;\psi_c)$, as a function of $\sigma^2_{\varepsilon,g_0}$, is given by

$$\sigma^2_{\varepsilon,g_0} = \frac{\sum_{i=1}^{N}p_{ig_0}u'_{i,g_0}(I_T - J_T/T)u_{i,g_0}}{(T-1)\sum_{i=1}^{N}p_{ig_0}}, \tag{A.53}$$

where $u_{i,g_0} = y_i - \Gamma_i\theta_{g_0} - z_i\eta$ and

$$p_{ig_0} = \frac{\pi_{ig_0}(\xi)m(y_i;\theta_{g_0},\sigma_{g_0},\eta)}{\sum_{j=1}^{G}\pi_{ij}(\xi)m(y_i;\theta_j,\sigma_j,\eta)}. \tag{A.54}$$

Let

$$\left(\sigma_{\varepsilon}^{i}\right)^{2}=\min_{\theta,\eta}\left(y_{i}-\Gamma_{i}\theta-z_{i}\eta\right)^{\prime}\left(I_{T}-J_{T}/T\right)\left(y_{i}-\Gamma_{i}\theta-z_{i}\eta\right), \tag{A.55}$$

which is independent of any parameter. Then

$$\sigma_{\varepsilon,g_0}^{2}=\frac{\sum_{i=1}^{N}p_{ig_0}u_{i,g_0}^{\prime}(I_{T}-J_{T}/T)u_{i,g_0}}{(T-1)\sum_{i=1}^{N}p_{ig_0}}\geq\frac{\sum_{i=1}^{N}p_{ig_0}\left(\sigma_{\varepsilon}^{i}\right)^{2}}{(T-1)\sum_{i=1}^{N}p_{ig_0}}. \tag{A.56}$$

Similarly, we can show that

$$\sigma_{o,g_0}^{2}=\frac{\sum_{i=1}^{N}p_{ig_0}u_{i,g_0}^{\prime}(J_{T}/T)u_{i,g_0}}{\sum_{i=1}^{N}p_{ig_0}}\geq\frac{\sum_{i=1}^{N}p_{ig_0}\left(\bar{\sigma}_{\varepsilon}^{i}\right)^{2}}{\sum_{i=1}^{N}p_{ig_0}}, \tag{A.57}$$

where

$$\left(\bar{\sigma}_{\varepsilon}^{i}\right)^{2}=\min_{\theta,\eta}\left(y_{i}-\Gamma_{i}\theta-z_{i}\eta\right)^{\prime}\left(J_{T}/T\right)\left(y_{i}-\Gamma_{i}\theta-z_{i}\eta\right). \tag{A.58}$$

Therefore

$$\begin{aligned}\left|\sigma_{\varepsilon,g_0}\sigma_{o,g_0}\right|&\geq\left(\frac{\sum_{i=1}^{N}p_{ig_0}\left(\sigma_{\varepsilon}^{i}\right)^{2}}{(T-1)\sum_{i=1}^{N}p_{ig_0}}\right)^{1/2}\left(\frac{\sum_{i=1}^{N}p_{ig_0}\left(\bar{\sigma}_{\varepsilon}^{i}\right)^{2}}{\sum_{i=1}^{N}p_{ig_0}}\right)^{1/2}\\&\geq\frac{1}{\sqrt{T-1}}\frac{\sum_{i=1}^{N}p_{ig_0}\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}}{\sum_{i=1}^{N}p_{ig_0}}\geq\frac{\sum_{i=1}^{N}p_{ig_0}\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|}{\sqrt{T-1}N},\end{aligned} \tag{A.59}$$

which implies that

$$\sum_{g_0=1}^{G}\left|\sigma_{\varepsilon,g_0}\sigma_{o,g_0}\right|\geq\frac{\sum_{i=1}^{N}\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|}{\sqrt{T-1}N}. \tag{A.60}$$

Since $\{(T-1)^{-1}\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|\}_{i=1}^{N}$ are iid, and $E\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|<\infty$, we have, by a strong law of large numbers,

$$\lim_{N\to\infty}\frac{\sum_{i=1}^{N}p_{ig_0}\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|}{\sqrt{T-1}N}=2r, \tag{A.61}$$

with probability one for some positive constant r. This is because when $T>k_1+k_2+k_3+3$, $\left|\sigma_{\varepsilon}^{i}\bar{\sigma}_{\varepsilon}^{i}\right|>0$ with probability one for all i. Therefore, with probability one, we have $\sum_{g_0=1}^{G}\left|\sigma_{\varepsilon,g_0}\sigma_{o,g_0}\right|\geq r$ when N is large enough. As a consequence,

$$\hat{\psi}_{c}=\arg\max_{\psi_c\in\Psi_c}\sum_{i=1}^{N}\log f_{c}(y_{i};\psi_{c})=\arg\max_{\psi_c\in\Psi_c^{0}}\sum_{i=1}^{N}\log f_{c}(y_{i};\psi_{c}), \tag{A.62}$$

with probability one when N is large enough, where

$$\Psi_{c}^{0}=\left\{\psi_{c}:\psi_{c}\in\Psi_{c},\sum_{g_0=1}^{G}\left|\sigma_{\varepsilon,g_0}\sigma_{o,g_0}\right|\geq r\right\}. \tag{A.63}$$

Let $q(U_i,\psi_c)=\log f_c(y_i;\psi_c)$, $K=\Psi_c^0$, and define $m(y_i;\theta_g,\sigma_g,\eta)=0$ if $\sigma_{\varepsilon,g}^2=0$ or ∞, or $\sigma_{o,g}^2=0$ or ∞. We now verify the conditions in Lemma 4. Condition (i) in Lemma 4 is obviously satisfied. Since $\sum_{g_0=1}^{G}\left|\sigma_{\varepsilon,g_0}\sigma_{o,g_0}\right|\geq r$, $f_c(y_i;\psi_c)>0$ with probability one. It follows

that $\log f_c(y_i;\psi_c)$ is continuous with probability one. Hence condition (ii) holds. Condition (iv) can be verified in the same way as in the proof of Lemma 5.

It remains to verify condition (iii). First, since

$$\exp(-x^2/2) \leq C\left(x^2\right)^{-T/2}, T \geq 2, \tag{A.64}$$

for some positive constant C, we have

$$\begin{aligned}
&m(y_i;\theta_g,\sigma_g,\eta)\\
&= (2\pi)^{-T/2}\,\sigma_{\varepsilon,g}^{-(T-1)}\exp\left(-\frac{u_{i,g}'(I_T-J_T/T)u_{i,g}}{2\sigma_{\varepsilon,g}^2}\right)\sigma_{o,g}^{-1}\exp\left(-\frac{u_{i,g}'(J_T/T)u_{i,g}}{2\sigma_{o,g}^2}\right)\\
&\leq (2\pi)^{-T/2}\,\sigma_{\varepsilon,g}^{-(T-1)}C\left(\frac{u_{i,g}'(I_T-J_T/T)u_{i,g}}{\sigma_{\varepsilon,g}^2}\right)^{-T/2}\sigma_{o,g}^{-1}\left(\frac{u_{i,g}'(J_T/T)u_{i,g}}{\sigma_{o,g}^2}\right)^{-1}\\
&\leq C\sigma_{o,g}\sigma_{\varepsilon,g}\left(u_{i,g}'(I_T-J_T/T)u_{i,g}\right)^{-T/2}\left(u_{i,g}'(J_T/T)u_{i,g}\right)^{-1}.
\end{aligned} \tag{A.65}$$

Therefore, with probability one,

$$\begin{aligned}
&\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)\right)\\
&\leq C+\log\sum_{g=1}^{G}\left(u_{i,g}'(I_T-J_T/T)u_{i,g}\right)^{-T/2}\left(u_{i,g}'(J_T/T)u_{i,g}\right)^{-1}\leq M,
\end{aligned} \tag{A.66}$$

for some constant $M>0$. This is because $\left[u_{i,g}'(I_T-J_T/T)u_{i,g}\right]\times\left[u_{i,g}'(J_T/T)u_{i,g}\right]>C$ for some small positive C with probability one.

Next,

$$\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)\right)\geq\log\left(\pi_{ig^*}(\xi)m(y_i;\theta_{g^*},\sigma_{g^*},\eta)\right) \tag{A.67}$$

for some g^* such that $\sigma_{\varepsilon,g^*}\sigma_{o,g^*}>0$. Since $\sum_{g=1}^{G}|\sigma_{\varepsilon,g}\sigma_{o,g}|\geq r>0$ on Ψ_c^0, there always exists such a g^*.

Using (A.66) and (A.67), we have

$$\left|\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)\right)\right|\leq M+\left|\log\left(\pi_{ig^*}(\xi)\right)\right|+\left|\log m(y_i;\theta_{g^*},\sigma_{g^*},\eta)\right|. \tag{A.68}$$

Since $\sigma_{\varepsilon,g^*}\sigma_{o,g^*}>0$, the largest eigenvalue of $\Sigma_{g^*}^{-1}$ is bounded above. Hence (A.50) holds. It then follows from the proof similar to that of Lemma 5 that

$$E\sup_{\psi\in K}\left|\log\left(\sum_{g=1}^{G}\pi_{ig}(\xi)m(y_i;\theta_g,\sigma_g,\eta)\right)\right|<\infty. \tag{A.69}$$

Since all the conditions in Lemma 4 are verified, $\hat{\psi}_c=\arg\max_{\psi_c\in\Psi_c}\sum_{i=1}^{N}\log f(y_i;\psi_c)$ is consistent. □

A.7 Proof of Theorem 7

Proof of Part (a)

For notational convenience, we write $\pi_{ij} = \pi_{ij}(\xi)$, $m_{ij} = m(y_i; \theta_j, \sigma_j, \eta)$ and $m^0_{ij} = m^0(y_{i0}; \varphi_j, \omega_j)$. Since the proofs are similar for different r's, we consider $\partial f(y_i, y_{i0}; \psi)/\partial \beta_j = \pi_{ij} m_{ij} m^0_{ij} u'_{i,j} \Sigma_j^{-1} y_{i,-1}$ as an example. The uniform integrability of other partial derivatives follows from the same argument.

Using the fact that Ψ_0 is bounded, let $\{B(\psi(\ell), \epsilon) : \ell = 1, ..., \mathcal{L}\}$ be a finite cover of Ψ_0. Then

$$
\begin{aligned}
&\int \sup_{\psi} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi) \right| d\lambda_0 d\lambda_1 d\lambda_2 \\
&\leq \int \sup_{\psi(\ell)} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 \\
&+ \int \sup_{\psi(\ell)} \sup_{\psi \in B(\psi(\ell), \epsilon)} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi) - \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 \\
&\leq \int \sup_{\psi(\ell)} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 \\
&+ \int \sup_{\psi(\ell)} \sup_{\psi \in B(\psi(\ell), \epsilon)} \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi) \left[\frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right]^{-1} - 1 \right| d\lambda_0 d\lambda_1 d\lambda_2 \qquad \text{(A.70)} \\
&\leq \int \sup_{\psi(\ell)} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 + C \int \sup_{\psi(\ell)} \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 \\
&\leq C \sum_{\ell=1}^{\mathcal{L}} \int \left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right| d\lambda_0 d\lambda_1 d\lambda_2 \leq C,
\end{aligned}
$$

where the third inequality follows because $\frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi) \left[\frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi(\ell)) \right]^{-1}$ is uniformly continuous on Ψ_0 and hence

$$
\left| \frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi) \left[\frac{\partial}{\partial \beta_j} f(y_i, y_{i0}; \psi') \right]^{-1} - 1 \right| \leq C \text{ for all } \psi \text{ and } \psi' \text{ such that } \left\| \psi - \psi' \right\| \leq \epsilon.
$$

Proof of Part (b)

Note that the second order derivatives with respect to regression parameters specific to different groups are always equal to zero, Part (b) automatically holds for these derivatives. For other second order derivatives, we first write down their explicit expressions and then use the same proof in part (a) to show that these second order derivatives are uniformly integrable. Details are omitted.

Proof of Part (c) Note that

$$\begin{aligned}
&\sup_{\psi\in\Psi_0}\left|\frac{\partial^2}{\partial\psi_r\partial\psi_s}\log f(y_i,y_{i0};\psi)\right| \\
&=\sup_{\psi\in\Psi_0}\left|\frac{\partial^2 f(y_i,y_{i0};\psi)}{\partial\psi_r\partial\psi_s}f-\frac{\partial f(y_i,y_{i0};\psi)}{\partial\psi_r}\frac{\partial f(y_i,y_{i0};\psi)}{\partial\psi_s}\right|\left|f^{-2}(y_i,y_{i0},\psi)\right| \\
&\leq\sup_{\psi\in\Psi_0}\left|\frac{\partial^2 f(y_i,y_{i0};\psi)}{\partial\psi_r\partial\psi_s}\frac{1}{f}\right|+\sup_{\psi\in\Psi_0}\left|\frac{\partial f(y_i,y_{i0};\psi)}{\partial\psi_r}\frac{\partial f(y_i,y_{i0};\psi)}{\partial\psi_s}\frac{1}{f^2}\right|, \qquad \text{(A.71)}
\end{aligned}$$

some manipulations show that $\sup_{\psi\in\Psi_0}\left|\frac{\partial^2}{\partial\psi_r\partial\psi_s}\log f(y_i,y_{i0};\psi)\right|$ is bounded by a finite order polynomial of $(|y_{it}|,|y_{it-1}|,|x_{it}|,|z_{it}|)$. The highest order is 4. Since all the random variables are assumed to have finite fourth moment, we have

$$E\sup_{\psi\in\Psi_0}\left|\frac{\partial^2}{\partial\psi_r\partial\psi_s}\log f(y_i,y_{i0};\psi)\right|<\infty.$$

Proof of Part (d)

Differentiating the identity $\int f(y_i,y_{i0},\psi_0)d\lambda_0 d\lambda_1 d\lambda_2 = 1$ twice, and interchanging the orders of differentiation and integration, as allowed by Parts (a) and (b), we have

$$Es_i(\psi_0)=0,\ Es_i(\psi_0)s_i'(\psi_0)=-Eh_i(\psi_0). \qquad \text{(A.72)}$$

From Part (c), we have

$$Es_i(\psi_0)s_i'(\psi_0)=-Eh_i(\psi_0)\leq E\sup_{\psi\in\Psi_0}\left|\frac{\partial^2}{\partial\psi\partial\psi'}\log f(y_i,y_{i0};\psi)\right|<\infty, \qquad \text{(A.73)}$$

where the inequality holds element by element. It then follows from the Lindberg-Levy central limit theorem that $\sqrt{N}S(\psi_0)\Rightarrow N(0,\mathcal{I})$.

Proof of Part (e)

We first show that, there exist a $\epsilon^*>0$ such that $H(\psi)\to E\frac{\partial^2}{\partial\psi\partial\psi'}\log f(y_i,y_{i0};\psi)$ uniformly for all $\psi\in B(\psi_0,\epsilon^*)\cap\Psi_0$. To prove this, we apply Lemma 4 with $q(U_i,\psi)=h_i(\psi)$. It is easy to see that Conditions (i) and (ii) in Lemma 4 are satisfied. Condition (iii) in Lemma 4 holds because of Part (c).

Taking a Taylor series expansion of $S(\hat{\psi})$, we have

$$S(\hat{\psi})=S(\psi_0)+H(\psi^*)(\hat{\psi}-\psi_0), \qquad \text{(A.74)}$$

where ψ^* is between ψ and $\hat{\psi}$. Since $\hat{\psi}$ is consistent, $\psi^*\in B(\psi_0,\epsilon^*)$, with probability approaching one. Combine this with the uniform convergence of $H(\psi)$ over $\psi\in B(\psi_0,\epsilon^*)$, we have $H(\psi^*)=H(\psi_0)+o_p(1)$ uniformly.

So

$$\sqrt{N}(\hat{\psi}-\psi_0)=-H^{-1}(\psi^*)\sqrt{N}S(\psi_0)=\mathcal{I}^{-1}\sqrt{N}S(\psi_0)+o_p(1)\Rightarrow N(0,\mathcal{I}^{-1}). \qquad \text{(A.75)}$$

□

A.8 Proof of Theorem 8

We start by showing that the quadratic approximations (5.23) and (5.26) hold uniformly over $\xi \in \Xi$. The uniform approximations hold if (i) $\sup_{\xi\in\Xi} \left\| \hat{\vartheta}_{ur}(\xi) - \vartheta_0 \right\| = o_p(1)$ under the null hypothesis; (ii) $\hat{\vartheta}_r - \vartheta_0 = o_p(1)$; (iii) $H_N(\xi,\vartheta) = N^{-1}\partial^2 L_{ur}(\xi,\vartheta)/\partial\vartheta\partial\vartheta'$ converges uniformly over $\xi \in \Xi$ and $\vartheta \in \Theta_0$ where Θ_0 is some neighborhood of ϑ_0; (iv) $\mathcal{I}_{ur}(\xi,\vartheta)$ is uniformly continuous in (ξ,ϑ) over $\Xi \times \Theta_0$ and (v) $\mathcal{I}_{ur}(\xi,\vartheta_0)$ is uniformly positive definite, i.e. $\inf_{\xi\in\Xi} \lambda_{\min}(\mathcal{I}_{ur}(\xi,\vartheta_0)) > 0$. Condition (i) can be verified by establishing the uniform convergence of the log-likelihood function to the limit function and the uniform identifiability condition on the limit function. The technical arguments are similar to those in the proofs of Lemma 5 and Theorem 7. Conditions (ii) and (iii) can be verified using Lemmas 4 and 5. Condition (iv) follows from condition (iii), the continuity of $H_N(\xi,\vartheta)$ and the compactness of $\Xi \times \Theta_0$. We now prove condition (v) by contradiction. Note that

$$\mathcal{I}_{ur}(\xi,\vartheta_0) = E\left(\begin{array}{c} \pi(\xi) \otimes \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\zeta} \\ \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\eta} \end{array} \right)\left(\begin{array}{c} \pi(\xi) \otimes \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\zeta} \\ \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\eta} \end{array} \right)'. \tag{A.76}$$

Assume that there is a $\xi^* \in \Xi$ such that $\mathcal{I}_{ur}(\xi^*,\vartheta_0)$ is singular. As a consequence, there exists a nonzero vector $c = (a',b')'$ where $a = (a_1',a_2',...,a_G')'$ such that $c'\mathcal{I}_{ur}(\xi^*,\vartheta_0)c = 0$. But the latter holds if and only if

$$\sum_{g=1}^{G} \pi_{\cdot,g}(\xi^*) a_g' \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\zeta} + b' \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\eta} = 0 \tag{A.77}$$

with probability one. Given that the support of $\tilde{\pi} = (\pi_{\cdot,1},...,\pi_{\cdot,G-1})'$ spans $\mathbb{R}^{G-1}$ and that $\sum_{g=1}^{G} \pi_{\cdot,g}(\xi^*) = 1$, equation (A.77) holds only in the following two cases. In the first case,

$$a_g' \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\zeta} = 0 \text{ for all } g, \tag{A.78}$$

and

$$b' \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\eta} = 0. \tag{A.79}$$

In the second case, $a_1 = a_2 = ... = a_G = a_0$ for some a_0, and

$$a_0' \frac{\partial \log m(y,y_0;\zeta_0,\eta_0)}{\partial\zeta} + b' \frac{\partial \log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta} = 0. \tag{A.80}$$

Note that the identifiability of the model implies that $(\partial \log m(y,y_0;\zeta_0,\eta_0)/\partial\zeta', \partial \log m(y,y_0;\zeta_0,\eta_0)/\partial\eta')$ are linearly independent functions. Therefore, in both cases, we have $a_g = 0$ for all g and $b = 0$, and hence that $c = 0$. This contradicts with the assumption that $c \neq 0$. We have thus proved that $\inf_{\xi\in\Xi} \lambda_{\min}(\mathcal{I}_{ur}(\xi,\vartheta_0)) > 0$.

The uniform quadratic approximations (5.23) and (5.26) imply that

$$2(L_{ur}(\xi,\hat{\vartheta}_{ur}) - L_r(\xi,\hat{\vartheta}_r)) = N S_N'(\xi)\, Q(\xi) S_N(\xi) + o_p(1), \tag{A.81}$$

uniformly over $\xi \in \Xi$.

Next, we show that $\sqrt{N} S_N(\xi)$ converges in distribution to a Gaussian process with mean zero and covariance kernel $\mathcal{C}(\xi_1,\xi_2)$. Note that for each $\xi \in \Xi$, $s_i(\xi,\vartheta_0)$ is an iid process with

finite second moment. Using the pointwise central limit theorem, we can show that the finite dimensional distributional convergence holds. It remains to show that $S_N(\xi)$ is stochastically equicontinuous. Given that Ξ is compact, $s_i(\xi,\vartheta_0)$ is continuous in ξ on Ξ and $s_i(\xi,\vartheta_0)$ is an iid process, it suffices to prove $E\sup_{\xi\in\Xi}||s_i(\xi,\vartheta_0)||<\infty$. By definition,

$$s_i(\xi,\vartheta_0)=\begin{pmatrix}\pi_i(\xi)\otimes\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\zeta}\\ \frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta}\end{pmatrix}, \tag{A.82}$$

so

$$\begin{aligned}E\sup_{\xi\in\Xi}||s_i(\xi,\vartheta_0)||&=E\sup_{\xi\in\Xi}\left|\pi_i'(\xi)\pi_i(\xi)\left(\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\zeta}\right)'\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\zeta}\right.\\&\quad\left.+\left(\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta}\right)'\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta}\right|\\&\le E\left|\left(\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\zeta}\right)'\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\zeta}\right|\\&\quad+E\left|\left(\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta}\right)'\frac{\partial\log m(y_i,y_{i0};\zeta_0,\eta_0)}{\partial\eta}\right|\\&<\infty\end{aligned} \tag{A.83}$$

as required. Hence $\sqrt{N}S_N(\xi)\Rightarrow W(\xi)$.

Invoking the continuous mapping theorem completes the proofs of (5.35) and (5.36).□

A.9 Proof of Theorem 9

For any given $\xi\in\Xi$, the probability limit of the unrestricted log-likelihood function $L_{ur}(\xi,\vartheta)$ under the sequence of local alternatives is the same as that under the fixed alternative ϑ_0. This is also true for the restricted log-likelihood function $L_r(\xi,\vartheta_r)$. Hence, $\sup_{\xi\in\Xi}\left\|\hat{\vartheta}_{ur}(\xi)-\vartheta_0\right\|=o_p(1)$ and $\hat{\vartheta}_r-\vartheta_0=o_p(1)$. It is not hard to show that under the local alternatives $H_N(\xi,\vartheta)=N^{-1}\partial^2L_{ur}(\xi,\vartheta)/\partial\vartheta\partial\vartheta'$ converges to $-\mathcal{I}_{ur}(\xi,\vartheta)$ uniformly over $\xi\in\Xi$ and $\vartheta\in\Theta_0$. Let

$$s_i(\xi,\vartheta)=\begin{pmatrix}\pi_{i1}(\xi)\frac{\partial m(y_i,y_{i0};\zeta_1,\eta)}{\partial\zeta_1}\left(\sum_{g=1}^G\pi_{ig}(\xi)m(y_i,y_{i0};\zeta_g,\eta)\right)^{-1}\\ \cdots\\ \pi_{iG}(\xi)\frac{\partial m(y_i,y_{i0};\zeta_G,\eta)}{\partial\zeta_G}\left(\sum_{g=1}^G\pi_{ig}(\xi)m(y_i,y_{i0};\zeta_g,\eta)\right)^{-1}\\ \sum_{g=1}^G\pi_{ig}(\xi)\frac{\partial\log m(y_i,y_{i0};\zeta_g,\eta)}{\partial\eta}\left(\sum_{g=1}^G\pi_{ig}(\xi)m(y_i,y_{i0};\zeta_g,\eta)\right)^{-1}\end{pmatrix} \tag{A.84}$$

then we have

$$2(L_{ur}(\xi,\hat{\vartheta}_{ur})-L_r(\xi,\hat{\vartheta}_r))=\left[\sqrt{N}S_N'(\xi)\right]Q(\xi)\left[\sqrt{N}S_N(\xi)\right]+o_p(1), \tag{A.85}$$

uniformly over $\xi\in\Xi$, where

$$\begin{aligned}\sqrt{N}S_N(\xi)&=\frac{1}{\sqrt{N}}\frac{\partial L_{ur}(\xi,\vartheta_0)}{\partial\vartheta}=\frac{1}{\sqrt{N}}\sum_{i=1}^N s_i(\xi,\vartheta_0)\\&=\frac{1}{\sqrt{N}}\sum_{i=1}^N\left[s_i(\xi,\vartheta_0)-Es_i(\xi,\vartheta_0)\right]+\sqrt{N}Es_i(\xi,\vartheta_0).\end{aligned} \tag{A.86}$$

It can be shown that $1/\sqrt{N}\sum_{i=1}^{N}\left[s_i(\xi,\vartheta_0)-Es_i(\xi,\vartheta_0)\right]$ converges in distribution to a Gaussian process $\bar{W}(\xi)$ with mean zero and covariance kernel $\bar{\mathcal{C}}(\xi_1,\xi_2)=cov(s_i(\xi_1,\vartheta_0),s_i(\xi_2,\vartheta_0))$. It remains to find the limit of $\sqrt{N}Es_i(\xi,\vartheta_0)$. To this end, we first obtain the limit of $\sqrt{N}Es_i(\xi_0,\vartheta_0)$. Note that

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}s_i(\xi_0,\vartheta_0)&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}s_i(\xi_0,\vartheta_N)+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial s_i(\xi_0,\vartheta_0)}{\partial\vartheta}\sqrt{N}(\vartheta_N-\vartheta_0)+o_p(1)\\&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}s_i(\xi_0,\vartheta_N)-\mathcal{I}_{ur}(\xi_0,\vartheta_0)d+o_p\left(1\right)\\&\to_d N\left(-\mathcal{I}_{ur}(\xi_0,\vartheta_0)d,\ \mathcal{I}_{ur}(\xi_0,\vartheta_0)\right),\end{aligned}\tag{A.87}$$

which implies that $\lim_{N\to\infty}\sqrt{N}Es_i(\xi_0,\vartheta_0)=-\mathcal{I}_{ur}(\xi_0,\vartheta_0)d$. Let $F(\xi,\xi_0)$ be a diagonal matrix such that

$$Es_i(\xi,\vartheta_0)=F(\xi,\xi_0)Es_i(\xi_0,\vartheta_0).$$

Since $d\in D_A$, $F(\xi,\xi_0)$ is always well defined. Now

$$\lim_{N\to\infty}\sqrt{N}Es_i(\xi,\vartheta_0)=-F\left(\xi,\xi_0\right)\mathcal{I}_{ur}(\xi_0,\vartheta_0)d.\tag{A.88}$$

As a consequence,

$$\begin{aligned}&2(L_{ur}(\xi,\hat{\vartheta}_{ur})-L_r(\xi,\hat{\vartheta}_r))\\&\Rightarrow\left[\bar{W}(\xi)-F\left(\xi,\xi_0\right)\mathcal{I}_{ur}(\xi_0,\vartheta_0)d\right]'Q(\xi)\left[\bar{W}(\xi)-F\left(\xi,\xi_0\right)\mathcal{I}_{ur}(\xi_0,\vartheta_0)d\right]',\end{aligned}\tag{A.89}$$

from which the theorem follows.

References

[1] Akerlof, G. (1997): "Social Distance and Social Decisions," *Econometrica*, 65, 1005–1028.

[2] Anderson, T. W., and C. Hsiao (1982): "Formulation and Estimation of Dynamic Models using Panel Data," *Journal of Econometrics*, 18, 47–82.

[3] Andrews, D. W. K. (2003): "Cross-section Regression with Common Shocks," Cowles Foundation Discussion Paper No. 1428. Yale University.

[4] Andrews, D. W. K. (2001): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 684–734.

[5] Andrews, D. W. K., and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

[6] Andrews, D. W. K., and W. Ploberger (1995): "Admissibility of the Likelihood Ratio Test When a Nuisance Parameter Is Present Only Under the Alternative," *The Annals of Statistics*, 23, 1609–1629.

[7] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

[8] Arthur, W. B. (1999): "Complexity and the Economy," *Science*, 284, 107–109.

[9] Baltagi, B. H., and J. M. Griffin (1997): "Pooled Estimators vs. Their Heterogeneous Counterpart in the Context of Dynamic Demand for Gasoline," *Journal of Econometrics*, 77, 303–327.

[10] Cho, J. S., and H. White (2003): "Testing for Regime Switching," mimeo, Department of Economics, University of California, San Diego.

[11] Conley, T. G. (1999): "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92 1–45.

[12] Conley, T. G., and W. D. Dupor (2003): "A Spatial Analysis of Sectoral Complementarity," *Journal of Political Economy*, 111, 311-352.

[13] Dacunha-Castelle, D., and E. Gassiat (1999): "Testing the Order of a Model Using Locally Conic Parametrization: Population Mixtures and Stationary ARMA processes," *Annals of Statistics*, 27, 1178–1209.

[14] Davies, R. B. (1977): "Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative," *Biometrika*, 64, 247–254.

[15] Davies, R.B. (1987): "Hypothesis testing when a Parameter is Present only under the Alternative," *Biometrika*, 74, 33–43.

[16] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood for Incomplete Data via the EM Algorithm," *Journal of the Royal Statistics Society*, Series B, 39, 1–38.

[17] Deissenberg, C., G. Feichtinger, W. Semmler, and F. Wirl (2001): "History Dependence and Global Dynamics in Models with Multiple Equilibria," *Computing in Economics and Finance 2001*, Paper #257, Society for Computational Economics.

[18] Fazzari, S., R. Hubbard, and B. Petersen (1988): "Financing Constraints and Corporate Investment," *Brookings Paper on Economic Activity*, 1988:1.

[19] Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96, 293–356.

[20] Hansen, B. (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis," *Econometrica*, 64, 413–430.

[21] Hansen, B. (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603.

[22] Haque, N. U., M. H. Pesaran, and S. Sharma (2000): "Neglected Heterogeneity and Dynamics in Cross-country Savings Regressions," in *Panel Data Econometrics: Future Directions, Papers in Honour of Professor Pietro Balestra*, Edited by Krishnakumar and Ronchetti, Elsevier.

[23] Hathaway, R. (1985): "A Constraint Formation of the Maximum Likelihood Estimation for Normal Mixture Distributions," *Annals of Statistics*, 13, 795–800.

[24] Hsiao, C. (2003), *Analysis of Panel Data*, New York: Cambridge University Press.

[25] Islam, N. (1998): "Growth Empirics: a Panel Data Approach — a Reply," *Quarterly Journal of Economics*, 113, 325–329.

[26] Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters," *Annals of Mathematical Statistics*, 27, 887–906.

[27] Laird, N. M. (1978): "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

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

[29] Lee, K., M. H. Pesaran, and R. Smith (1997): "Growth and Convergence in a Multi-country Empirical Stochastic Solow Model," *Journal of Applied Econometrics*, 12, 357-392.

[30] Lee, K., M. H. Pesaran, and R. Smith (1998): "Growth Empirics: a Panel Data Approach — a comment," *Quarterly Journal of Economics*, 113, 319–323.

[31] Lehmann, E. L. (1983): *Theory of Point Estimation*, New York: John Wiley & Sons.

[32] McLachlan, G. J., and T. Krishman (1996): *The EM Algorithm and Extensions*, New York: John Wiley & Sons.

[33] Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

[34] Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Ch. 36, Vol. 4., New York: North Holland.

[35] Neyman, J., and E. L. Scott (1948): "Consistent Estimation from Partially Consistent Observations," *Econometrica*, 16, 1–32.

[36] Phillips, P. C. B., and D. Sul (2003): "Dynamic Panel Estimation and Homogeneity Testing under Cross Sectional Dependence," *The Econometrics Journal*, 6(1) 217–259.

[37] Pollard, D. (1990): *Empirical Process: Theory and Applications*, Hayward: Institute of Mathematical Statistics.

[38] Redner, R. A. (1981): "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions," *Annals of Statistics*, 9, 225–228.

[39] Sun, Y. (2001): "Catching up, Forging ahead, and Falling behind: A Panel Structure Analysis of Convergence Clubs," mimeo, Department of Economics, University of California, San Diego.

[40] Teicher, H. (1963): "Identification of Finite Mixtures," *Annals of Mathematical Statistics*, 34, 1265–1269.

[41] Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985): *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley & Sons.

[42] Train, K. (2002): *Discrete Choice Methods with Simulation*, New York: Cambridge University Press.

[43] White, H. (1994): *Estimation, Inference and Specification Analysis*, New York: Cambridge University Press.

[44] Yakowitz, S., and J. Spargins (1967): "On the Identification of Finite Mixtures," *Annals of Mathematical Statistics*, 39, 209–214.

[45] Zeldes, S. P. (1989): "Consumption and Liquidity Constraints: an Empirical Investigation," *Journal of Political Economy*, 97(2), 305–346.