Consistent Estimation of Finite Mixtures : An Application to Latent Group Panel Structures*

Raphaël Langevin†

July 25, 2024

Abstract

Finite mixtures are often used in econometric analyses to account for unobserved heterogeneity. This paper shows that maximizing the likelihood of a finite mixture of parametric densities leads to inconsistent estimates under weak regularity conditions. The size of the asymptotic bias is positively correlated with the degree of overlap between the densities within the mixture. In contrast, I show that maximizing the max-component likelihood function equipped with a consistent classifier leads to consistency in both estimation and classification as the number of covariates goes to infinity while leaving group membership completely unrestricted. Extending the proposed estimator to a fully nonparametric estimation setting is straightforward. The inconsistency of standard maximum likelihood estimation (MLE) procedures is confirmed via simulations. Simulation results show that the proposed algorithm generally outperforms standard MLE procedures in finite samples when all observations are correctly classified. In an application using latent group panel structures and health administrative data, estimation results show that the proposed strategy leads to a reduction in out-of-sample prediction error of around 17.6% compared to the best results obtained from standard MLE procedures.

Keywords: Panel data, Finite mixtures, EM algorithm, CEM algorithm, K-means, Healthcare expenditures, Unobserved heterogeneity

JEL Codes: C14, C23, C51, I10

*I am grateful to Erin Strumpf, Saraswata Chaudhuri, Philippe Goulet Coulombe, William MacCausland, Victoria Zinde-Walsh, Byoung Park, JoonHwan Cho, Marine Carrasco, Lynda Khalaf, Dante Amengual, Fabian Lange, Enrique Pinzon, David Stephens, Abbas Khalili, Gabriel Rondon Rodriguez, Bryan Graham, Koen Jochmans, Nicholas Brown, Brantly Callaway, Peng Shao, Jad Beyhum, Masamune Iwasawa, and Aristide Houndetoungan for helpful comments on early drafts of this paper. Special thanks to Pierre-Carl Michaud, David Boisclair, and François Laliberté-Auger for providing access to computational hardware and technical advice. I also thank the participants of NY Camp Econometrics XVII, the 18^{th} CIREQ PhD Student Conference, the 57^{th} Annual Meeting of the Canadian Economics Association, the 2023 Stata Conference in Stanford, the 2024 RCEA International Conference at Brunel University London, the 63^{th} SCSE Annual Congress at HEC Montréal, the 2024 NASMES at Vanderbilt University, and the 2024 IAAE Annual Conference in Thessaloniki for advice and comments. This work was supported by the grant no.767-2020-2809 of the Social Sciences and Humanities Research Council of Canada (SSHRC).

†Department of Economics, McGill Unversity, 855 Sherbrooke W Street, H3A 2T7, Montréal, Canada. E-mail: raphael.langevin@mail.mcgill.ca.

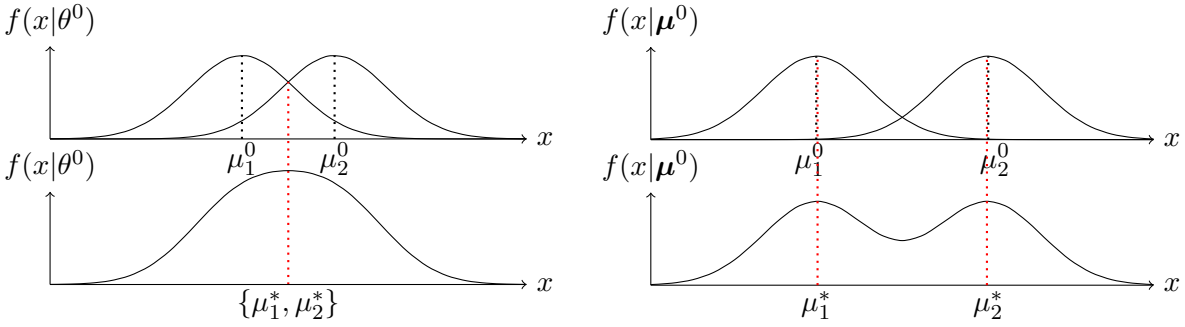
1 Introduction

Finite mixtures are extensively used in statistics, computer science, and machine learning for pattern recognition and unsupervised classification to account for various types of unobserved heterogeneity (Bishop, 2006; Frühwirth-Schnatter, 2006). Several applications of finite mixtures can also be found in labor and health economics (Heckman and Singer, 1984; Deb and Trivedi, 1997; Keane and Wolpin, 1997; Jones et al., 2015). For instance, Deb and Trivedi (1997) use finite mixtures to distinguish two unobserved, latent types (i.e. the "healthy" and the "ill") regarding the demand for medical care and find substantial differences in fitted distributions for each type. Methods to account for unobserved heterogeneity can reduce bias (Hsiao, 2014), improve inference and forecast (Boot and Pick, 2018), and also allow for the estimation of heterogeneous treatment effects (Ahn and Kasahara, 2024).

Conceptually speaking, a finite mixture distribution is a convex combination of a small number of distinct parametric densities where the combination weights, known as *mixing weights*, correspond to the proportion of observations that originate from each density in the population. The resulting density, known as the *mixture density*, corresponds to a well-defined density that fully describes the distribution of the observed data. Finite mixtures are related to unsupervised clustering methods given that each *component density* within the mixture fully describes its corresponding group of observations. Finite mixtures are also known to be highly flexible since they can easily accommodate the presence of covariates and nonlinear models. Estimation of the parameters contained in the mixture density is usually performed using maximum likelihood estimation (MLE) procedures, including nonparametric MLE (Compiani and Kitamura, 2016).

In this paper, I show that maximizing the likelihood of a mixture density using standard parametric MLE methods leads to inconsistent estimates of all parameters in the mixture if the distance between each component density is finite. The main issue resides in the estimation of the mixing weights : the MLE of the mixing weights is not well-defined and conventional estimates will not converge to their true values unless all component densities are infinitely distant from each other. I am not aware of any estimator of the mixing weights that would be consistent regardless of the distance between the component densities. The issue is similar to the case of profile maximum likelihood when the nuisance parameters are not consistently estimated, hence leading to an asymptotic bias in the parameters of interest. The issue is also similar to the incidental parameter problem initially described by Neyman and Scott (1948) in the sense that the inconsistencies vanish as the asymptotic distributions of all component densities get further away from each other.

Figure 1 illustrates the problem using a simple mixture model of two normal densities with different mean values and equal mixing weights/variances. The true mean values are denoted by μ_1^0 and μ_2^0 on both graphs, while maximum likelihood estimates for the means are denoted by μ_1^* and μ_2^*. When the two densities are very close to each other (panel (a)), the mixture is confounded with



(a) True mean values: $\mu^0 = (-0.35, 0.35)$ (b) True mean values: $\mu^0 = (-0.65, 0.65)$

Figure 1: Two mixture distributions of two normal densities with different mean values, equal mixing weights, and equal variances. The estimates provided by MLE in each case are represented by μ_1^* and μ_2^*, whereas the true mean values are represented by μ_1^0 and μ_2^0.

a single-component normal density unless the true mixing weights, π_1^0 and π_2^0, are known.[1] As a result, one of the two mixing weights will be very close to unity and its corresponding estimate for the mean will be approximately equal to the (weighted) average of the true mean values. Panel (b) of Figure 1 shows that the maximum likelihood estimates of the means (and the mixing weights) will converge to their true values as the densities get further away from each other. Those statements are confirmed by simulation results that are presented in Section 4.

The contributions of the paper are twofold. First, I show why the standard MLE of finite mixtures leads to inconsistent estimates of all parameters in the mixture under weak regularity conditions.[2] By standard MLE of finite mixtures, I refer to procedures that attempt to globally maximize the mixture likelihood function, as defined in Section 3.1. The well-known expectation-maximization (EM) algorithm of Dempster et al. (1977) and all Newton-type algorithms (when applied to the mixture likelihood) fall into this category. The inconsistency proof developed in this paper is general with respect to the component densities, meaning that it applies to any finite mixtures regardless of the nature of the component densities.

The second contribution of the paper is to show under which conditions the maximization of a different objective function, the *max-component likelihood* (MCL) function, can lead to consistent estimation of all parameters in the mixture. Similar to Dzemski and Okui (2021), I show that consistency in estimation can be obtained by maximizing the MCL function provided the use of a consistent classifier, thus implying that the proportion of misclassified observations goes to zero in the limit. Efficient estimation of all parameters is also possible under standard regularity conditions

[1] One could test, for instance, the kurtosis of the empirical distribution to assess the presence of a mixture. However, such a test is likely to have a low power against the null hypothesis of no mixture in finite samples when the true mean and variance values are close to each other. For more details on tests for finite mixtures, see Amengual et al. (2024) and references therein.

[2] By "all parameters" in the mixture, I always refer to the mixing weights and the set of parameters governing each density in the mixture. Note also that "components" and "groups" have the same meaning and are interchangeable.

if all observations are simultaneously classified within their true component as the sample size goes to infinity (Su et al., 2016). Contrary to the literature on latent group panel structures, the proposed estimation strategy has the benefit of leaving group membership completely unrestricted over any dimension of the dataset, as is usually the case in finite mixture models.

I illustrate those two theoretical contributions by comparing the finite-sample performance of the proposed estimation strategy to standard MLE procedures through the use of the EM algorithm with both simulated and real-world data. Simulation results confirm the inconsistency of standard MLE of finite mixtures under finite distance between each component density. Simulation results also show that maximizing the MCL using the proposed estimation algorithm outperforms the EM algorithm in terms of estimation error when the mixture features no small-sized component and the misclassification rate is near (or equal to) zero at the true parameter values. More precisely, those results show that the mean and variance parameters estimated by each algorithm become more biased as the misclassification rate goes up. Those estimation errors do not vanish as the sample size increases, hence confirming the presence of an asymptotic bias when the component densities are not sufficiently distant from each other. Simulation results also show that no algorithm clearly outperforms the other as the misclassification rate increases away from zero.

The empirical part of the paper uses the EM algorithm and the proposed estimation algorithm to model individual healthcare expenditure (HCE) from administrative data over time. Both algorithms use a latent group two-part model (LGTPM) for estimation and identical specification of each component density. Empirical results show that the proposed algorithm leads to a reduction in out-of-sample prediction error of 17.6% compared to the best result obtained when using the EM algorithm, and of 56.6% compared to the single-component two-part model. The proposed algorithm also allows to recover the true group memberships for almost every observation in the sample, which can be dynamically modeled in a second step for forecasting purposes.

The remainder of the paper is as follows. Section 2 briefly reviews the related literature. Section 3 shows why maximizing the mixture likelihood leads to inconsistencies under weak regularity conditions. It also details the proposed estimation strategy that is used to solve the inconsistency issue. Section 4 presents simulation scenarios and results, confirming the theoretical insights from the previous section. Section 5 presents the empirical application and the related results, while Section 6 concludes. All proofs and additional results can be found in the appendices.

2 Related Literature

Finite Mixtures and the EM Algorithm. In practice, finite mixture models are almost always estimated via the EM algorithm although other numerical optimization algorithms can be used.[3] The EM algorithm consists of the consecutive repetition of an expectation step (i.e. the

[3]It is important to note that the EM algorithm is not a maximization algorithm *per se*, but a general strategy for the maximization of any incomplete-data likelihood function. Indeed, it is possible and often desirable to use a

E-step) and a maximization step (i.e. the M-step), each conditional on the results obtained from the previous step. The E-step computes assignment probabilities for each observation and each component density, whereas the M-step maximizes the likelihood function conditional on the most recent assignment probabilities. In their seminal paper, Dempster et al. (1977) showed that such an algorithm never decreases the *incomplete-data likelihood* function between two consecutive iterations, the mixture likelihood being a special case of the former when the unobserved information is assumed to have discrete support.

However, the multimodal nature of the mixture likelihood makes it difficult for any optimization algorithm to find its global maximum, giving rise to various procedures to reduce the probability of being "stuck" in a local spurious maximum (Celeux, 2019). Moreover, it is acknowledged that "the sample size [...] has to be very large, before asymptotic theory of maximum likelihood applies" (McLachlan and Peel, 2000) when applying MLE to the mixture likelihood, especially when the "component densities are poorly separated" (Redner and Walker, 1984; Aragam and Yang, 2023). Given that the mixture density is well-defined, it has been claimed that maximizing the mixture likelihood function will necessarily lead to consistent estimates of all the mixture parameters (Redner and Walker, 1984; Chen, 2017). Nonetheless, the practical difficulties often encountered with standard MLE of finite mixtures did not cast doubt on its ability to converge to the true parameter values. The inconsistency shown in Section 3.2 helps to explain why it is frequent to observe such practical difficulties while maximizing the mixture likelihood function.

The results developed in this paper generalize the conclusion of Kwon and Caramanis (2019) to any kind of mixture. The authors show that the EM algorithm will converge to the true parameter values in the context of a mixture of linear regressions if the component densities are sufficiently distant from each other and if the algorithm is initialized with parameters that are sufficiently close to the true values. The main contribution of this paper is the development of an estimation procedure that is able to converge to the true parameter values when the EM algorithm and standard MLE of finite mixtures cannot.

K-means and the CEM Algorithm. The K-means algorithm is one of the most widely used clustering algorithms in unsupervised machine learning (Hastie et al., 2009). It can be used to estimate finite mixture models just as the EM algorithm, but is less flexible than the EM. Moreover, it is known to yield *inconsistent* estimates of all parameters in the mixture (Pollard, 1981; Bryant, 1991). The CEM algorithm (for classification EM) is the likelihood generalization of K-means and is as flexible as the EM, but still leads to inconsistent estimates of the mixture parameters (Bryant and Williamson, 1978; Celeux and Govaert, 1992). Given that the K-means algorithm corresponds to the special case of the CEM algorithm when all errors are independent, identically distributed, and homoskedastic, the rest of the paper will focus on the general, more flexible CEM algorithm.

numerical optimization method at each M-step of the EM algorithm. See Section 5 of this paper and Section 2.4.4 of Frühwirth-Schnatter (2006) for more details on the subject.

Contrary to the EM algorithm, the CEM algorithm classifies each observation to the group that maximizes its corresponding density value. This practical difference leads to a theoretical difference, which is that the CEM algorithm maximizes the MCL function rather than the mixture likelihood function. The MCL function is the estimated counterpart of the *complete-data* likelihood function, the latter being the likelihood function one would get if all group memberships were known (Gepperth and Pfülb, 2021). If this is the case, then it is said that the estimator achieves the *oracle property* and all estimated parameters are efficient and asymptotically normal as a consequence of standard, single-component MLE (Su et al., 2016).

Akin to the EM algorithm, it has been shown that the MCL function never decreases between two consecutive iterations of the CEM algorithm (Bottou and Bengio, 1994). Given that the MCL function is also multimodal, convergence of the CEM algorithm to the global maximum of the MCL function is rarely guaranteed. Therefore, similar practical issues are encountered when maximizing either the mixture likelihood or the MCL function (Samé et al., 2007). It is however worth noting that global convergence of the estimation algorithm is not very helpful if the global maximum of the objective function is not located at the true parameter values asymptotically.

Latent Group Panel Structures. Several recent studies in econometrics have used different variants of the K-means algorithm to account for unobserved heterogeneity in panel data through the introduction of grouped fixed-effects (GFE) and/or group-specific coefficients (Bonhomme and Manresa, 2015; Bonhomme et al., 2019, 2022; Okui and Wang, 2021; Lumsdaine et al., 2023; Liu et al., 2020; Wang et al., 2024). The GFE estimator is closely related to factor models since it allows time-fixed effects to vary across groups, as in models with interactive fixed effects (Bai, 2009). However, factor models typically assume homogeneity of all the other parameters in the model, whereas the GFE estimator impose restrictions on group memberships over time or on the form taken by the unobserved heterogeneity to achieve consistency (Bonhomme and Manresa, 2015; Okui and Wang, 2021; Bonhomme et al., 2022; Lumsdaine et al., 2023; Wang et al., 2024). The estimation procedure proposed in this paper completely relaxes those two assumptions without sacrificing consistency under the "many covariates" asymptotics where the number of covariates p is allowed to grow at a strictly slower rate than the sample size (Cattaneo et al., 2018b,a).

Instead of employing the K-means algorithm, several authors relied on the LASSO device or on binary segmentation algorithm for combining unit-level coefficients into group-level coefficients (Su et al., 2016, 2019; Qian and Su, 2016; Wang et al., 2018, 2019; Wang and Su, 2021). All estimation strategies presented in this kind of papers impose certain restrictions on group memberships over time. For instance, in the case of the classifier-LASSO of Su et al. (2016), leaving group membership completely unrestricted would imply the estimation of NT initial parameters for each time-varying covariate in the sample, which is impractical in regular panel datasets. This is why this paper focuses exclusively on regular finite mixtures where no restriction is imposed on group memberships along any of the dimensions of the dataset.

3 Maximum Likelihood Estimation of Finite Mixtures

This section is divided into three parts. Section 3.1 presents the general framework that will be used throughout the paper. The general case considered is panel data when the ratio N/T is relatively large. Section 3.2 then shows why standard MLE of finite mixtures is inconsistent under weak regularity conditions. Extension of all inconsistency proofs to cross-sections, with or without covariates, is straightforward. Finally, Section 3.3 describes the proposed estimation strategy and shows under which conditions it will lead to consistent estimation of all parameters in the mixture.

3.1 General Setup and Notation

The mixture density function of any observation $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X} \subseteq \mathbb{R}^{p+1}$ is generally defined as follows

$$f(y_{it}|x_{it};\theta,\pi) := \sum_{g=1}^{G} \pi_g f_g(y_{it}|x_{it};\theta_g) \equiv \sum_{g=1}^{G} \pi_g f_g(y_{it}|x_{it};\theta), \tag{1}$$

where y_{it} is the observed univariate outcome of individual i at period t with $i = 1, ..., N$ and $t = 1, ..., T$, and where x_{it} is a p-sized column vector of strictly exogenous covariates.[4] The function $f : \mathcal{Y} \times \mathcal{X} \subseteq \mathbb{R}^{p+1} \to \mathbb{R}_{>0}$ is the mixture density, defined as a function of $\theta = (\theta_1, ..., \theta_g, ..., \theta_G)$, the set of *component parameters*, and of $\pi = (\pi_1, ..., \pi_G)$, the vector of mixing weights. The set of mixing weights π correspond to the *mixing distribution*, where G is the total number of components. The set of component parameters θ is assumed to lie inside the compact parameter space Θ, whereas the vector of mixing weights π is assumed to lie within the open space $\Pi = (0,1)^G$ with the unit constraint $\sum_{g=1}^{G} \pi_g = 1$. Such a setup is also called a *mixture of experts* when the vector of mixing weights π is allowed to depend on the covariates (Bishop, 2006).

Each component density $f_g(.)$ corresponds to a well-defined parametric density with respect to a σ-finite measure, generally denoted by $\upsilon(dy_{it})$. All component densities do not necessarily belong to the same family of distributions, although this is common in practice and also facilitates the estimation of θ. In this paper, the vector of covariates x_{it} is treated as a p-variate random variable that is drawn from the population. All x_{it}'s are assumed to be independent of each other but are not necessarily identically distributed (see Assumption 2 for more details). For simplicity, it is also assumed that $\mathcal{Y}$ is independent of the values taken by $x_{it} \in \mathcal{X}$, such that $\int_{\mathcal{Y}} f_g(y_{it}|x_{it};\theta_g)\upsilon(dy_{it}) = 1$ for any value of $g \in \{1, ..., G\} = \mathbb{G}$ and any value of $x_{it} \in \mathcal{X}$.

Throughout the paper, the set of true parameter values is denoted by (π^0, θ^0) and lies in the interior of the parameter space $\Pi \times \Theta$, where $\pi^0 = (\pi_1^0, ..., \pi_g^0, ..., \pi_G^0)^\top$ and $\theta^0 = (\theta_1^0, ..., \theta_g^0, ..., \theta_G^0)$. Hence, the corresponding true mixture density is denoted by $f(y_{it}|x_{it};\theta^0,\pi^0) \equiv \sum_{g=1}^{G} \pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)$,

[4] For simplicity, it is assumed for now that all observations are i.i.d., but serial correlation within units is taken into account in the (asymptotic) distribution of the proposed estimator.

where the true values (π^0, θ^0) are unobserved by the econometrician. Any dataset generated by the true mixture density is denoted by $(\mathbf{y}, \mathbf{x})$, where $\mathbf{y} = (y_{11}, ..., y_{it}, ..., y_{NT})^\top$, and $\mathbf{x} = (x_{11}, ..., x_{it}, ..., x_{NT})^\top$. Note that the set of true mixing weights π^0 needs not be a function of (a subset of) θ^0. Note also that, for convenience, the true number of components G is assumed to be discrete, finite, and known by the econometrician unless stated otherwise.

In the finite mixture framework, each observation is assumed to originate from only one of the G densities. The unobserved, true binary assignment (or grouping) variable, denoted by z^0_{itg}, is defined as follows

$$z^0_{itg} := \begin{cases} 1 & \text{if and only if } y_{it} \text{ is generated by } f_g(\cdot|x_{it}; \theta^0), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Analogously, the true categorical assignment variable (or group membership), z^0_{it}, is defined such that $z^0_{it} = g$ if and only if $z^0_{itg} = 1$. Consequently, the true mixing weight of the g^{th} component density, π^0_g, can be defined as follows

$$\pi^0_g := \operatorname*{plim}_{N,T\to\infty} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{z^0_{itg}}{NT} = \mathbb{P}[z^0_{itg} = 1], \tag{3}$$

which corresponds to the unconditional probability of any observation to belong to the g^{th} component/group. Note that $z^0_{itg} = 1$ for at least one pair (i, t) and for any $g \in \mathbb{G}$ since $\pi^0_g \in (0, 1)$ for any $g \in \mathbb{G}$. Such a definition of π^0_g also implies that the value z^0_{itg} can be seen as the realization of the random variable Z_{it} drawn from a univariate multinational distribution with G categories and vector of probabilities $(\pi^0_1, ..., \pi^0_G)$ (McLachlan et al., 2019).

The infeasible maximum likelihood estimator where all true group memberships are known is the estimator that maximizes the *complete-data log likelihood* function, which is defined as follows

$$l^C(\theta, \pi) := \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} z^0_{itg} \log(\pi_g f(y_{it}|x_{it}; \theta)) \equiv \sum_{i=1}^{N}\sum_{t=1}^{T} \log(\pi_{z^0_{it}} f_{z^0_{it}}(y_{it}|x_{it}; \theta)). \tag{4}$$

Maximizing $l^C(\theta, \pi)$ with respect to θ is similar to maximizing G distinct log likelihood functions given that all observations are associated with their true component. The vector of mixing weights may then be estimated using eq.(3). Since z^0_{itg} is unobserved, two alternative objective functions have been used to estimate both π^0 and θ^0. The first objective function corresponds to the mixture (log) likelihood function and is defined as follows

$$l(\theta, \pi) := \sum_{i=1}^{N}\sum_{t=1}^{T} \log(\sum_{g=1}^{G} \pi_g f_g(y_{it}|x_{it}; \theta)), \tag{5}$$

whereas the second objective function corresponds to the MCL function

$$l^{MC}(\theta, \mathbf{z}) := \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} z_{itg} \log(f_g(y_{it}|x_{it}; \theta)), \tag{6}$$

where $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_g, ..., \mathbf{z}_G)$, with $\mathbf{z}_g = (z_{11g}, ..., z_{itg}, ..., z_{NTg})^\top$, and where each $z_{itg} = \{0, 1\}$ with $\sum_{g=1}^{G} z_{itg} = 1$. It is easy to see that if $z_{itg} = z^0_{itg}$ for all triples (i, t, g), then $l^{MC}(\theta, \mathbf{z})$ would be equal to the infeasible estimator and the oracle property would be achieved. Given that a non-null proportion of observations are asymptotically misclassified (unless the component densities are infinitely distant from each other), maximizing the MCL function yields asymptotically biased estimates of all parameters in the mixture (Celeux and Govaert, 1992; McLachlan et al., 2019).

It is commonly acknowledged that maximizing the mixture likelihood function with respect to (θ, π) will yield consistent and asymptotically normal estimates of the mixture parameters (Redner and Walker, 1984; Chen, 2017). The next subsection shows precisely why such a statement is false unless all component densities are infinitely distant from each other, as is the case when maximizing $l^{MC}(\theta, \mathbf{z})$ with respect to θ when $\mathbf{z} \neq \mathbf{z}^0$, where $\mathbf{z}^0 = (\mathbf{z}^0_1, ..., \mathbf{z}_g, ..., \mathbf{z}^0_G)$, with $\mathbf{z}^0_g = (z^0_{11g}, ..., z^0_{itg}, ..., z^0_{NTg})^\top$, is the set containing all true group memberships.[5]

3.2 Inconsistency of Standard MLE of Finite Mixtures

This subsection shows why maximizing the mixture likelihood leads to inconsistent estimation of all parameters in the mixture under weak regularity conditions. Those regularity conditions are described below in Assumption 1.

Assumption 1.

(i) $\mathbb{E}_0[\log(f(y_{it}|x_{it}; \theta, \pi))] < \infty$ *for any* $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$, *any* $\theta \in \Theta$, *and any* $\pi \in \Pi$, *where* $\mathbb{E}_0[\cdot]$ *stands as the expected value with respect to the true mixture density, and where* x_{it} *is a set of strictly exogenous covariates.*

(ii) $f_g(y_{it}|x_{it}; \theta_g) > 0$ *for any* $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$, *for any* $\theta_g \subset \theta \in \Theta$, *and for any* $g \in \mathbb{G}$, *where all* (y_{it}, x_{it}) *are i.i.d. conditional on belonging to the same component, with* $\theta_g = \theta_j \Leftrightarrow g = j$, *and where all component densities share the same support.*

(iii) $l(\theta_g) := \sum_{i=1}^{N} \sum_{t=1}^{T} \log(f_g(y_{it}|x_{it}; \theta_g))$ *features a unique maximum with respect to* θ_g *for any* $g \in \mathbb{G}$ *and any dataset* $(\mathbf{y}, \mathbf{x})$.

(iv) $l(\theta, \pi) = l(\theta', \pi') \Leftrightarrow \theta = \theta'$ *and* $\pi = \pi'$ *up to any permutation in the labels of* (θ, π) *and* (θ', π'), *for any pair* $(\theta, \theta') \in \Theta \times \Theta$, *any pair* $(\pi, \pi') \in \Pi \times \Pi$, *and any dataset* $(\mathbf{y}, \mathbf{x})$.

(v) $l(\theta, \pi)$ *features a unique global maximum with respect to* (θ, π) *for any dataset* $(\mathbf{y}, \mathbf{x})$.

[5] Infinite distance between each component density implies that $\int_{\mathcal{Y}} f_g(y_{it}|x_{it}; \theta^0_g) \times f_j(y_{it}|x_{it}; \theta^0_j) \upsilon(dy_{it}) \to 0$ as $||\theta^0_g - \theta^0_j|| \to \infty$ for any pair $(g, j) \in \mathbb{G} \times \mathbb{G} \backslash g$, and any $x_{it} \in \mathcal{X}$, where $||\cdot||$ denotes the Euclidean norm.

(vi) $f(y_{it}|x_{it};\theta,\pi)$ *is continuously differentiable with respect to both* θ *and* π *for all pairs* $(y_{it},x_{it}) \in \mathcal{Y}\times\mathcal{X}$.

Assumption 1(i) rules out cases where the log likelihood function is not bounded from above. In practice, this is satisfied provided appropriate constraints on key parameters of all component densities or by the use of penalized MLE (McLachlan et al., 2019; Tanaka, 2009). Note that this assumption directly implies that $\mathbb{E}_{0,g}[\log(f_g(y_{it}|x_{it};\theta_g))] < \infty$ for all $g \in \mathbb{G}$, where $\mathbb{E}_{0,g}[\cdot]$ is the expected value with respect to the true density of the g^{th} component density.

Assumption 1(ii) states that the possible values taken by all component densities are restricted to be strictly positive on the whole support of each density, where this support is assumed to be identical across densities. Assumption 1(ii) also assumes that all observations are independently and identically distributed conditional on belonging to the same component. Such an assumption is made for simplicity since it implies that each y_{it} generated by the same component density follows a stationary process over time. Stationarity can however be relaxed at the expense of additional modeling complexities. Assumption 1(ii) also implies that two different component densities cannot identical parameter values.

Assumption 1(iii) is standard in most MLE problems. Unimodality of each component density ensures that the maximization of any single-component likelihood function with respect to θ_g will always yield a single set of estimated parameter values, which is located at the global maximum of the corresponding likelihood function. Assumption 1(iv) is commonly known as *generic identifiability* of the mixture density. Section 1.3 of Frühwirth-Schnatter (2006) describes the three types of identification issues that can arise when modeling finite mixtures, including generic identifiability. The two other identification issues are characterized as "weak" since those issues can be overcome by the use of appropriate constraints on the component labels and the number of components.

Assumption 1(v) is similar to Assumption 1(iii), but applied to the mixture likelihood function. It is different from Assumption 1(iii) since it eliminates situations where the empirical mixture likelihood would have two identical global maxima, but at two different locations in $\Theta\times\Pi$ for a given dataset. Finally, Assumption 1(vi) is not crucial for the rest of the paper, although it greatly simplifies the proof of Corollary 3.1, which is stated below along with the first lemma of the paper.

Lemma 3.1. *Let Assumption 1 hold, and let* ζ *be any open subset of the space* Π *that does not contain the set of true mixing weights,* π^0. *Then the following inequality always holds :*

$$\mathbb{E}_0[\log(f(y_{it}|x_{it};\theta^0,\tilde{\pi}))] < \mathbb{E}_0[\log(f(y_{it}|x_{it};\theta^0,\pi^0))],$$

for any $\tilde{\pi} \in \zeta$.

Corollary 3.1. *Let Assumption 1 hold, and let* ζ *be any open subset of the space* Π *that does not*

contain the set of true mixing weights, π^0. Let also the estimator $\hat{\theta}_{NT}(\pi)$ be defined as follows

$$\hat{\theta}_{NT}(\pi) := \arg\max_{\theta \in \Theta} l(\theta, \pi),$$

such that $\text{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\pi^0) = \theta^0$. *Then, for almost every vector $\tilde{\pi} \in \zeta$, we have*

$$\operatorname*{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\tilde{\pi}) \neq \theta^0.$$

Lemma 3.1 and Corollary 3.1 formalize the logic behind the (in)consistency of pseudo maximum likelihood estimates in the sense of Gong and Samaniego (1981) : for almost every $\tilde{\pi} \neq \pi^0$, the maximum likelihood estimates $\hat{\theta}_{NT}(\tilde{\pi})$ will not converge in probability to θ^0 as the sample size increases. Although there could exist a finite, countable set $A \subset \zeta$ such that every $\pi_A \in A$ could satisfy $\text{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\pi_A) = \theta^0$, this set is of measure zero under Assumption 1(vi). Analogously, it can be shown that maximizing the mixture likelihood with respect to π will not converge to π^0 for almost every value of $\theta \neq \theta^0$. A concrete example of non-convergence of θ when $\tilde{\pi} \neq \pi^0$ is given in Appendix A.19 using a mixture of two exponential densities.

Therefore, it is important to know whether π converges to π^0 or not as the sample size increases. This depends on the choice of estimator for π conditional on $\theta = \theta^0$. The next proposition shows that applying a profile maximum likelihood strategy on the mixture likelihood conditional on θ leads to a corner solution in terms of π.

Proposition 3.1. *Let Assumption 1 hold, and define the mixture log likelihood function as in eq.(5). Then*

$$\hat{\pi}_\theta := \arg\max_{\pi \in \Pi} l(\theta, \pi) = e_G,$$

for almost every value of $\theta \in \Theta$, where e_G is G-sized vector with all elements equal to zero, except for a single element which is equal to one.

Proposition 3.1 shows that the "exact" MLE of π is not well-defined. Instead of maximizing the profile likelihood, each one of the G mixing weights can be estimated by averaging over the posterior probability that each observation belongs to each component. For the it^{th} observation, such a posterior probability is defined as follows (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006)

$$\tau_g(y_{it}|x_{it};\theta,\pi) \equiv \tau_{itg}(\theta,\pi) := \frac{\pi_g f_g(y_{it}|x_{it};\theta_g)}{\sum_{l=1}^{G} \pi_l f_l(y_{it}|x_{it};\theta_l)}. \tag{7}$$

The posterior probability $\tau_{itg}(\theta,\pi)$ comes from the application of Bayes' rule on $\mathbb{P}[z^0_{itg} = 1|y_{it}, x_{it}, \theta]$ for a given observation and a given value of $g \in \mathbb{G}$. This definition naturally leads to the following

estimator for the g^{th} mixing weight π_g

$$\hat{\pi}_g(\theta) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\hat{\pi}), \tag{8}$$

where the presence of $\hat{\pi}_g$ on both sides of the equation is circumvented by the iterative nature of the estimation algorithm. Redner and Walker (1984) refer to this estimator as the "approximate MLE" of π_g given that it maximizes the conditional expectation of the complete-data log likelihood (also called the "Q-function", (Celeux, 2019)) rather than the mixture log likelihood, as is shown in the next lemma.

Lemma 3.2. *Let Assumption 1 hold, and define the conditional expectation of the complete-data log likelihood as follows*

$$\mathbb{E}_z[l^C(\theta,\pi)] := \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G}\tau_{itg}(\theta,\pi)\log(\pi_g f(y_{it}|x_{it};\theta)),$$

where $\tau_{itg}(\theta,\pi)$ is defined as in eq.(7). Then the estimator $\hat{\pi}(\theta) = (\hat{\pi}_1(\theta),...,\hat{\pi}_G(\theta))$ as shown in eq.(8) maximizes $\mathbb{E}_z[l^C(\theta,\pi)]$ conditional on θ and conditional on all posterior probabilities.

Other estimators of the mixing weights are sometimes used in applied studies, but those estimators often require the estimation of an additional set of parameters and the imposition of a specific relationship between the extra parameters and the mixing weights.[6] The main benefit of eq.(8) is that the mixing weights depend only on θ and on previous estimates of π (or a prior distribution of π). The main theoretical result of this subsection is stated in Theorem 3.1.

Theorem 3.1. *Let Assumption 1 hold, and define $\hat{\pi}_g(\theta)$ as in eq.(8). Then*

$$\operatorname*{plim}_{N,T\to\infty}\hat{\pi}_g(\theta^0) \neq \pi_g^0$$

for any $g \in \mathbb{G}$ unless all component densities are infinitely distant from each other.

Corollary 3.2. *Let Assumption 1 hold, and define $\hat{\pi}_g(\theta)$ and $\tau_{itg}(\theta,\pi)$ as in eq.(8) and eq.(7) respectively. Then*

$$\operatorname*{plim}_{N,T\to\infty}\hat{\theta}_{NT}(\hat{\pi}(\theta^0)) = \operatorname*{plim}_{N,T\to\infty}\arg\max_{\theta\in\Theta} l(\theta,\hat{\pi}(\theta^0)) \neq \theta^0,$$

unless all component densities are infinitely distant from each other.

Theorem 3.1 shows that $\hat{\pi}_g(\theta^0)$ will not converge to π_g^0 when all component densities are finitely distant from each other. Consequently, θ cannot be consistently estimated by maximizing $l(\theta,\hat{\pi}(\theta^0))$

[6] See, for instance, Neelon et al. (2011) and Kasteridis et al. (2022) for alternative estimators of π_g in applied health studies.

with respect to θ unless all component densities are infinitely distant from each other, as shown in Corollary 3.2. Therefore, π can be regarded as incidental parameters that will bias the estimation of θ when they are not consistently estimated. To my knowledge, no estimator of π has been proven to be consistent under standard regularity conditions when the true group memberships are unknown unless the proportion of misclassified observations goes to zero in the limit. This latter avenue is explored in more detail in Section 3.3.

The next two propositions describe some properties of standard MLE of finite mixtures when the g^{th} estimated mixing weight is defined as follows

$$\hat{\pi}_g(\theta) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{f_g(y_{it}|x_{it};\theta_g)}{\sum_{l=1}^{G} f_l(y_{it}|x_{it};\theta_l)}. \tag{9}$$

Proposition 3.2. *Let Assumption 1 hold. Let also $\hat{\pi}_g(\theta)$ be defined as in eq.(9) for any $g \in \mathbb{G}$, and let $\pi_g^0 = 1/G$ for any $g \in \mathbb{G}$. Then $\operatorname{plim}_{N,T\to\infty} \hat{\pi}_g(\theta^0) = \pi_g^0$ for any $g \in \mathbb{G}$.*

Proposition 3.3. *Let Assumption 1 hold, and let $\hat{\pi}_g(\theta)$ be defined as in eq.(9). Then*

$$\lim_{f_g(y_{it}|x_{it};\theta_g^0)\to f_j(y_{it}|x_{it};\theta_j^0)} \operatorname*{plim}_{N,T\to\infty} \hat{\pi}_g(\theta^0) = 1/G$$

for any pair $(g, j) \in \mathbb{G} \times \mathbb{G}\backslash g$.

Proposition 3.2 shows that the estimated mixing weights will necessarily converge to their true values if the mixing weights are estimated using eq.(9) and if the true mixing weights are all equal to $1/G$. This case is similar to the one where the prior distribution of the parameters of interest is identical to their true distribution. However, it occurs too infrequently to be considered relevant in practice. Nonetheless, if one has good reason to think that all component densities are of similar size, then using eq.(9) instead of eq.(8) and eq.(7) to estimate π should lead to smaller bias, especially when the component densities are very close to each other. The intuition behind both propositions is confirmed by simulations that are described in Section 4.1.1.

3.3 Consistent Estimation of Finite Mixtures

As described in the literature on latent group panel structures, consistent estimation of the parameters within each component is possible if the estimation procedure correctly classifies all observations as $N, T \to \infty$ (Su et al., 2016). In this subsection, I present a general strategy for the consistent estimation of finite mixtures that relies on the availability of a consistent, unsupervised classification procedure. More precisely, I show that the use of the Mahalanobis distance leads to a consistent classifier such that, under the additional assumptions stated below, its misclassification rate goes to (or equals) zero as $p \to \infty$. For brevity, some important assumptions described in Assumption 1 are not copied below in Assumption 2.

Assumption 2.

(i) $\mathbb{E}_0[\log(\sum_{g=1}^{G} z_{itg} f_g(y_{it}|x_{it};\theta_g))] < \infty$ *for any* $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$ *and any* $\theta \in \Theta$*, where* x_{it} *is a set of strictly exogenous covariates with at least one continuous covariate.*

(ii) $f_g(\mathbf{y}|\mathbf{x};\theta_g) = f_j(\mathbf{y}|\mathbf{x};\theta_j) \Leftrightarrow \theta_g = \theta_j$ *for any dataset* $(\mathbf{y}, \mathbf{x})$ *and any pair* $(j, g) \in \mathbb{G} \times \mathbb{G}$*.*

(iii) $f_g(y_{it}|x_{it};\theta_g)$ *is continuously differentiable with respect to* θ_g *for all pairs* $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$*.*

(iv) There exists a set of true parameter values $\xi^0 = (\xi_1^0, ..., \xi_g^0, ..., \xi_G^0) \in \Xi$*, where* Ξ *is a compact parameter space, and a p-variate density, denoted by* $f_g(\cdot|\xi_g) > 0$*, that generates* x_{it} *if and only if* $z_{itg}^0 = 1$ *for each* $g \in \mathbb{G}$ *such that each density* $f_g(\cdot|\xi_g^0)$ *shares the same p-variate support, is continuously differentiable with respect to any* $\xi \in \Xi$*, and such that each* $f_g(x_{it}|\xi_g)$ *integrates to unity over* $\mathcal{X}$ *with respect to a* σ*-finite measure denoted by* $\nu(dx_{it})$*.*

(v) $\mathbb{E}[x_{it}] = \boldsymbol{\mu}_g^0 = (\mu_{g1}^0, ..., \mu_{gp}^0)^\top$ *and* $Var[x_{it}] = \Sigma_g^0$ *if and only if* $z_{itg}^0 = 1$ *such that* $(\boldsymbol{\mu}_g^0, \Sigma_g^0) \subseteq \xi_g^0$*, where* $||\boldsymbol{\mu}_g^0||^2 < \infty$ *for any* $g \in \mathbb{G}$*, and where* Σ_g^0 *is a* $p \times p$ *positive-definite matrix with diagonal elements* $0 < \sigma_{g,ii}^2 < \infty$ *and off-diagonal elements* $0 < |\sigma_{g,ij}| < \sigma_{g,ii}\sigma_{g,jj}$ *with* $Cov[x_{it}, x_{js}] = 0$ *for any pair* $(j, s) \neq (i, t)$*.*

(vi) $\Sigma_j^0 \neq \Sigma_g^0$ *for any pair* $(g, j) \in \mathbb{G} \times \mathbb{G} \backslash g$*, and for a nonvanishing proportion of the elements in each covariance matrix as* $p \to \infty$*.*

Assumption 2(i) is the analog of Assumption 1(i) in the context of the MCL, but also assumes that at least one of the covariates in x_{it} is continuous. This assumption avoid situations where $\mathbb{P}[x_{it} = x_{js}] > 0$ for any $(i, t) \neq (j, s)$ but can be relaxed in practice without any major practical implication. Assumption 2(ii) is analogous to Assumption 1(iv) but now expressed at the level of the components. It implies that the MCL function will be generically identifiable if each θ_j is point identified. Contrary to the mixture likelihood, overspecifying the number of components in the MCL function does not lead to any identification issue. In practice, it is possible to test for the presence of a "useless" component by testing if $\theta_g = \theta_j$ for any pair $(g, j) \in \mathbb{G} \times \mathbb{G} \backslash g$. This idea is reinforced by the fact that latent group estimators tend not to mix observations from different groups when G is overspecified (Liu et al., 2020). Assumption 2(iii) is the component-level analog of Assumption 1(vi) and is also slightly weaker than Assumption 1(vi).

Assumption 2(iv) states that x_{it} is generated by a multivariate density that varies according to the value of z_{it}^0. Assumption 2(v) implies that no covariate is perfectly collinear with any other (set of) covariate(s) within each group, and that all covariates show minimal variation within groups (apart from the group intercepts). This assumption is similar to the classical full-rank assumption in linear regression but at the true group level. It can be relaxed to allow groups to depend exclusively on one or several covariate(s) taking a specific value or values, but care must be taken in order to avoid perfect collinearity during the estimation process. The same assumption also implies that the vector x_{it} is i.i.d. conditional on the true group membership, which could be relaxed at the expense of tedious theoretical complications (see the proof of Corollary 3.3). Finally, Assumption

2(*vi*) implies that the proportion of elements that differs between each true covariance matrix Σ_g^0 does not go to zero as $p \to \infty$. Those last three assumptions are essential in order to show that the proposed estimator is consistent.

I now generally define a *binary classifier* for any observation in the sample.

Definition 1. *Let Assumptions 1(ii)-(iii) and 2 hold. Then the binary classifier $z_{itg}(\theta, \xi)$ is defined as follows*

$$z_{itg}(\theta, \xi) := \begin{cases} 1 & \text{if } g = \arg\max_{j \in \mathbb{G}} h_j(y_{it}, x_{it} | \theta_j, \xi_j) \\ 0 & \text{otherwise,} \end{cases}$$

where $h_j : \mathbb{R}^{p+1} \to \mathbb{R}$ is a categorizing function (or discriminant function) conditional on (θ_j, ξ_j) and which is continuous with respect to both θ_j and ξ_j for any $j \in \mathbb{G}$.

Using Definition 1, the bias of any binary classifier $z_{itg}(\theta, \xi)$ is defined as follows.

Definition 2. *Let Assumptions 1(ii)-(iii) and 2 hold. A binary classifier $z_{itg}(\theta, \xi)$ is said to be unbiased if and only if*

$$\arg\max_{j \in \mathbb{G}} \mathbb{E}_0[h_j(y_{it}, x_{it} | \theta_j^0, \xi_j^0)] = z_{it}^0,$$

for any pair (i, t) and any value of $g \in \mathbb{G}$, where $z_{it}^0 = \{1, ..., G\}$ is the true group membership indicator for the it^{th} observation after a suitable permutation in the labels of the groups, and where $\mathbb{E}_0$ denotes the expected value with respect to the true joint density of both y_{it} and x_{it}.

Remark 1. An unbiased binary classifier *does not* correspond to a binary classifier such that $\mathbb{E}_0[z_{itg}(\theta^0, \xi^0)] = \mathbb{E}_0[\arg\max_{j \in \mathbb{G}} h_j(y_{it}, x_{it} | \theta_j^0, \xi_j^0)] = z_{itg}^0$, which actually corresponds to a consistent classifier (see Definition 3). Instead, the bias of any classifier refers to the capacity of its discriminant function to correctly classify any observation after averaging $h_j(\cdot)$ over multiple draws.

A *consistent* classifier is then defined as follows.

Definition 3. *A consistent binary classifier is a binary classifier such that*

$$\begin{aligned} \hat{E}_{NTp}(\theta^0, \xi^0) &:= \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} \frac{\mathbb{1}[\hat{z}_{NTp,it(g)}(\theta^0, \xi^0) \neq z_{itg}^0]}{2NT}, \\ &\equiv \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\mathbb{1}[\hat{z}_{NTp,it}(\theta^0, \xi^0) \neq z_{it}^0]}{NT} \xrightarrow{p} 0, \end{aligned} \tag{10}$$

as $N, T \to \infty$, where $\hat{E}_{NTp}(\theta, \xi)$ is the misclassification rate of any binary classifier $\hat{z}_{NTp,it(g)}(\theta, \xi)$ that is based on a sample of NT observations and p covariates, and where (g) is a permutation from $g \in \mathbb{G} \to j \in \mathbb{G}$ that minimizes $\hat{E}_{NTp}(\theta, \xi)$ for given values of (θ, ξ).

Akin to Su et al. (2016), I now define a binary classifier that is *uniformly consistent*.

Definition 4. *A uniformly consistent binary classifier is a binary classifier such that*

$$\hat{E}^{u}_{NTp}(\theta^0, \xi^0) := \mathbb{P}[\cup_{i=1}^{N} \cup_{t=1}^{T} \cup_{g=1}^{G} (\hat{z}_{NTp,it(g)}(\theta^0, \xi^0) \neq z^0_{itg})] \to 0,$$

as $N, T \to \infty$*, where* $\hat{E}^{u}_{NTp}(\theta, \xi)$ *is the probability of misclassifying at least one observation in the sample,* $\hat{z}_{NTp,it(g)}(\theta, \xi)$ *is defined as in Definition 3, and where* (g) *is a permutation from* $g \in \mathbb{G} \to j \in \mathbb{G}$ *that minimizes* $\hat{E}^{u}_{NTp}(\theta, \xi)$ *for given values of* (θ, ξ)*.*

Remark 2. A uniformly consistent classifier is a classifier that groups every observation into its true group with probability approaching 1 as $N, T \to \infty$. Such a classifier is "consistent" uniformly over all possible realizations of (y_{it}, x_{it}). Note also that a uniformly consistent classifier will not necessarily yield a misclassification rate that is equal to zero in finite samples. Analogously, a misclassification rate equal to zero does not imply that the employed classifier is uniformly consistent. If the rate of convergence of a uniformly consistent classifier is (very) fast, then it should lead to a value of $\hat{E}_{NTp}(\theta, \xi)$ that is (very) close to zero when θ and ξ are close or equal to their true values.

Unlike Su et al. (2016), the latter definition does not distinguish between Type I and Type II errors since they can be controlled simultaneously in finite samples by minimizing the misclassification rate if no observation is left unclassified, which is implicit throughout the paper.

I now define three different binary classifiers : the joint density classifier, the Euclidean distance classifier, and the Mahalanobis distance classifier.

Definition 5. *Let Assumptions 1(ii)-(iii) and 2 hold, and define* $z_{itg}(\theta, \xi)$ *as in Definition 1.*
(a) The joint density classifier $z^{D}_{itg}(\theta, \xi)$ *is defined such that*

$$z^{D}_{itg}(\theta, \xi) : h_j(y_{it}, x_{it}|\theta_j, \xi_j) = f_j(y_{it}, x_{it}|\theta_j, \xi_j),$$

where $f_j(y_{it}, x_{it}|\theta_j, \xi_j) = f_j(y_{it}|x_{it}; \theta_j) f_j(x_{it}|\xi_j)$ *represents the joint density of the* it^{th} *observation as a function of* (θ_j, ξ_j)*.*
(b) The Euclidean distance classifier $z^{E}_{itg}(\boldsymbol{\mu})$ *is defined such that*

$$z^{E}_{itg}(\boldsymbol{\mu}) : h_j(y_{it}, x_{it}|\theta_j, \xi_j) = -||x_{it} - \boldsymbol{\mu}_j||^2.$$

(c) The Mahalanobis distance classifier $z^{M}_{itg}(\boldsymbol{\mu}, \Sigma)$ *is defined such that*

$$z^{M}_{itg}(\boldsymbol{\mu}, \Sigma) : h_j(y_{it}, x_{it}|\theta_j, \xi_j) = -d^2(x_{it}, \boldsymbol{\mu}_j, \Sigma_j),$$

where $d^2(x_{it}, \boldsymbol{\mu}_j, \Sigma_j) = (x_{it} - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (x_{it} - \boldsymbol{\mu}_j)$ *denotes the squared Mahalanobis distance.*

Using the joint density of the outcome and the covariates as a classifier is a natural choice when distributional assumptions are made on the joint density. This classifier is known as the Bayes' classification rule and reduces to a naive Bayes' classifier (with prior equals to $f_j(y_{it}|x_{it};\theta_j)$) when all covariance terms in $\Sigma = (\Sigma_1, ..., \Sigma_G)$ are assumed to be null (Hastie et al., 2009). The Euclidean distance classifier is a nonparametric classifier that relies on the first moment of x_{it} and is the one normally used with K-means. Finally, the Mahalanobis distance classifier is also nonparametric but relies on the first and the second moments of x_{it} to classify each observation. Several other classifiers exist and are used in practice, but those three classifiers are the more prevalent in the literature on unsupervised classification and clustering techniques.

The next two lemmas show that all three classifiers are unbiased under specific conditions.

Lemma 3.3. *Let Assumptions 1(ii)-(iii) and 2 hold. Then all three classifiers defined in Definition 5 are unbiased if* $\boldsymbol{\mu}_j^0 = \boldsymbol{\mu}_g^0 \Leftrightarrow j = g$.

Lemma 3.4. *Let Assumptions 1(ii)-(iii) and 2 hold. Then the Mahalanobis distance classifier is unbiased if p is sufficiently large and if* $\Sigma_g \neq \Sigma_j$ *for any pair* $(g, j) \in \mathbb{G} \times \mathbb{G} \backslash g$, *and for a sufficiently large proportion of the elements in each covariance matrix.*

Lemma 3.4 shows that the Mahalanobis distance classifier is unbiased even if the vectors of true mean values are identical across components. This makes this classifier more robust to cases where the component densities feature strong overlap between each other compared to the Euclidean distance classifier. Although the joint density classifier is unbiased if $\boldsymbol{\mu}_j^0 = \boldsymbol{\mu}_g^0 \Leftrightarrow j = g$, it is generally not possible to show that it is (uniformly) consistent without specifying the distribution of the joint density. On the other hand, Theorem 3.2 and Corollary 3.3 show that the Mahalanobis distance classifier is uniformly consistent under Assumption 2 provided that the number of covariates grows at a faster rate than the number of observations for a fixed value of G.

Theorem 3.2. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Mahalanobis distance classifier as in Definition 5(c) with a fixed number of groups G. Then*

$$\mathbb{P}[\cup_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) \neq z_{itg}^0)] = O_p(p^{-1}),$$

for any pair (i, t), *where* (g) *corresponds to a suitable permutation in the labels of the groups.*

Theorem 3.3. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Mahalanobis distance classifier as in Definition 5(c). Then* $z_{itg}^{M}(\boldsymbol{\mu}^0, \Sigma^0)$ *is a uniformly consistent classifier if the number of covariates p increases at a strictly higher-order rate than the rate at which the number of groups G increases relative to the sample size.*

Corollary 3.3. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Mahalanobis distance classifier as in Definition 5(c) with a fixed number of groups G. Then* $z_{itg}^{M}(\boldsymbol{\mu}^0, \Sigma^0)$ *is a uniformly consistent classifier if* $p/NT \to \infty$ *as* $N, T, p \to \infty$.

Corollary 3.4. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Mahalanobis distance classifier as in Definition 5(c). Then* $z^M_{itg}(\boldsymbol{\mu}^0, \Sigma^0)$ *is a consistent classifier if* $p/G \to \infty$ *as* $N, T, p, G \to \infty$.

Corollary 3.3 implies that the Mahalanobis distance classifier may not be uniformly consistent if the number of covariates increases at a slower rate than the sample size. This is problematic since the estimation of each θ_g cannot be performed using MLE procedures or ordinary least squares (OLS) in this case. On the other hand, Corollary 3.4 shows that the misclassification rate given by the Mahalanobis distance classifier will go to zero as $p \to \infty$ if the number of covariates grows at a strictly faster rate than the number of groups, regardless from the rate at which the sample size increases. As emphasized by Dzemski and Okui (2021), uniform consistency of a classifier is not necessary to obtain consistent estimates of θ and ξ when estimating latent group panel structures (or finite mixtures). Such a statement is confirmed below by Theorem 3.5. Note also that (strict) exogeneity of x_{it} is not necessary for the last two theorems and the last two corollaries to be true.

Contrary to the Mahalanobis distance classifier, the next theorem and corollary show that the Euclidean distance classifier will always misclassify a non-null proportion of the observations in the sample as $N, T, p \to \infty$ if the distance between each vector $\boldsymbol{\mu}^0_j$ does not become sufficiently large as p tends to infinity. Consequently, the Euclidean distance classifier cannot be consistent in general, which is shown below in Corollary 3.5. This conclusion reinforces the idea according to which the K-means algorithm does not converge to the true values if no restriction is imposed on group memberships over (at least) one dimension of the dataset (Pollard, 1981; Bryant, 1991).

Theorem 3.4. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Euclidean distance classifier as in Definition 5(b) with a fixed number of groups G. Then*

$$\mathbb{P}[\cup_{g=1}^{G}(z^E_{it(g)}(\boldsymbol{\mu}^0) \neq z^0_{itg})] \xrightarrow{p} c_{it} \text{ as } p \to \infty$$

for any pair (i,t)*, where* $c_{it} \geq 0$ *is random.*

Corollary 3.5. *Let Assumptions 1(ii)-(iii) and 2 hold, and define the Euclidean distance classifier as in Definition 5(b) with a fixed number of groups G. Then* $z^E_{itg}(\boldsymbol{\mu}^0)$ *is not a consistent classifier unless the ratio* $\frac{||\boldsymbol{\mu}^0_j - \boldsymbol{\mu}^0_g||^2}{tr(\Sigma^0_g)} \to \infty$ *as* $p \to \infty$ *for any pair* $(g,j) \in \mathbb{G} \times \mathbb{G} \backslash g$*, where* $tr(\cdot)$ *denotes the trace operator.*

Remark 3. When all covariates in x_{it} are normally distributed, the joint density classifier can be expressed as a function of the Mahalanobis distance given that the squared Mahalanobis distance is embedded into the exponential part of the multivariate normal density. In this case, the joint density classifier is both unbiased and consistent under the same conditions as the ones stated in Lemma 3.4 and Corollary 3.4. Simulations confirm that the joint density classifier leads to a smaller misclassification rate in finite samples than the Mahalanobis distance classifier when the normality assumption is satisfied (results not shown).

Algorithm 1 The "consistent" CEM algorithm

Let $z_{itg}(\theta, \xi)$ be any binary classifier that is *consistent*, as described in Definition 3, as $N, T, p \to \infty$. Given initial/previous values $(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$ (the NT subscript is dropped for brevity), the "consistent" CEM algorithm consists of the consecutive repetition of the two following steps until convergence :

1. The Classification Step (the CE-step) :
   Compute $z_{itg}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$ for all pairs (i, t) and all values of $g \in \mathbb{G}$.
2. The Maximization Step (the M-step) :
   (a) Compute $\hat{\theta}^{(k+1)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} z_{itg}(\theta^{(k)}, \xi^{(k)}) \log(f_g(y_{it}|x_{it}, \theta))$.
   (b) Compute $\hat{\xi}^{(k+1)} = \arg\max_{\xi \in \Xi} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} z_{itg}(\theta^{(k)}, \xi^{(k)}) \log(f_g(x_{it}|\xi))$ or estimate consistently $\hat{\boldsymbol{\mu}}^{(k+1)}$ and $\hat{\Sigma}^{(k+1)}$ conditional on $z_{itg}(\hat{\boldsymbol{\mu}}^{(k)}, \hat{\Sigma}^{(k)})$ depending on the chosen classifier.

If the normality assumption does not hold, the Mahalanobis distance classifier offers a very good nonparametric alternative to the joint density classifier. Extending the use of the Mahalanobis distance classifier to a consistent, fully nonparametric estimation strategy is straightforward via the use of a kernel regression estimator for the outcome at each M-step of the "consistent" CEM algorithm (see Algorithm 1). Quotation marks are used to name the algorithm since it is not different from the original CEM algorithm, except that the employed classifier is consistent. This is why, from now on, mention of the CEM algorithm will always refer to its consistent version, when possible.

We can now derive the asymptotic distribution of the estimates given by Algorithm 1 under the following assumptions.

Assumption 3.

(i) $z_{itg}(\theta^0, \xi^0)$ *is a classifier that is unbiased and consistent as* $N, T, p \to \infty$.

(ii) $\operatorname{plim}_{N,T\to\infty} \arg\max_{\theta \in \Theta} l^{MC}(\theta, \mathbf{z}^0) = \theta^0$.

(iii) $\lim_{N,T\to\infty} n_g^0 = \infty$ *for all* $g \in \mathbb{G}$, *where* $n_g^0 = \sum_{i=1}^{N} \sum_{t=1}^{T} z_{itg}^0$.

(iv) Each component density $f_g(y_{it}|x_{it}; \theta)$ *is twice continuously differentiable with respect to* θ *for all pairs* $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$.

(v) $\mathcal{I}_g := \mathbb{E}_0 \left[\left\{ s_{ig}(\theta^0, \xi^0) \right\} \left\{ s_{ig}(\theta^0, \xi^0) \right\}^\top \right] < \infty$, *where* $s_{ig}(\theta, \xi) = \sum_{t=1}^{T} z_{itg}(\theta, \xi) \frac{\partial}{\partial \theta_g} \log f_g(y_{it}|x_{it}; \theta_g)$, *with* $\bar{n}_g^{-1}(\theta^0, \xi^0) \sum_{i=1}^{N} \left\{ s_{ig}(\theta^0, \xi^0) \right\} \left\{ s_{ig}(\theta^0, \xi^0) \right\}^\top \xrightarrow{p} \mathcal{I}_g$ *and* $\bar{n}_g(\theta, \xi) = \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{itg}(\theta, \xi)$ *for each* $g \in \mathbb{G}$, *and where all observations across units and across groups are assumed to be independent from each other.*

(vi) $\mathcal{H}(\theta_g) := \mathbb{E}_0[\frac{\partial}{\partial \theta_g} s_{ig}(\theta, \xi)]$ *is the expected Hessian matrix of the* g^{th} *group, where* $\mathcal{H}_g \equiv \mathcal{H}(\theta_g^0)$ *is finite and non-singular for all* $g \in \mathbb{G}$, *and where* $\bar{n}_g^{-1}(\theta, \xi) \sum_{i=1}^{N} \frac{\partial}{\partial \theta_g} s_{ig}(\theta, \xi) \xrightarrow{p} \mathcal{H}(\theta_g)$ *uniformly in* θ_g *in an open neighbourhood of* θ_g^0 *for all* $g \in \mathbb{G}$ *as* $N, T \to \infty$.

Assumption 3(i) and Assumption 3(ii) are necessary to ensure that the algorithm leads to consistent estimates of θ (and ξ). Those assumptions can be adapted to the case of fixed-T asymptotics by using a bias-corrected estimator if the component densities are plagued by the incidental parameter problem (Hahn and Newey, 2004). Assumption 3(iii) implies that the number of observations within each group goes to infinity as the sample size increases. Assumption 3(iv) is standard in most MLE problems and is necessary to compute the Hessian matrix for each group. Finally, Assumption 3(v) and Assumption 3(vi) respectively describe the information and Hessian matrices of the estimator for each $g \in \mathbb{G}$. Note that the score function $s_{ig}(\theta, \xi)$ is computed at the unit level to account for serial correlation within each unit-group. Independence of the same unit across groups is assumed for simplicity since cross-group correlation would require the estimation of additional covariance parameters.

The next two theorems follow from Assumption 3 and Algorithm 1, and can be adapted to cross-sections by modifying the score function appropriately. Theorem 3.5 shows that the CEM algorithm will yield consistent estimates of all parameters in the mixture after a finite number of iterations provided appropriate initial parameter values, and a sufficiently large sample size and number of covariates. If those conditions are satisfied, Theorem 3.6 shows that each estimated component parameters $\hat{\theta}_g^{(k)}$ will be normally distributed asymptotically.

Theorem 3.5. *Let Assumptions 1(ii)-(iii), 2 and 3 hold. Given a fixed number of groups G, initial values $(\theta^{(0)}, \xi^{(0)})$ sufficiently close to (θ^0, ξ^0), and values of (N, T, p) sufficiently large, the CEM algorithm equipped with a consistent classifier will converge to $\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)}$ and $\hat{\xi}^{(k+1)} = \hat{\xi}^{(k)}$ such that $(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \xrightarrow{p} (\theta^0, \xi^0)$ as $N, T, p \to \infty$, where k is a positive, discrete, finite number.*

Theorem 3.6. *Let Assumptions 1(ii)-(iii), 2 and 3 hold, and let $(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \xrightarrow{p} (\theta^0, \xi^0)$ as $N, T, p \to \infty$ as described in Theorem 3.5. Then*

$$\sqrt{n_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})}(\hat{\theta}_g^{(k)} - \theta_g^0) \xrightarrow{d} \mathcal{N}(0, \mathcal{H}_g^{-1}\mathcal{I}_g\mathcal{H}_g^{-1}),$$

as $N, T, p \to \infty$, where $n_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) = \sum_{i=1}^{N}\sum_{t=1}^{T} z_{itg}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$.

Remark 4. The convergence of $\hat{\theta}^{(k)}$ is not uniform over π^0 since arbitrarily small values of π_g^0 can make the convergence of $\hat{\theta}_g^{(k)}$ arbitrarily slow. This is why all convergence results presented here are "groupwise" under Assumption 3(iii). Note also that the estimates provided by the CEM algorithm will not be efficient if the employed binary classifier is not uniformly consistent. Developing an inference procedure that would be uniform over π^0 as well as developing a binary classifier that would be uniformly consistent and reach the semiparametric efficiency bound under weaker conditions than the ones presented in Theorem 3.3 are left to future research.

As in standard MLE problems, misspecification of the component densities will make the CEM algorithm converges to a pseudo-true value $\bar{\theta} \neq \theta^0$ that minimizes the Kullback-Leibler divergence

between the "working" and the true component densities (Gourieroux et al., 1984). However, the Mahalanobis distance classifier remains consistent under Assumption 1(ii)-(iii) and Assumption 2 even if the component densities are misspecified. This is one of the main benefits of using the Mahalanobis distance classifier compared to the other ones presented in Definition 5.

Remark 5. Under misspecification of the component densities and/or lack of efficiency, estimation of $\text{Var}[\hat{\theta}_g^{(k)}]$ can be done using the sample counterpart of the asymptotic variance given in Theorem 3.6. This estimated variance is defined as follows

$$\text{Var}[\hat{\theta}_g^{(k)}] = \left\{ s'_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \right\}^{-1} \sum_{i=1}^{N} \left[\left\{ s_{ig}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \right\} \left\{ s_{ig}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \right\}^{\top} \right] \left\{ s'_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) \right\}^{-1}, \quad (11)$$

where $s'_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)}) = \sum_{i=1}^{N} \frac{\partial}{\partial \hat{\theta}_g^{(k)}} s_{ig}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$, and where $\hat{\theta}^{(k)}$ and $\hat{\xi}^{(k)}$ are the parameters that maximize the MCL function after searching for the global maximum of the objective function. Note also that the expressions for both the asymptotic variance and its estimated counterpart will not account for the uncertainty in group memberships. This uncertainty can be taken into account by the use of resampling techniques, such as the bootstrap, or by constructing confidence sets that account for group memberships (Dzemski and Okui, 2024; Higgins and Jochmans, 2022). Implementing such methods goes however beyond the scope of this paper.

The next section confirms the theoretical results obtained in this section, which are summarized as follows : standard MLE of finite mixtures cannot yield consistent estimates of any parameter in the mixture unless one relies on a consistent classification procedure whose misclassification rate will go to or be equal to zero as the sample size and the number of covariates tend to infinity.

4 Monte Carlo Simulations

4.1 Data-generating Processes and Estimation Strategies

In this section, I describe the objectives and details of two distinct simulation exercises. The main objective of the first simulation exercise is to confirm the conclusion of Corollary 3.2. The main objective of the second simulation exercise is to compare the finite-sample performance of the EM and the CEM algorithms using a mixture of linear panel data.

4.1.1 Simple Gaussian Mixture Model with no Covariate

This first simulation exercise uses a simple Gaussian mixture of two and three components with different means and equal variances. The maximization of the mixture likelihood and the max-component likelihood functions is performed via the EM and the CEM algorithms respectively for each scenario. Given that this simulation exercise does not employ any covariate, the CEM algorithm cannot yield consistent estimates, just as the EM algorithm.

The data-generating process (DGP) of the current simulation exercise is represented by the following equation

$$y_i = \mu^0_{z_i^0} + \epsilon_i, \tag{12}$$

with $\boldsymbol{\mu}^0 = (\mu^0_1, ..., \mu^0_G)$ denoting the vector of true mean values, and where $\epsilon_i \sim N(0, 1)$ for all $i \in \{1, ..., N\}$. The following steps explain how the simulation exercise was conducted.

1. 500 observations were drawn from each component of the specified mixture.
2. 500 random initial mean values were drawn from a centered normal distribution with a relatively large variance ($\approxeq 100$). Initial values for σ^2_ϵ and π^0 were set to their true values for simplicity.
3. Both the EM and CEM algorithms were performed on the simulated dataset using each set of initial values (with equations (7) and (8) when using the EM).
4. The estimates associated with the highest log likelihood value among the 500 runs for each algorithm were selected (for all parameters, i.e. $\hat{\boldsymbol{\mu}}$, $\hat{\sigma}^2_\epsilon$ and $\hat{\pi}$).
5. The initial sample size was increased by a factor $r = \{2, 10, 20, 50, 200, 1000\}$ in increasing order while keeping the same random generation seed.
6. Steps 3 to 6 were repeated until $r = 1000$ with the selected estimates in step 4 as updated initial values.

The first three steps of the above procedure help to find a local optimum in small samples that is close to the global maximum of the corresponding likelihood function. Once this local optimum is found, increasing the sample size should not fundamentally change the structure of this maximum, so that the estimates associated with the previously identified maximum can be used as starting values for the next, larger dataset.

If globally maximizing the mixture likelihood yields consistent estimates of the mean parameters, the distance between the estimated parameters and the true parameter values should diminish as N increases. The value of the objective function evaluated at the true parameter values should also become larger than when it is evaluated at any other point as N increases if the estimator is consistent. If the value of the objective function evaluated at the true parameter values remains lower than the value of the same objective but evaluated at another point that does not converge to the true parameter value as N increases, then such a situation is indicative of an inconsistent estimation strategy. Therefore, such a simulation setup can indicate under which conditions the MLE will yield (in)consistent estimates of the true parameter values.

To bind the likelihood function from above, a minimum variance of 0.01 was enforced for all components. The true variance is equal to one for all components while the mean values vary across simulation scenarios. The maximum number of iterations for each run of the CEM/EM algorithm

is set to 100,000, and convergence is reached when the change in the average log likelihood value is smaller than 1E-10 between two consecutive iterations.

4.1.2 Latent Group Linear Panel Structure

I consider here the typical case in latent panel structure where the conditional density of the outcome value follows a normal distribution for all values of $g \in \mathbb{G}$. The main goal of this simulation exercise is to determine which one of the two algorithms, the EM or the CEM algorithm, yields the best finite-sample results, in terms of estimation biases, when looking at the estimates that provide the highest log likelihood value among randomly chosen initial values. The exact DGP for the outcome variable used for this simulation exercise is described as follows

$$\begin{aligned} y_{it} &= x_{it}^{\top}\beta_{z_{it}^0} + \bar{x}_i^{\top}\gamma_{z_{it}^0} + \delta_{tz_{it}^0} + \alpha_{iz_{it}^0} + \epsilon_{it}, \\ &= X_{it}\tilde{\beta}_{z_{it}^0} + \alpha_{iz_{it}^0} + \epsilon_{it}, \end{aligned} \tag{13}$$

where z_{it}^0 is generated by a categorical distribution where the vector of probabilities follows an AR(1) process, $\tilde{\beta}_{z_{it}^0} = (\beta_{z_{it}^0}^{\top}, \gamma_{z_{it}^0}^{\top}, \delta_{1z_{it}^0}, ..., \delta_{Tz_{it}^0})^{\top}$, $X_{it} = (x_{it}^{\top}, \bar{x}_i^{\top}, \mathbb{1}[t=1], ..., \mathbb{1}[t=T])$, $\bar{x}_i^{\top}$ is a p-sized row-vector of time-averaged covariates' values, $\beta_{z_{it}^0}$ and $\gamma_{z_{it}^0}$ are both p-sized column-vectors of parameters to be estimated for each value of z_{it}^0, $\alpha_{iz_{it}^0} \sim N(0, \sigma^2_{\alpha,z_{it}^0})$ where $\sigma^2_{\alpha,z_{it}^0} = z_{it}^0 \in \mathbb{G}$, and where $\delta_{tz_{it}^0}$ is a time-fixed effect that varies arbitrarily with the value of z_{it}^0. For simplicity, we have that $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 1$, and also $\mathbb{E}[\epsilon_{it}\epsilon_{ls}] = 0$ for any pair $(l,s) \neq (i,t)$.

The vector of covariate x_{it} is generated according to

$$x_{it} \sim N_p(\boldsymbol{\mu}_{z_{it}^0}^0, \Sigma_{z_{it}^0}^0), \tag{14}$$

where N_p denotes the p-variate normal distribution, and where the ij^{th} entry of Σ_g^0 is equal to

$$\sigma_{g,lj} = (a \times g)^{|l-j|}, \text{ for any pair } (l,j) \in \{1,...,p\} \times \{1,...,p\},$$

with $a = 0.16$ such that all elements in the main diagonal of each Σ_g^0 is equal to one and all covariances are bounded from above. Each element in the p-sized vector $\boldsymbol{\mu}_g^0$ is drawn from a normal distribution with unit variance and centered at $\eta_g^0 \in [-4,4]$ to guarantee that $\boldsymbol{\mu}_g^0 \neq \boldsymbol{\mu}_l^0$ for any $l \neq g$. The number of covariates, p, and the true parameter values are modified in order to set the misclassification rate, $\hat{E}_{NTp}(\theta^0, \xi^0)$ (where $\theta^0 = (\tilde{\beta}, \sigma_\alpha^2, \sigma_\epsilon^2)$ and $\xi^0 = (\boldsymbol{\mu}^0, \Sigma^0)$), at the desired level for each simulation scenario.

The specification described in equation (13) corresponds to the Mundlak specification where the time-average values of the covariates plus unit-random effects are used in place of unit-specific fixed effects. It has been shown that it is equivalent to the *least-square dummy variable* (LSDV) estimator (Yang, 2022) but with the advantage of being more computationally tractable since it

greatly reduces the number of parameters to estimate. Moreover, using the Mundlak approach allows the non-random part of the unit-fixed effects, $\bar{x}_i^\top \gamma_{z_{it}^0}$, to vary across groups for the same individual. Such a feature is also feasible with the LSDV estimator, but each unit would have to remain in the same group for at least two periods to avoid identification issues. The Mundlak specification also facilitates cross-validation since the non-random part of the unit-fixed effect can be predicted for any unit in the test set. This framework also avoids the biases generated by the incidental parameter problem encountered in nonlinear panels with fixed effects.

Estimation is carried out using both the EM and the CEM algorithms, where the EM algorithm uses equations (7) and (8) at each E-step, whereas the CEM algorithm uses the joint density classifier as shown in Definition 5. All models are correctly specified, including the use of normal distributions for the random-unit effects α_{ij}, the error term ϵ_{it}, and the density of x_{it}. All simulation scenarios use the same set of 1,000 different, randomly generated, initial values $(\hat{\theta}^{(0)}, \hat{\xi}^{(0)})$ to assess the sensitivity of each algorithm to the same set of initial parameter values. More details regarding the two estimation procedures for this second simulation exercise are given in Appendix B.

To ensure the boundedness of all likelihood functions, minimum and maximum variance values of respectively 0.001 and 1000 are enforced for all diagonal elements in all covariance matrices. Non-singularity of $\hat{\Omega}_g^{(k)}$ and $\hat{\Sigma}_g^{(k)}$ is obtained by ensuring that all eigenvalues remain above 0.001 for all $g \in \mathbb{G}$ at each iteration. The maximum number of iterations for all simulation runs is set to 100. Convergence of both algorithms is assumed to be reached when the relative change between two consecutive log likelihood values is less than 0.01%. Contrary to more traditional simulation exercises, the dataset does not change across the thousand sets of initial values; only the initial values are allowed to vary to find the global maximum of the corresponding objective function. The simulated dataset however changes across simulation scenarios.

4.2 Simulation Results

4.2.1 Simple Gaussian Mixture Model

Table 1 shows the results of the first simulation exercise when $G = 2$ and when $\pi^0 = (0.5, 0.5)$. The first column of Table 1 shows the true mean values, $\boldsymbol{\mu^0}$, used to generate the dataset. The second column of Table 1 shows the asymptotic misclassification rate at the true parameter values, $E(\theta^0, \pi^0)$, while the sample size N is shown in column (3). The fourth column shows the distance between the mixture log likelihood function when evaluated at the estimated parameter values and at the true parameter values. The fifth column shows the root mean square error (RMSE) of the estimated mean values $\hat{\boldsymbol{\mu}}$ for each scenario. An increasing, positive distance $l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$ observed along an increasing sample size and RMSE is indicative of the inconsistency of the estimation procedure. If the estimates are consistent, then $l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$ should be negative and decreases as both the RMSE and N rise. Finally, the sixth and seventh columns of Table 1 are the

$\boldsymbol{\mu}^0$	$E(\theta^0, \pi^0)$ (%)	N	$l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$	RMSE, EM	$l^{MC}(\hat{\theta}, \hat{\pi}) - l^C(\theta^0, \pi^0)$	RMSE, CEM
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(-0.25, 0.25)	**40.1**	2,000	2.573	1.036	953.1	0.599
		10,000	5.217	0.994	4776.8	0.582
		20,000	0.691	1.060	9471.9	0.573
		50,000	2.234	0.739	23942.6	0.577
		200,000	1.612	0.351	95070.9	0.575
		1,000,000	-0.498	0.435	475990.9	0.572
(-0.5, 0.5)	**30.9**	2,000	1.917	0.411	810.6	0.424
		10,000	3.024	0.309	4045.6	0.403
		20,000	2.199	0.274	7974.6	0.396
		50,000	2.269	0.310	20313.3	0.400
		200,000	1.559	0.160	80164.5	0.398
		1,000,000	-1.504	0.103	400923.9	0.396
(-1, 1)	**15.9**	2,000	1.424	0.023	474.9	0.184
		10,000	3.427	0.010	2298.3	0.174
		20,000	4.243	0.089	4377.7	0.169
		50,000	1.927	0.028	11423.2	0.168
		200,000	1.968	0.022	44902.4	0.169
		1,000,000	0.737	0.003	223462.2	0.167
(-2, 2)	**2.3**	2,000	2.681	0.014	80.5	0.033
		10,000	4.781	0.006	343.7	0.015
		20,000	3.148	0.017	634.4	0.027
		50,000	3.273	0.008	1842.1	0.016
		200,000	1.976	0.001	7017.0	0.019
		1,000,000	0.427	0.000	34913.3	0.017

Table 1: Root mean square errors (RMSEs) of the estimated mean values and differences in log likelihood values with $G = 2$ when $\pi^0 = (0.5, 0.5)$; the true variances are all equal to one.

analogs of columns (4) and (5), but applied to the standard, inconsistent CEM algorithm.

The fourth and fifth columns of Table 1 show that the RMSE of the mean values obtained with the EM algorithm mostly decreases as the sample size increases, except in some cases where it remains constant or slightly increases with N. This is the case, for instance, when $\boldsymbol{\mu}^0 = (-0.5, 0.5)$ and N goes from 20,000 to 50,000, where both $l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$ and the RMSE associated with the EM algorithm increase as N increases. A similar case is also observed when $\boldsymbol{\mu}^0 = (-1, 1)$ and N goes from 50,000 to 200,000. Such situations indicate that the estimation procedure is inconsistent when $E(\theta^0, \pi^0)$ is too high. On the other hand, $l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$ becomes negative when $N = 1,000,000$ in the first two scenarios. This does not invalidate the inconsistency hypothesis since it can be caused by the presence of a local maximum near the global maximum. Nonetheless, the general

Algorithm (1)	$\boldsymbol{\mu}^0$ (2)	$\hat{\mu}_1$ (3)	$\hat{\mu}_2$ (4)	$\hat{\sigma}_1$ (5)	$\hat{\sigma}_2$ (6)	$\hat{\pi}_1$ (7)	$\hat{\pi}_2$ (8)
EM	**(-0.25, 0.25)**	-0.018	0.819	1.026	0.905	0.979	0.021
	(-0.5, 0.5)	-0.399	0.605	1.019	0.984	0.603	0.397
	(-1, 1)	-1.002	0.996	1.000	1.003	0.498	0.502
	(-2, 2)	-2.000	2.001	1.001	1.001	0.500	0.500
CEM	**(-0.25, 0.25)**	-0.823	0.822	0.621	0.621	0.500	0.500
	(-0.5, 0.5)	-0.895	0.897	0.670	0.669	0.501	0.499
	(-1, 1)	-1.169	1.165	0.799	0.801	0.499	0.501
	(-2, 2)	-2.018	2.017	0.965	0.966	0.500	0.500

Table 2: Estimated values for each scenario with $G = 2$, when $\pi^0 = (0.5, 0.5)$ and $N = 1,000,000$. The estimated mixing weights do not always sum to one due to rounding.

evolution of the RMSE in the fifth column shows that the estimation bias is substantially larger in cases where the overall degree of overlap between the component densities is larger.

Even though those results do not completely confirm that the whole estimation procedure is inconsistent, it directly questions this aspect, especially when the degree of overlap between the component densities is large. The last rows of Table 1 also show that maximizing the likelihood of a mixture of distant component densities will yield good approximations of the true parameter values when N is large, thus strengthening the notions conveyed by Corollary 3.2 and Figure 1.

The sixth and seventh columns of Table 1 also show that the standard CEM algorithm yields inconsistent estimates when the mean values are too close to each other. The large positive distances observed in column (6) are explained by the fact that the max-component log likelihood is computed only with the component that maximizes the density value of the observation. As the overlap between the components shrinks, the distance between the two functions shrinks as well, implying that the standard CEM algorithm will yield consistent estimates if the component densities get infinitely distant from each other, just as the EM. Note also that the RMSE values provided by the CEM algorithm are much more stable than those provided by the EM, and that the CEM algorithm yields less biased estimates than the EM algorithm in small samples when the components are very close to each other (first four rows of Table 1).

Table 2 shows the estimated parameters associated with the results of Table 1 for both algorithms when $N = 1,000,000$. It shows that the mean values provided by the EM algorithm get more distant from their true values as the component densities are closer to each other. However, the estimates shown in Table 2 might correspond to a local optimum of the objective function since the associated distance $l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$ is negative at $N = 1,000,000$ for the first two scenarios. Notwithstanding this possibility, the estimated mean values for the first two scenarios are in line with the intuition given in panel (a) of Figure 1. Contrary to the EM algorithm, the mixing weights

estimated by the CEM algorithm are stable and very close to the true values. Overall, results from Table 2 confirm the higher stability of the CEM algorithm while showing the inconsistency of the "approximate MLE" of the mixing weights when the true densities are too close to each other.

Additional simulation results can be found in Appendix C.1, where the results are shown for different values of G, different mixing weights, and different true mean values. Similar conclusions as those drawn from Tables 1 and 2 can be drawn from the tables in Appendix C.1, but with more examples confirming the inconsistency of standard MLE of finite mixtures, especially when $G = 3$.

4.2.2 Latent Group Linear Panel Structure

The upper and lower graphs of Figure 2 respectively show the weighted RMSEs associated with the estimated conditional mean parameters $\hat{\tilde{\beta}}$ and the estimated variance parameters $(\hat{\omega}_{\alpha+\epsilon}^{2(k)}, \hat{\omega}_{\alpha}^{2(k)})$ when $\hat{E}_{NTp}(\theta^0, \xi^0)$ is equal or close to zero (between 0.00% and 0.05%). The estimation errors shown in Figure 2 are based on the estimates that maximize the log likelihood function associated with each algorithm. The weights used to compute the weighted RMSEs are the true mixing weights π_g^0 after a suitable permutation in the labels of the groups. The weighted RMSEs associated with $\hat{\xi}^{(k)} = (\hat{\boldsymbol{\mu}}^{(k)}, \hat{\Sigma}^{(k)})$ are not shown since they do not represent parameters of interest and are useful only to reduce the misclassification rate $\hat{E}_{NTp}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$. Note that the vector of mixing weights based on the true group membership varies across simulation scenarios, but all "true" mixing weights lie between 0.234 and 0.504 for all simulation scenarios.[7]

The weighted RMSEs shown in Figure 2 illustrate that the CEM algorithm always features lower estimation errors than the EM algorithm for both the mean and variance parameters, except for the mean parameters when $T = 4$. Nonetheless, the upper graphs of Figure 2 clearly show that the estimation errors of the mean and variance parameters obtained from the EM algorithm generally *increase* as T goes up, which is the opposite of what we would expect from a consistent estimator. On the contrary, the estimation errors of the mean and variance parameters obtained from the CEM algorithm tend to decrease as both N and/or T increase, which clearly illustrates the usefulness of the proposed estimator. Note also that the weighted RMSEs for the mean parameters averaged over all sets of results (one for each set of initial parameter values; excluding those who did not converge) were much lower for the CEM algorithm than for the EM algorithm (results not shown). This strengthens the idea that the CEM algorithm provides results that are less sensitive to initial parameter values than those provided by the EM algorithm.

Table 3 presents the misclassification rates for each scenario. Results from Table 3 show that the estimated misclassification rate $\hat{E}_{NTp}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$ equals the "true" misclassification rate $\hat{E}_{NTp}(\theta^0, \xi^0)$ in all scenarios when the CEM algorithm is used. Unlike the EM algorithm, the CEM algorithm can correctly classify all observations when $\hat{E}_{NTp}(\theta^0, \xi^0) = 0$ even if the AR(1) nature of z_{it}^0 had

[7] By "true" mixing weights, I am not referring to the true population weights, but rather to their estimated counterparts when all group memberships are known.

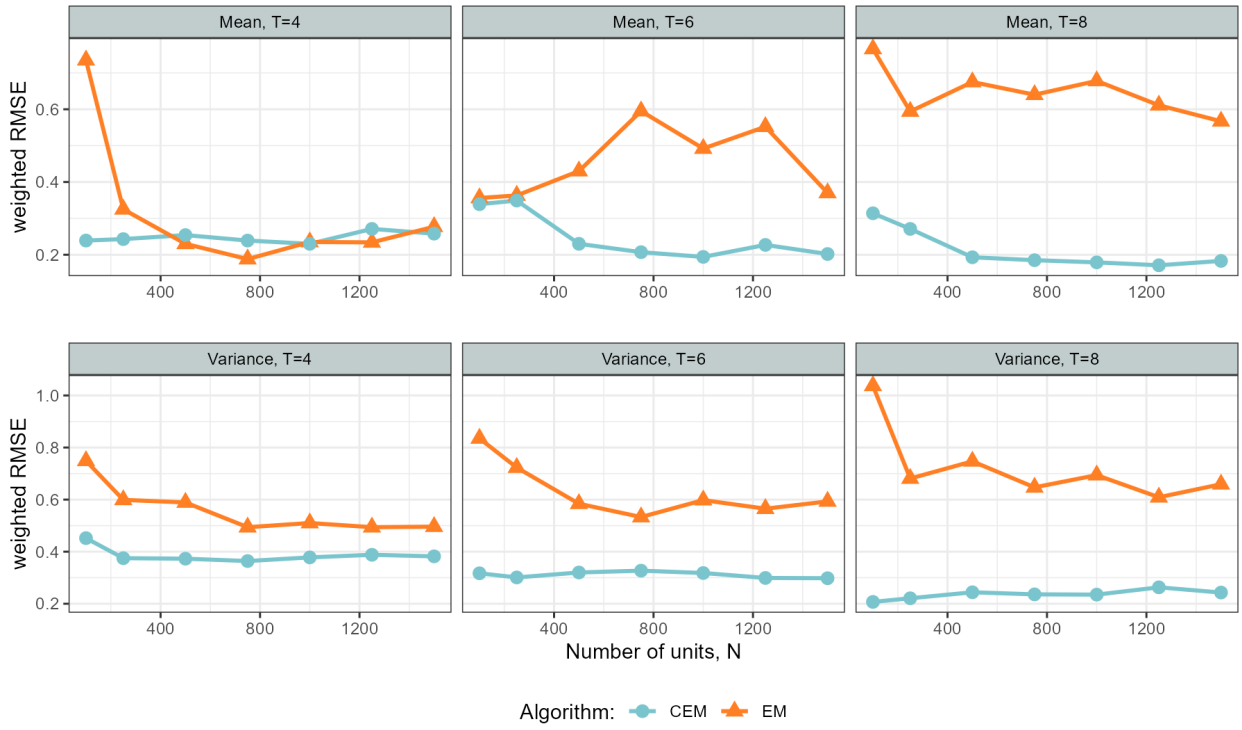


Figure 2: Evolution of the estimation error when $G = 3$ and when $\hat{E}_{NTp}(\theta^0, \xi^0) = [0.00\%, 0.05\%]$ with $p = 4$. The estimates selected to compute the weighted RMSEs are the ones that maximize the log likelihood function associated with each algorithm. The y-axis stands as the weighted RMSE for each total number of periods T, and each type of parameter (mean and variance). The "true" mixing weights all lie between 0.234 and 0.504.

not been taken into account for classification. It also shows that small deviations of $\hat{E}_{NTp}(\theta^0, \xi^0)$ away from zero should not substantially increase the estimated misclassification rates when using the CEM algorithm.

If the general form of the joint density $f(y_{it}, x_{it}|\theta, \xi)$ is known, the joint density classifier leads to good classification performance in finite samples, as shown in Table 3. Relying on the Mahalanobis distance classifier in the context of the current simulation exercise leads to higher values of $\hat{E}_{NTp}(\theta^0, \xi^0)$ for all scenarios (results not shown). This is not surprising given that, when $p = 4$, a large share of the performance of the joint density classifier is explained by the j^{th} outcome density $f_j(y_{it}|x_{it}; \theta_j)$. A hybrid classifier using both the Mahalanobis distance $d^2(x_{it}, \boldsymbol{\mu}_j, \Sigma_j)$ and the outcome density $f_j(y_{it}|x_{it}; \theta_j)$ for each $j \in \mathbb{G}$ could still lead to good classification performance if one is willing to make distributional assumptions for the outcome only. Exploring the performance of such a hybrid classifier goes however beyond the scope of this paper.

Columns (3), (5), and (7) of Table 3 show that the misclassification rates obtained with the EM algorithm generally *increase* with T, which explains the results depicted in Figure 2. This is a consequence of the inconsistency of the estimation procedure : as the sample size increases, the estimates remain biased but are estimated more precisely, which leads to an increase in the

N	$\hat{E}(\cdot)$	$T=4$		$T=6$		$T=8$	
		EM (%)	CEM (%)	EM (%)	CEM (%)	EM (%)	CEM (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
100	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	5.75	0.00	11.33	0.00	12.75	0.00
250	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	7.50	0.00	10.33	0.00	10.60	0.00
500	$\hat{E}(\theta^0, \xi^0)$	0.05	0.05	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	8.10	0.05	9.43	0.00	11.50	0.00
750	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	6.70	0.00	9.36	0.00	10.20	0.00
1000	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	6.78	0.00	9.35	0.00	11.10	0.00
1250	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	8.32	0.00	10.05	0.00	11.20	0.00
1500	$\hat{E}(\theta^0, \xi^0)$	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	7.67	0.00	10.26	0.00	10.94	0.00

Table 3: Misclassification rates evaluated at the true parameter values and evaluated at the parameter values that maximize the log likelihood function for each algorithm when $\hat{E}_{NTp}(\theta^0, \xi^0) = [0.00\%, 0.05\%]$. The NTp subscripts are dropped for brevity. The misclassification rates obtained with the EM algorithm are computed using the maximum posterior probability.

misclassification rates, which leads to more biased estimates, and so on. Those results confirm the idea that maximizing the mixture likelihood leads to an asymptotic bias that will not vanish as the sample size increases if the component densities are not infinitely distant from each other.

Two additional sets of simulation results are shown in Appendix C.2 where $\hat{E}_{NTp}(\theta^0, \xi^0) > 0$. Those results show that no algorithm is superior to the other when $\hat{E}_{NTp}(\theta^0, \xi^0)$ is larger than 4.0% and the mixing weights are well-balanced. While Figures 7 and 8 tend to show that the EM algorithm produces less biased mean estimates than the CEM algorithm, this is generally the opposite for the variance parameters. Those figures also confirm that the EM algorithm produces more unstable estimates than the CEM algorithm, especially when looking at the variance parameters.

Note that the weighted RMSEs portrayed in the graphs of Appendix C.2 are almost always higher than their analogs from Figure 2. This advocates for the use of the CEM algorithm when the number of covariates is sufficiently large such that it is reasonable to assume that $\hat{E}_{NTp}(\theta^0, \xi^0) = 0$. If it is not reasonable to make such an assumption, then additional covariates should be included in the model to reduce classification error. Other simulation results with binary choice models, unbalanced mixing weights, and misspecification of G are available from the author upon request.

5 Empirical Analysis

5.1 Objectives

The main objective of the empirical analysis is to group within the same component all the observations that share the same latent, unobserved individual characteristics. In this context, it is assumed that the unobserved characteristics explain a non-null proportion of the observed variation in individual healthcare expenditures (HCE), and that the unobserved characteristics can also influence the relationships between the outcome and the observed individual characteristics. For instance, individuals who are more frail than others, which is unobserved, are more likely to develop adverse health outcomes and become more dependent on the healthcare system after experiencing a minor injury. Such poor underlying characteristics will be reflected in the data by the presence of higher individual HCE in the periods following the minor injury.

This is similar to health state modeling in the literature on hidden Markov models (HMMs) (Luo et al., 2021; Komariah and Sin, 2019) but using a two-step approach where the first consistently estimates the group membership (i.e. the "health state") and the second step models the dynamic behavior of the group membership. Unlike this literature, I will refer to health groups rather than "health state" given that the unobserved individual characteristics explaining HCE can be related to non-health factors (e.g. accessibility to healthcare services, peer effects, etc.).

The second objective of the empirical analysis is to compare the results coming from both algorithms (CEM and EM) when using real-world data with missing information. The predictive performance of the results obtained from each algorithm is compared using cross-validation, as detailed in Section 5.5.

5.2 Data Sources and Characteristics

The employed dataset is the Québec's portion of the Canadian Emergency departments Team Initiative (CETI). The CETI research program on mobility and aging is a Canadian clinical program "which aims to improve emergency department (ED) care for older adults with minor injuries" (Provencher et al., 2015). In the province of Québec, the CETI research program followed 1,391 patients with a medical consultation at the ED of one of three hospitals (Hôpital du Sacré-Coeur de Montréal, Hôpital de l'Enfant-Jésus and Hôpital du Saint-Sacrement) after a minor injury that occurred between 2009 and 2015. Individuals were included in the cohort if they were aged 65 and older and suffered from a minor injury that did not lead to hospitalization. A detailed description of the eligibility criteria and monitoring schedules can be found in Provencher et al. (2015).

The health administrative data for each participant has been extracted from Québec's physician claims database, which is managed by the Régie de l'Assurance-Maladie du Québec (RAMQ). Each claim included a unique patient identification number, billing codes corresponding to the services rendered, the diagnostic code according to the International Classification of Diseases, 9th Revision

(ICD-9-CM), dates and locations of the services provided, an identification code for the physician making the claim and the amount paid to the physician. The period covered by the dataset goes from January 1, 2008, to June 30, 2016. Note that physician claims do not include all the HCE that an individual could incur in a given period. However, physician claims are strongly correlated with total HCE while being usually more precise than other cost measures (e.g. hospitalization costs, costs of treatment episodes, etc.).

An unbalanced panel dataset containing 1,330 individuals and seven three-month periods per individual was created using the information described above. Some individuals died before the end of the last period, hence leading to 64 observations with no information. A weight of zero was attributed to those 64 observations at each (C)E-step of both algorithms. Consequently, the total number of observations with non-null weights in the dataset is 9,246. The initial ED visit occurred at the end of the second period while two follow-up visits occurred at the end of the third and fourth periods respectively. Price effects (inflation) were accounted for by adjusting all costs with Québec's all-item consumer price index (CPI) between 2008 and 2016.

5.3 Estimation Strategy

The employed model is a mixture of two-part models, where the first part corresponds to a Probit binary choice model and where the second part uses a lognormal density. The lognormal density accounts for the heavy-tailed distribution that is usually observed in individual HCE (Manning and Mullahy, 2001). Conceptually speaking, the first part models the "decision" of the i^{th} participant to incur a strictly positive amount of HCE at the t^{th} period, whereas the second part models the total amount, if any, that was incurred at this period by the participant. Two-part models have been extensively used in econometrics and health economics to model the decision-making of agents that lead to the generation of a strictly positive, continuous outcome depending on the initial decision of the agent (Norton et al., 2008; Neelon et al., 2011).

The specification used within each part of the model is similar to the one used for the second simulation exercise, as formulated by eq.(13). For simplicity, it is assumed that all covariates follow a normal distribution with means and covariance matrices varying across components. The general form of the g^{th} joint component's density for the it^{th} observation is written as follows

$$f_g(y_{it}^c, x_{it}|\theta_g, \xi_g) = \left[\left(1-\Phi(\eta_{itg}^b)\right) f_g(x_{it}^b|\boldsymbol{\mu}_g^b, \Sigma_g^b)\right]^{1-d_{it}} \left[\Phi(\eta_{itg}^b)\phi\left(\eta_{itg}\right) f_g(x_{it}|\boldsymbol{\mu}_g, \Sigma_g)\right]^{d_{it}}, \quad (15)$$

with $\eta_{itg}^b = X_{it}^b\tilde{\beta}_g^b$ and $\eta_{itg} = \frac{X_{it}\tilde{\beta}_g - y_{it}^c}{\omega^2_{\alpha+\epsilon,g}}$, where $f_g(\cdot|\boldsymbol{\mu}_g, \Sigma_g)$ corresponds to the multivariate normal probability density function (pdf) with mean $\boldsymbol{\mu}_g$ and variance-covariance matrix Σ_g, $\Phi(\cdot)$ and $\phi(\cdot)$ respectively refer to the cumulative density function and the pdf of the standard normal distribution, $y_{it}^c = \log(y_{it})$ stands as the log value of the HCE for the it^{th} observation (and is set to zero if $y_{it} = 0$), and $d_{it} = \mathbb{1}[y_{it} > 0]$ is equal to one if $y_{it} > 0$ and zero otherwise. All other coefficients have the

same interpretation as in Section 4.1.2, with the b superscript denoting the estimates of the binary part. Note that all covariates and parameters associated to the binary part need not be equal to those of the continuous part even if the true distributions of the common elements in x_{it}^b and x_{it} are identical. More details on the covariates included in each part are provided in the next subsection.

Estimation of the model is carried out using the same approaches as the ones described in Appendix B. Two differences are however worth noting. First, each M-step uses a Newton-Raphson procedure to update the estimates from both binary parts (one for each algorithm). Second, the weights $w_{itg}^{(k)}$ used for the empirical analysis are defined in Appendix B, where both the joint density classifier $z_{itg}^D(\theta, \xi)$ and the posterior probabilities $\tau_{itg}(\theta, \xi)$ are based on the joint density. Formally speaking, this means that $h_g(y_{it}^c, x_{it}|\theta, \xi) = f_g(y_{it}^c, x_{it}|\theta_g, \xi_g)$ and that $\tau_{itg}(\theta, \xi) = \frac{f_g(y_{it}^c, x_{it}|\theta_g, \xi_g)}{\sum_{j=1}^G f_j(y_{it}^c, x_{it}|\theta_j, \xi_j)}$, where $f_g(y_{it}^c, x_{it}|\theta_g, \xi_g)$ is defined as in eq.(15). This definition of the posterior probability avoids the need to choose between the binary and the continuous density to perform the EM algorithm. Not introducing any mixing weight in the RHS of the posterior probability $\tau_{itg}(\theta, \xi)$ also makes the two algorithms more comparable to each other. Such a formulation also suggests that the densities $f_g(x_{it}^b|\boldsymbol{\mu}_g^b, \Sigma_g^b)$ and $f_g(x_{it}|\boldsymbol{\mu}_g, \Sigma_g)$ are used as prior grouping information for computing $\tau_{itg}(\theta, \xi)$.

For simplicity, the unit-random effects of the binary and the continuous parts are assumed to be independent of each other. It has been shown that assuming independence between the unit-random effects of a two-part model is likely to introduce bias in the continuous part of the model (Su et al., 2009). Note that this bias is different from the sample selection bias often encountered in econometrics (Heckman, 1979). If the probability for every individual of generating a non-null amount of HCE is never equal to zero over time, then the sample selection bias will vanish as T increases if the sample is representative of the population of interest. Therefore, the bias introduced by the independence assumption between the two parts of the model may still remain even in the absence of any sample selection bias. Nonetheless, the finite mixture setup naturally (partially) accounts for correlation that might be caused by correlated group-specific intercepts or time-fixed effects. Introducing correlated random effects is also possible using Bayesian estimation methods or simulated maximum likelihood, but this goes beyond the scope of this paper.

If the independence assumption between the two unit-random effects is satisfied, then consistent estimation of the parameters can be done by estimating each part of the model separately. This implies that only the observations with a strictly positive outcome value are used to estimate the continuous parts of the model. On the other hand, the estimation the binary parts is performed using an iterative weighted Newton-Raphson procedure that uses all the observations in the sample. Convergence of the whole estimation procedure is assumed to be achieved when the relative change between two consecutive log likelihood values is less than 0.01%. A maximum number of 100 iterations is enforced for each single iterative procedure. All computations have been performed with Python 3.9.13 and Numpy 1.21.5.

5.4 Covariates

The list of covariates included in X_{it}^b and X_{it} are resumed in Appendix D. The included covariates are the same across all component densities. To proxy for frailty, I use the Elder's Risk Assessment (ERA) index, which predicts the hospitalization and health risks among elders (Crane et al., 2010). Frailty is known to be an important predictor of medical resource consumption and individual HCE over time (Sirven and Rapp, 2017). The composition of the ERA index is shown in Appendix E and slightly differs from the original index due to data availability issues.

In addition to frailty, the global amount of comorbidity is also known to be an important predictor of individual HCE (Charlson et al., 2008). Consequently, the Charlson index has been used to control for the overall burden of comorbidities in both parts of the model. The composition of the Charlson index is shown in Appendix E. Given that the ERA and Charlson indices have comorbidities in common, the overlapping covariates were removed from the Charlson index to limit collinearity issues. Note that only the time-averaged Charlson index was used in every specification due to strong collinearity between the time-varying and the time-averaged Charlson indices.

Finally, the Continuity of Care Index (COCI) developed by Bice and Boxerman (1977) was also introduced in the model to account for the peculiar relationships between each patient and the healthcare system. Continuity of care is defined as "how one patient experiences care over time as coherent and linked". Values of the index range from zero to one with a zero value referring to a total absence of continuity of care (i.e. each visit is associated with a different provider), whereas a value of one represents perfect continuity of care (i.e. all visits are associated with the same unique provider). Several studies have shown that the COCI is strongly associated with individual HCE and that even modest variations in the index value are associated with large variations in medical costs (Chu et al., 2012; Hussey et al., 2014). Given that no COCI can be computed when $y_{it} = 0$, the time-varying COCI is only used in the continuous part of the model.[8]

5.5 Selection of the Initial Parameter Values and Number of Groups

Between 300 and 2,000 different sets of initial parameters were assessed to estimate the model for each algorithm and each value of G.[9] The selection of the "optimal" initial parameter values and number of groups G was performed using the maximum value of the corresponding likelihood function and cross-validation (CV). If some observations are misclassified, then maximizing the likelihood might select parameter estimates that are inconsistent and that are far away from the true parameters. This is why CV was also used to select the optimal set of initial parameter values and number of groups.

[8] To maintain comparability with the other coefficients, the COCI has been rescaled from 0 to 10.

[9] The number of initial parameter values is different for each algorithm since it is more frequent for the CEM algorithm to experience convergence issues due to the presence of empty component(s). The exact number of initial parameter values used for each value of G is indicated in the code provided by the author upon request.

The goal of the CV procedure is to estimate the out-of-sample prediction error associated with the estimated parameters and the chosen number of groups. Following the recommendations of Zhang and Yang (2015), a 2-fold CV procedure with 10 different data splittings was performed for for each value of $G \in \{2, 6\}$ and each set of estimates associated with the 15 highest likelihood values.[10] All data splittings randomly allocated each unit to the training set or the test set, which maintains the correlation structure among units in the training set. To limit the probability of being "trapped" in a local maximum of the objective function, the estimated parameter values of the complete dataset were used as the initial values for each repetition during the CV procedure.

The predicted values of the test set for both algorithms are computed as follows

$$\hat{y}_{it}^{c} = \widehat{\log(y_{it})} = \sum_{g=1}^{G} w_{itg}^{(k)} \Phi(X_{it}^{b} \hat{\tilde{\beta}}_{g}^{b,(k)}) X_{it} \hat{\tilde{\beta}}_{g}^{(k)},$$

where $w_{itg}^{(k)}$ is defined as in Section 5.3, and where $\hat{\tilde{\beta}}_{g}^{b,(k)}$ and $\hat{\tilde{\beta}}_{g}^{(k)}$ are the respective estimated analogs of $\tilde{\beta}_{g}^{b}$ and $\tilde{\beta}_{g}$ after convergence of the algorithm. The cross-validated RMSEs were computed by combining the prediction errors of all observations for each test set and each data splitting. Note that all $\hat{\tilde{\beta}}_{g}^{(k)}$ correspond to semi-elasticities and that retransformation to the original scale is not necessary to perform the CV procedure.

5.6 Results

5.6.1 Selection of the Initial Parameter Values and Number of Groups

Figure 3 shows the relative, cross-validated RMSE values obtained by the CV procedure described in Section 5.5 for each algorithm, each value of G, and each set of initial parameter values that produced the 15 highest likelihood values. The values shown in Figure 3 are relative to the RSME computed when $G = 1$, which corresponds to a value of 2.05.

The results illustrated in Figure 3 show that the EM algorithm minimizes the cross-validated RMSE when $G = 4$ while the CEM algorithm minimizes the cross-validated RMSE when $G = 5$. Several repetitions of the CV procedure have been performed to ensure the stability of the lowest RMSE values for each algorithm (results not shown). The lowest RMSE value obtained from the CEM algorithm is equal to 0.89, which is 17.6% smaller than the lowest RMSE value obtained from the EM algorithm. This also corresponds to a reduction of 56.6% compared to the standard, single-component two-part model (2.05 vs. 0.89). Figure 3 also shows that the out-of-sample prediction errors obtained from the CEM algorithm are generally lower and less variable than those obtained from the EM algorithm, especially when $G = 2$. For parsimony, the subsequent subsections will focus on the set of parameter estimates that yielded the lowest RMSE among all values.

[10] The 15 highest likelihood values were chosen to reduce the computational burden of the CV procedure for the selection of the initial parameter values.

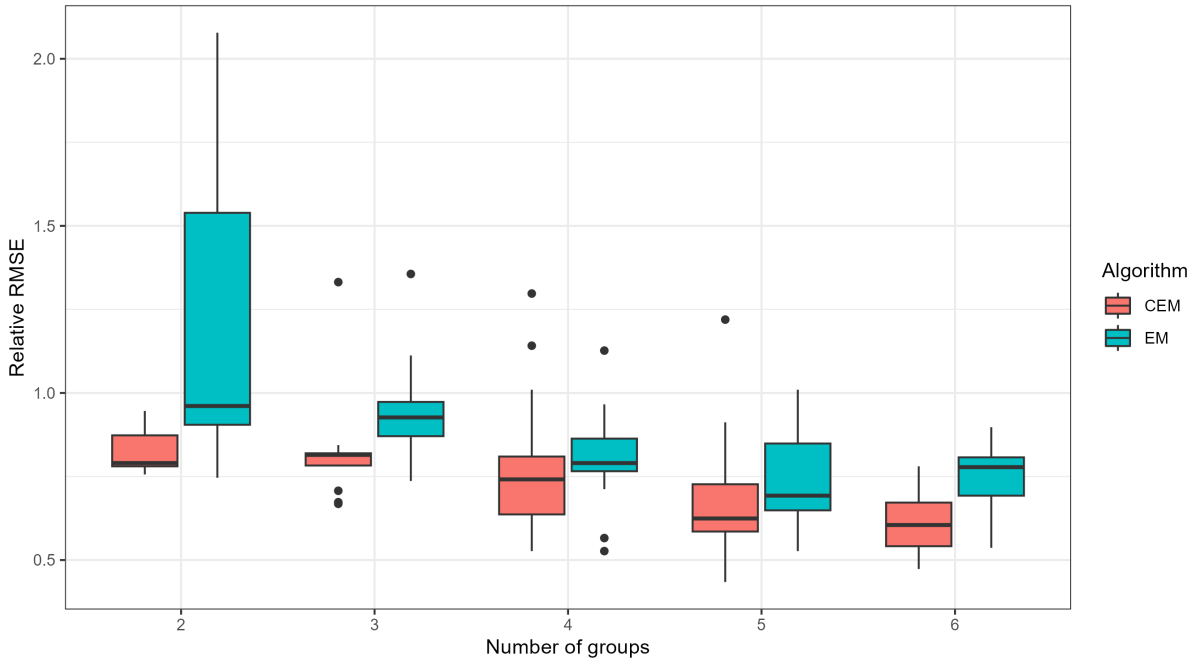


Figure 3: Relative cross-validated RMSE values obtained by repeated 2-fold cross-validation (with 10 repetitions) for each one of the 15 highest likelihood values obtained by random initialization.

5.6.2 Groups' Analysis

Table 4 shows the estimated mean and variance values of each covariate for each group associated to the optimal set of estimates according to the CV procedure. Note that the distributions of the covariates in each group are likely to overlap with each other, as denoted by the large within-group variance values. Although each group contains observations that are homogeneous with respect to their unobserved characteristics, this means that all groups might contain observations that appear very different from each other based on their *observed* characteristics. Because of this *observed* heterogeneity, it is hard to broadly characterize the observations contained within each group. This is why the expression "mostly feature" is used below for the description of each group.

The first group contains observations that mostly feature a moderately high number of comorbidities, as indicated by columns (5) and (9), and low continuity of care. The second group contains observations that mostly feature a larger number of comorbidities compared to the first group, but also a higher contemporaneous level of continuity of care, as indicated by column (7). The third group contains observations that mostly feature very low continuity of care and a moderately high number of comorbidities. Note that this group is the largest in size and also has the largest proportion of males. The fourth group contains observations that mostly feature a very low number of comorbidities and a relatively high continuity of care. This group also features the smallest proportion of males. Finally, the fifth group contains exclusively observations that feature perfect continuity of care. It is also the smallest group in size and observations contained in this

Group number	Estimated Mixing Weights	Moment	Male	ERA	Time-averaged ERA	COCI	Time-averaged COCI	Time-averaged Charlson
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All groups	1.000	Mean	0.38	2.05	2.05	3.19	4.00	1.35
		Variance	0.23	2.96	2.61	9.81	4.30	1.34
1	0.224	Mean	0.38	1.46	1.57	1.34	4.01	1.50
		Variance	0.24	1.28	1.41	0.60	5.33	0.94
2	0.192	Mean	0.39	3.90	3.68	3.06	3.25	2.41
		Variance	0.24	3.29	2.85	3.19	2.42	2.32
3	0.232	Mean	0.40	2.03	2.06	0.99	3.49	0.93
		Variance	0.24	2.64	2.38	0.31	3.66	0.63
4	0.201	Mean	0.35	1.22	1.30	3.49	1.30	0.82
		Variance	0.23	0.86	0.91	1.39	2.86	0.32
5	0.151	Mean	0.37	1.83	1.80	10.00	5.67	1.18
		Variance	0.23	2.46	2.08	0.00	3.65	1.00

Table 4: Descriptive statistics of the covariates contained within each group created by the optimal set of estimates according to the CV procedure.

group mostly feature a relatively low number of comorbidities. Given that one covariate is constant within this group, this creates a perfect collinearity issue that can be dealt with by removing the group's intercept.

Note that there exist large differences between the time-varying and the time-averaged COCI in most groups. For instance, the first and third groups contain observations that mostly feature low continuity of care at the current period, but high continuity of care on average over time. This is the opposite for the fourth and fifth groups. Therefore, it is not unreasonable to think that the level of continuity of care, both contemporaneously and over time, is an important factor in determining group memberships in the sample.

Figure 4 shows the estimated time-fixed effects for each group and each part of the model that are associated to the same set of estimates. Confidence intervals on the time-fixed effects (and all other coefficients) are computed using eq.(11). The first period, ranging from 6 months to 3 months before the initial ED visit, is set as the reference value for the other time-fixed effects. The shaded areas correspond to the 95% cluster-robust confidence intervals. Given that the continuous part of the model is in the log scale, each mean value in the graphs on the RHS of Figure 5 corresponds to the average, relative increase in individual HCE (in %) for each period. For instance, the observations in Group 1 feature an average increase of 70% in their individual HCE at $t = 5$ compared to $t = 1$, which is explained by the low levels of continuity of care and the poor health of the observations contained within this group.

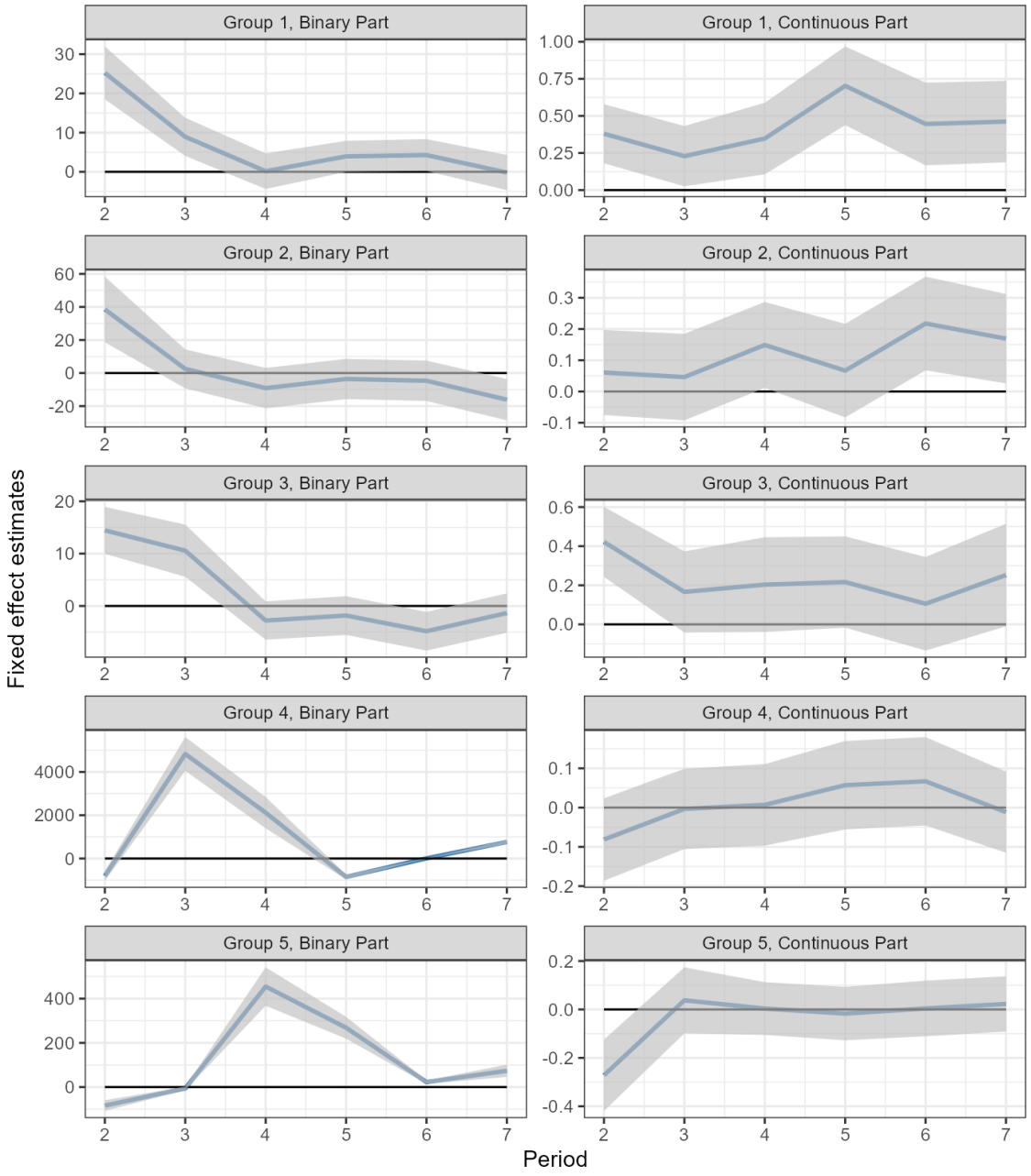


Figure 4: Time-fixed effects associated to the optimal set of estimates according to the CV procedure. The value of the first time-fixed effect is equal to zero and is set as the reference value. The initial visit to the ED occurs at the end of the second period. The shaded areas correspond to the 95% cluster-robust confidence interval and do not account for uncertainty in group memberships.

The marginal effects of the time-fixed effects in the binary parts can be obtained using the estimated coefficient values shown in Appendix F.1 and the covariates' average values presented in Table 4. Nonetheless, the sign of the mean values of the time-fixed effects on the LHS of Figure 4 still indicates the direction of each effect on the probability of consuming medical resources. The

very large time-fixed effects depicted in Group 4 at $t = 3$ and $t = 4$ indicate that every or almost every observation in this group consumed a strictly positive amount of medical resources during those two periods. Note that the initial ED visit after the minor injury did not necessarily lead to a visit to the physician, thus explaining why some time-fixed effects at $t = 2$ are significantly negative in both parts of the model.

The analogous set of time-fixed effects that are associated with the "best" parameters obtained from the EM algorithm is shown in Appendix F.3. Trying to make formal connections between the estimates obtained from each algorithm is, according to me, hazardous given that the two methods do not treat the observed information similarly, and can produce results that are substantially different from each other. Such differences can also be appreciated by examining all other parameters estimated by each algorithm (see Appendix F).

5.6.3 Transition Between Groups

Using the results from Table 4 and Figure 4, it is possible to broadly characterize each group. This is shown below in Figure 5, along with the proportions of the total number of observations contained in each group at each period. This "naming" exercise is highly subjective and does not rely on a comprehensive analysis of all observed characteristics and results. The proposed "names" for the groups rely exclusively on frailty and continuity of care, which are known to be good predictors of individual HCE (Sirven and Rapp, 2017; Chu et al., 2012; Hussey et al., 2014).[11] Note that the group numbers have been previously arranged in descending order so that the first group corresponds to the "high-cost" (i.e. frail with low continuity of care) group while the fifth group corresponds the "low-cost" group (i.e. robust with perfect continuity of care).

Figure 5 shows that the initial ED visit substantially reduced the number of units in the group with perfect continuity of care at $t = 2$. This is not surprising since the initial ED visit was associated with an unexpected minor trauma, therefore increasing the chance for the patient's regular healthcare provider to be unavailable at this precise moment. On the other hand, the initial ED visit seemed to have substantially increased the number of patients in Group 3. This can be explained by the fact that (robust) individuals who rarely consume medical resources can easily switch from perfect continuity to low continuity of care if they see a different healthcare provider each time they consume medical resources.

Figure 5 also shows that Group 4 experienced the largest reduction in size with a decrease of 3.4

[11] More precisely, it was assumed that a group is composed of frail individual-periods if at least one of the time-fixed effects of the continuous part is significantly positive after the initial injury, and if the group mean value for the ERA or the Charlson index is larger than its global average. It was assumed to be composed of robust individual-periods if no time-fixed effect of the continuous part is significantly different from zero after the initial injury. Overall continuity of care was based exclusively on the average value of the time-varying COCI covariate, as shown in column (7) of Table 4, where the attribution of the "type" of continuity of care (i.e. low, moderate, and perfect) is obvious. Note that the overlap between the distributions of the COCI covariate for each type of continuity of care is very small compared to the other covariates, which advocates for the use of this covariate to broadly define the groups.

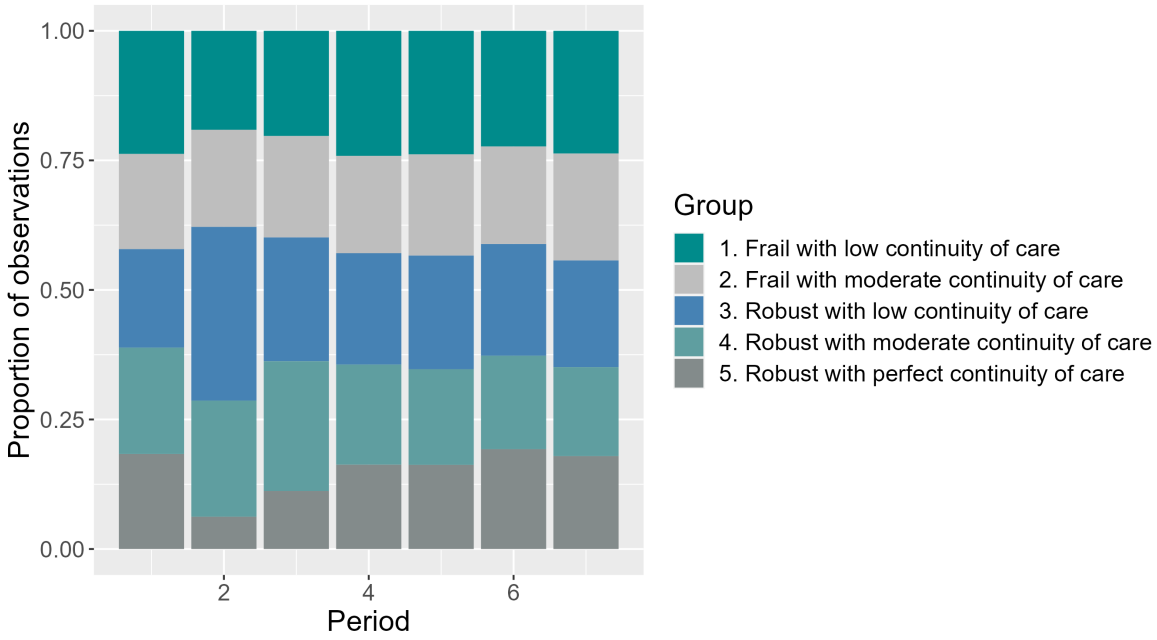


Figure 5: Proportions of the total number of observations in each group at each period.

percentage points between period 1 and period 7, while Group 2 is the group whose size increased the most with 2.3 additional percentage points between period 1 and period 7. This can be easily explained by a natural transition from robustness to frailty as time passes after the minor injury.

Figure 6 depicts the estimated transition matrix of the group memberships over time. This transition matrix shows that robustness and frailty are persistent characteristics given that transitions from a robust group to another robust group account for 74% of all transitions out of a robust group on average, whereas transitions from a frail group to another frail group account for 65% of all transitions out of a frail group on average. Therefore, robustness appears more persistent than frailty, which is in line with the fact that frailty is reversible provided appropriate care and healthy lifestyle habits (Kolle et al., 2023).

Such a transition matrix can also be used to predict the group membership at period $t+1$ given the group membership at period t. Using the Bayes' rule of categorical assignment and the probabilities shown in Figure 6, group membership at period $t+1$ is correctly predicted 3,226 times out of 7,980 total memberships, which corresponds to a "success" rate of 40.4%. This success rate could be easily improved upon using more sophisticated approaches such as dynamic multinomial logit models, namely to include additional observed factors that might affect future group memberships. An applied version of this paper details the results of such a categorical model.

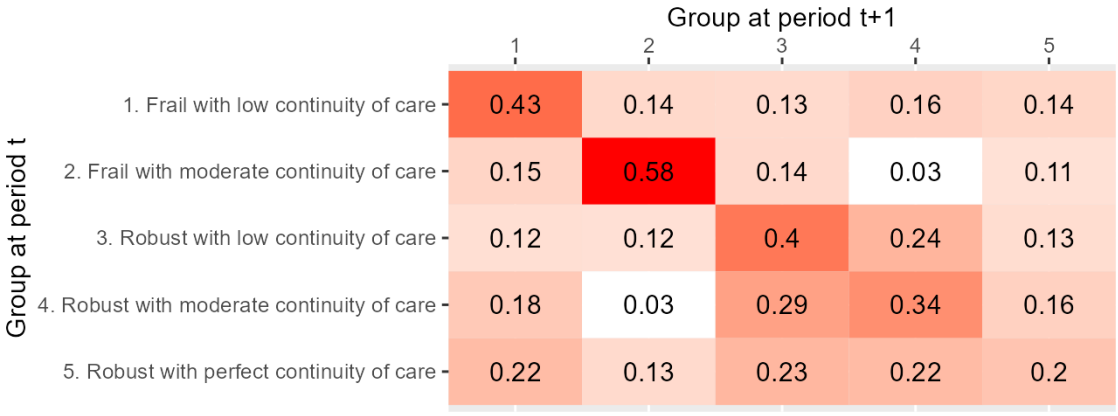


Figure 6: Estimated transition matrix of the group memberships based on the optimal set of estimates according to the CV procedure.

6 Conclusion

This paper showed that maximizing the likelihood function of a mixture density, as generally defined by eq.(5), leads to inconsistent estimates of the parameters governing any kind of finite mixtures under weak regularity conditions. It is worth emphasizing that this inconsistency does not depend on the chosen algorithm to maximize the mixture likelihood. Indeed, the motivation behind the widespread use of the EM algorithm is that it maximizes the mixture likelihood function just as any other numerical optimization method, but is easier to implement in practice. Still, it is true that for the same initial parameter values, different algorithms might converge to different local maxima. This does not however invalidate any of the claims made in this paper.

Rather than relying on the mixture likelihood function, this paper showed that maximizing the max-component log likelihood function combined with a consistent classifier leads to the consistent estimation of all parameters in the mixture. As shown in Section 3.3, consistency of the classifier is essential to obtain consistent estimates of the component parameters and the mixing weights. Although uniform consistency of the chosen classifier is a desirable property, Theorem 3.5 showed that it is not necessary to obtain consistent estimates of all parameters in the mixture. Nonetheless, Corollary 3.3 showed that the Mahalanobis distance classifier is uniformly consistent as the number of covariates goes to infinity at a faster rate than the sample size for a fixed number of groups, which leads to estimation issues. Given that the covariates used for each step of the CEM algorithm need not be the same, one could still use a setup where $p >> NT$ to compute the Mahalanobis distance classifier, and then select only a relevant subset of the available covariates at each M-step using standard regularization techniques.

Contrary to the recent papers on the subject, the estimation strategy proposed in this paper has the benefit of leaving group membership completely unrestricted. This has important implications

from a policy perspective given that transitions between groups over time allow for the prediction of future group memberships, which can then be used to improve decision-making in healthcare. Achieving such an objective implies that group membership has to be consistently estimated in the first step. The dynamic behavior of the group membership can then modeled in the second step using a consistent estimator for categorical outcomes. Such a two-step procedure is advocated by Bonhomme et al. (2019) in the context of matched employer-employee panel data analysis while being more in line with the literature on finite mixtures where, typically, no restriction is imposed on group memberships to estimate the mixture parameters.

A general recommendation concerning the estimation of any kind of finite mixture models (e.g. Gaussian mixture models, mixture of experts, latent group panel structure, etc.) with unrestricted group membership can thus be formulated : instead of using the EM algorithm, one should always use the CEM algorithm combined to a consistent classifier with as many covariates as possible. If the large number of covariates makes the computational burden too heavy, then reducing the number of covariates should be considered during the realization of each M-step. If the number of covariates is still too large, then one should remove most of the covariates that have a very high degree of collinearity between each other. It is however important to recognize that high collinearity does not necessarily imply classification irrelevance, and doing so might increase the misclassification rate. In the context where no covariate is available, a multivariate outcome can be used to reduce the misclassification rate under the assumption that group memberships do not change within any single observation. If the outcome is univariate and no covariate is available, then it is impossible to obtain consistent estimates of the mixture parameters and the size of the asymptotic bias will depend on the degree of overlap between the component densities.

Results from the simulation exercises and a real-world application showed that the estimation procedure leads to less biased and more stable estimates than those obtained by standard MLE procedures such as the EM algorithm. Results from the empirical analysis also show that the proposed estimation strategy identified five heterogeneous health groups that account for a large part of the unobserved heterogeneity in the sample. The use of the proposed estimation strategy reduced the out-of-sample prediction error by more than 55% compared to the single-component model and by 17.6% compared to the best results obtained from standard, widely used MLE procedures.

Compared to other algorithms, the proposed estimation strategy contrasts with HMMs where the dynamic behavior of the latent variable is estimated simultaneously with the parameters of each component density. Although this method has been proven to be consistent (Douc and Matias, 2001), the two-step estimation procedure proposed in this paper seems more efficient and more robust to random initialization than HMMs. Comparison of the proposed two-step approach to other machine learning techniques, such as random forests, and inclusion of feedback effects remain open areas for future research (Chamberlain, 2022).

References

Ahn, Y. and H. Kasahara (2024, January). Difference in Differences with Latent Group Structures. *Working Paper*.

Amengual, D., X. Bei, M. Carrasco, and E. Sentana (2024, March). Score-type tests for normal mixtures. *Journal of Econometrics*, 105717.

Aragam, B. and R. Yang (2023, February). Uniform consistency in nonparametric mixture models. *The Annals of Statistics 51*(1).

Bai, J. (2009). Panel Data Models With Interactive Fixed Effects. *Econometrica 77*(4), 1229–1279.

Bice, T. W. and S. B. Boxerman (1977). A Quantitative Measure of Continuity of Care. *Medical Care 15*(4), 347–349. Publisher: Lippincott Williams & Wilkins.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information science and statistics. New York: Springer.

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica 87*(3), 699–739.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing Unobserved Heterogeneity. *Econometrica 90*(2), 625–643.

Bonhomme, S. and E. Manresa (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica 83*(3), 1147–1184. Publisher: [Wiley, The Econometric Society].

Boot, T. and A. Pick (2018, October). Optimal Forecasts from Markov Switching Models. *Journal of Business & Economic Statistics 36*(4), 628–642.

Bottou, L. and Y. Bengio (1994). Convergence Properties of the K-Means Algorithms. In *Advances in Neural Information Processing Systems*, Volume 7. MIT Press.

Bryant, P. and J. A. Williamson (1978). Asymptotic Behaviour of Classification Maximum Likelihood Estimates. *Biometrika 65*(2), 273–281.

Bryant, P. G. (1991, January). Large-sample results for optimization-based clustering methods. *Journal of Classification 8*(1), 31–44.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a, April). Alternative Asymptotics and the Partially Linear Model with Many Regressors. *Econometric Theory 34*(2), 277–301.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018b, July). Inference in Linear Regression Models with Many Covariates and Heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

Celeux, G. (2019, January). EM Methods for Finite Mixtures. In S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert (Eds.), *Handbook of Mixture Analysis* (1 ed.)., pp. 21–39. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC.

Celeux, G. and G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis 14*(3), 315–332.

Chamberlain, G. (2022, January). Feedback in panel data models. *Journal of Econometrics 226*(1), 4–20.

Charlson, M. E., R. E. Charlson, J. C. Peterson, S. S. Marinopoulos, W. M. Briggs, and J. P. Hollenberg (2008, December). The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. *Journal of Clinical Epidemiology 61*(12), 1234–1240.

Chen, J. (2017, February). Consistency of the MLE under Mixture Models. *Statistical Science 32*(1).

Chu, H.-Y., C.-C. Chen, and S.-H. Cheng (2012, November). Continuity of Care, Potentially Inappropriate Medication, and Health Care Outcomes Among the Elderly: Evidence From a Longitudinal Analysis in Taiwan. *Medical Care 50*(11), 1002–1009.

Compiani, G. and Y. Kitamura (2016). Using mixtures in econometric models: a brief review and some new results. *The Econometrics Journal 19*(3), C95–C127.

Crane, S. J., E. E. Tung, G. J. Hanson, S. Cha, R. Chaudhry, and P. Y. Takahashi (2010, December). Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits: The elders risk assessment index. *BMC Health Services Research 10*(1), 338.

Deb, P. and P. K. Trivedi (1997). Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics 12*(3,), 313–336.

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

Douc, R. and C. Matias (2001, June). Asymptotics of the Maximum Likelihood Estimator for General Hidden Markov Models. *Bernoulli 7*(3), 381.

Dzemski, A. and R. Okui (2021, June). Convergence rate of estimators of clustered panel models with misclassification. *Economics Letters 203*, 109844.

Dzemski, A. and R. Okui (2024). Confidence Set for Group Membership. *Quantitative Economics 15*(2), 245–277.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer series in statistics. New York: Springer. OCLC: ocm71262594.

Gepperth, A. and B. Pfülb (2021, December). Gradient-Based Training of Gaussian Mixture Models for High-Dimensional Streaming Data. *Neural Processing Letters 53*(6), 4331–4348.

Gong, G. and F. J. Samaniego (1981, July). Pseudo Maximum Likelihood Estimation: Theory and Applications. *The Annals of Statistics 9*(4).

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica 52*(3), 681–700. Publisher: [Wiley, Econometric Society].

Hahn, J. and W. Newey (2004). Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica 72*(4), 1295–1319. Publisher: [Wiley, Econometric Society].

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2 ed.). Springer series in statistics. New York: Springer International Publishing.

Heckman, J. and B. Singer (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica 52*(2), 271–320.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–161. Publisher: [Wiley, Econometric Society].

Higgins, A. and K. Jochmans (2022, April). Bootstrap inference for fixed-effect models. *Working Paper Toulouse School of Economics*(1328), 28.

Hsiao, C. (2014). *Analysis of Panel Data* (3rd ed.). New York: Cambridge University Press.

Hussey, P. S., E. C. Schneider, R. S. Rudin, D. S. Fox, J. Lai, and C. E. Pollack (2014, May). Continuity and the Costs of Care for Chronic Disease. *JAMA Internal Medicine 174*(5), 742.

Jones, A. M., J. Lomas, and N. Rice (2015). Healthcare Cost Regressions: Going Beyond the Mean to Estimate the Full Distribution. *Health Economics 24*(9), 1192–1212.

Kasteridis, P., N. Rice, and R. Santos (2022, December). Heterogeneity in end of life health care expenditure trajectory profiles. *Journal of Economic Behavior & Organization 204*, 221–251.

Keane, M. and K. Wolpin (1997, June). The Career Decisions of Young Men. *Journal of Political Economy 105*(3), 473–522.

Kolle, A. T., K. B. Lewis, M. Lalonde, and C. Backman (2023, November). Reversing frailty in older adults: a scoping review. *BMC Geriatrics 23*(1), 751.

Komariah, K. S. and B.-K. Sin (2019, July). Health State Modeling and Prediction based on Hidden Markov Models. In *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 245–250. ISSN: 2165-8536.

Kwon, J. and C. Caramanis (2019, November). EM Converges for a Mixture of Many Linear Regressions. arXiv:1905.12106 [cs, stat].

Liu, R., Z. Shang, Y. Zhang, and Q. Zhou (2020, April). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics 215*(2), 574–590.

Lumsdaine, R. L., R. Okui, and W. Wang (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics 233*(1), 45–65.

Luo, Y., D. A. Stephens, A. Verma, and D. L. Buckeridge (2021). Bayesian latent multi-state modeling for nonequidistant longitudinal electronic health records. *Biometrics 77*(1), 78–90.

Manning, W. G. and J. Mullahy (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 34.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019, March). Finite Mixture Models. *Annual Review of Statistics and Its Application 6*(1), 355–378.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models* (1 ed.). John Wiley & Sons, Ltd.

Neelon, B., A. J. O'Malley, and S.-L. T. Normand (2011). A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity. *Biometrics 67*(1), 280–289.

Neyman, J. and E. L. Scott (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica 16*(1), 1–32. Publisher: [Wiley, Econometric Society].

Norton, E. C., W. H. Dow, and Y. K. Do (2008, December). Specification tests for the sample selection and two-part models. *Health Services and Outcomes Research Methodology 8*(4), 201–208.

Okui, R. and W. Wang (2021, February). Heterogeneous structural breaks in panel data models. *Journal of Econometrics 220*(2), 447–473.

Pollard, D. (1981, January). Strong Consistency of K-Means Clustering. *The Annals of Statistics 9*(1).

Provencher, V., M.-J. Sirois, M.-C. Ouellet, S. Camden, X. Neveu, N. Allain-Boulé, and M. Emond (2015). Decline in Activities of Daily Living After a Visit to a Canadian Emergency Department for Minor Injuries in Independent Older Adults: Are Frail Older Adults with Cognitive Impairment at Greater Risk? *Journal of the American Geriatrics Society 63*(5), 860–868.

Qian, J. and L. Su (2016, March). Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *Journal of Econometrics 191*(1), 86–109.

Redner, R. A. and H. F. Walker (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review 26*(2), 195–239.

Samé, A., C. Ambroise, and G. Govaert (2007, August). An online classification EM algorithm based on the mixture model. *Statistics and Computing 17*(3), 209–218.

Sirven, N. and T. Rapp (2017, March). The cost of frailty in France. *The European Journal of Health Economics 18*(2), 243–253.

Su, L., Z. Shi, and P. C. B. Phillips (2016). Identifying Latent Structures in Panel Data. *Econometrica 84*(6), 2215–2264.

Su, L., B. D. M. Tom, and V. T. Farewell (2009, April). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics 10*(2), 374–389.

Su, L., X. Wang, and S. Jin (2019, April). Sieve Estimation of Time-Varying Panel Data Models With Latent Structures. *Journal of Business & Economic Statistics 37*(2), 334–349.

Tanaka, K. (2009). Strong Consistency of the Maximum Likelihood Estimator for Finite Mixtures of Location-Scale Distributions When Penalty is Imposed on the Ratios of the Scale Parameters. *Scandinavian Journal of Statistics 36*(1), 171–184.

Wang, W., P. C. Phillips, and L. Su (2019, January). The heterogeneous effects of the minimum wage on employment across states. *Economics Letters 174*, 179–185.

Wang, W., P. C. B. Phillips, and L. Su (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics 33*(6), 797–815.

Wang, W. and L. Su (2021, February). Identifying latent group structures in nonlinear panels. *Journal of Econometrics 220*(2), 272–295.

Wang, Y., P. C. Phillips, and L. Su (2024, March). Panel data models with time-varying latent group structures. *Journal of Econometrics 240*(1), 105685.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). Cambridge, Massachusetts: The MIT Press.

Yang, Y. (2022, April). A correlated random effects approach to the estimation of models with multiple fixed effects. *Economics Letters 213*, 110408.

Zhang, Y. and Y. Yang (2015, July). Cross-validation for selecting a model selection procedure. *Journal of Econometrics 187*(1), 95–112.

A Appendix - Proofs

A.1 Lemma 3.1

Proof. Maximizing $\mathbb{E}_0[\log f(y_{it}|x_{it};\theta,\tilde{\pi})]$ with respect to θ is similar to maximizing $\mathbb{E}_0\left[\log\left(\frac{f(y_{it}|x_{it};\theta,\tilde{\pi})}{f(y_{it}|x_{it};\theta^0,\pi^0)}\right)\right]$ with respect to θ given that $f(y_{it}|x_{it};\theta^0,\pi^0)>0$ by Assumption 1(ii). By Jensen's inequality, we have that

$$\begin{aligned}
\mathbb{E}_0\left[\log\left(\frac{f(y_{it}|x_{it};\theta,\tilde{\pi})}{f(y_{it}|x_{it};\theta^0,\pi^0)}\right)\right] &\leq \log\left(\mathbb{E}_0\left[\frac{f(y_{it}|x_{it};\theta,\tilde{\pi})}{f(y_{it}|x_{it};\theta^0,\pi^0)}\right]\right), \\
&\leq \log\left(\int_{\mathcal{Y}}\frac{f(y_{it}|x_{it};\theta,\tilde{\pi})}{f(y_{it}|x_{it};\theta^0,\pi^0)}f(y_{it}|x_{it};\theta^0,\pi^0)\upsilon(dy_{it})\right), \\
&\leq \log\left(\int_{\mathcal{Y}}f(y_{it}|x_{it};\theta,\tilde{\pi})\upsilon(dy_{it})\right), \\
&\leq 0,
\end{aligned}$$

given that $f(y_{it}|x_{it};\theta,\tilde{\pi})$ is a well-defined density that integrates to one over $\mathcal{Y}$. Therefore, we have that

$$\mathbb{E}_0[\log(f(y_{it}|x_{it};\theta,\tilde{\pi}))] \leq \mathbb{E}_0[\log(f(y_{it}|x_{it};\theta^0,\pi^0))],$$

with equality if and only if $\theta = \theta^0$ and $\tilde{\pi} = \pi^0$, otherwise Assumption 1(v) would be violated. Therefore, we have that

$$\mathbb{E}_0[\log(f(y_{it}|x_{it};\theta^0,\tilde{\pi}))] < \mathbb{E}_0[\log(f(y_{it}|x_{it};\theta^0,\pi^0))],$$

for any $\tilde{\pi} \in \zeta$. A similar argument can be made by looking at the expected value of the score function when evaluated at θ^0. The first-order condition for consistency implies that

$$\mathbb{E}_0[s_{it}(\theta)]\big|_{\theta=\theta^0} = \mathbb{E}_0\left[\frac{\partial \log f(y_{it}|x_{it};\theta,\pi)}{\partial\theta}\right]\Bigg|_{\theta=\theta^0} = 0,$$

by Assumption 1(vi), which can be rewritten as

$$\int_{\mathcal{Y}} \frac{f(y_{it}|x_{it};\theta^0,\pi^0)}{f(y_{it}|x_{it};\theta^0,\pi)} \frac{\partial f(y_{it}|x_{it};\theta,\pi)}{\partial\theta} \upsilon(dy_{it})\Bigg|_{\theta=\theta^0} = 0.$$

It is easy to see that, under interchangeability of the derivative and the integral, this condition is satisfied for any $(\theta,\pi) \in \Theta \times \Pi$ if $\pi = \pi^0$ given that this reduces to

$$\frac{\partial}{\partial\theta}\int_{\mathcal{Y}} f(y_{it}|x_{it};\theta,\pi)\upsilon(dy_{it})\Bigg|_{\theta=\theta^0} = \frac{\partial 1}{\partial\theta} = 0.$$

If $\pi \neq \pi^0$, this condition does not generally hold. However, it is true that there could exists a $\tilde{\pi} \in \zeta$ such that this condition holds, but such a $(\tilde{\pi},\theta^0)$ would be located at a local maximum of the expected log likelihood function as a consequence of Assumption 1(v). □

A.2 Corollary 3.1

Proof. Define $\pi^* \notin \Pi$ such that $\pi^* = (0,...,1,...,0)$, where the mixture problem is treated as a single-component density, non-mixture estimation problem. Clearly, we have that $\text{plim}_{N,T\to\infty}\hat{\theta}_{NT}(\pi^*) \neq \theta^0$, otherwise a single-component density would suffice to get consistent estimates of θ. Therefore, we have that

$$\lim_{\tilde{\pi}\to\pi^*} \operatorname*{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\tilde{\pi}) \neq \theta^0,$$

and that

$$\operatorname*{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\tilde{\pi}_{\epsilon,\pi^*}) \neq \theta^0,$$

where $\tilde{\pi}_{\epsilon,\pi^*}$ corresponds to any $\tilde{\pi} \in \zeta$ that lies within a ball $B(\epsilon,\pi^*)$ centered around π^* and of radius $\epsilon > 0$, otherwise Assumption 1(vi) would not be satisfied. However, it is true that there

could exist a finite number of values $\tilde{\pi}_{\delta,\pi^*} \in B(\delta, \pi^*)$ with $\delta > \epsilon$ such that $\pi^0 \notin B(\delta, \pi^*)$, and such that

$$\operatorname*{plim}_{N,T\to\infty} \hat{\theta}_{NT}(\tilde{\pi}_{\delta,\pi^*}) = \theta^0,$$

while still satisfying Assumption 1(vi). This implies that for almost every vector $\tilde{\pi} \in \zeta$, $\hat{\theta}_{NT}(\tilde{\pi})$ will not converge in probability to θ^0 as $N, T \to \infty$. □

A.3 Proposition 3.1

Proof. Maximizing the log likelihood with respect to π leads to a corner solution given that

$$\frac{\partial l(\theta,\pi)}{\partial \pi} = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\sum_{g=1}^{G} \frac{\partial \pi_g f_g(y_{it}|x_{it};\theta_g)}{\partial \pi}}{\sum_{j=1}^{G} \pi_j f_j(y_{it}|x_{it};\theta_j)} = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\sum_{g=1}^{G} f_g(y_{it}|x_{it};\theta_g)}{\sum_{j=1}^{G} \pi_j f_j(y_{it}|x_{it};\theta_j)}.$$

By Assumption 1(ii), the NT terms in the sum are all strictly positive, including the mixing weight π_j in the denominator. Therefore, $\frac{\partial l(\theta,\pi)}{\partial \pi}$ can never be equal to zero, except in cases where $||\theta_g||^2 \to \infty$ implies that $f_g(y_{it}|x_{it};\theta_g) \to 0$ for all $g \in \mathbb{G}$. This means that the log likelihood function has no well-defined maximum nor minimum in the interior of Π, which is indicative of a corner solution. This can also be deduced by looking at the expression of $l(\theta,\pi)$ when θ and $(\mathbf{y},\mathbf{x})$ are taken as given : under Assumption 1, the vector of estimated mixing weights $\hat{\pi}_\theta$ will put a weight of one on the component that yields the largest log likelihood value, and a weight of zero on all the other components. □

A.4 Lemma 3.2

Proof. Taking all $\tau_{itg}(\theta,\pi)$ as constants, the derivative of $\mathbb{E}_z[l^C(\theta,\pi)]$ with respect to π_g leads to

$$\frac{\partial \mathbb{E}_z[l^C(\theta,\pi)]}{\partial \pi} = \frac{\partial}{\partial \pi}\sum_{g=1}^{G} \log \pi_g \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{itg}(\theta,\pi).$$

This expression corresponds to the derivative of the negative cross entropy function between π_g and $\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi)$. From information theory, we know that the cross entropy function is minimized only when the two variables are identically distributed. Therefore, the negative cross entropy function is maximized if and only if $\pi_g = \alpha\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi)$ for all values of $g \in \mathbb{G}$, where α is a normalizing constant. Since the mixing weights *have* to sum to one, this implies that

$$\alpha\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi) = 1,$$

thus leading to

$$\alpha = \frac{1}{\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi)} = \frac{1}{NT},$$

given that $\sum_{g=1}^{G}\tau_{itg}(\theta,\pi)=1$ by construction. Therefore, we have that $\pi_g(\theta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi)$ is the MLE of π_g when $\mathbb{E}_z[l^C(\theta,\pi)]$ is used as the objective function. □

A.5 Theorem 3.1

Proof. From equations (7) and (8), we have that

$$\hat{\pi}_g(\theta^0)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\hat{\pi}_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}.$$

Since both $\hat{\pi}_g(\theta)$ and $f_g(y_{it}|x_{it};\theta_g)$ are continuous functions of θ for any $g\in\mathbb{G}$, we can apply the WLLN and Slutsky's theorem on $\hat{\pi}_1(\theta^0)$ such that

$$\begin{aligned}
\hat{\pi}_1(\theta^0)&\xrightarrow{p}\mathbb{E}_0\left[\frac{\hat{\pi}_1(\theta^0)f_1(y_{it}|x_{it};\theta_1^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\right],\\
&=\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=1}^{G}\pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it}),\\
&=\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)\pi_1^0(f_1(y_{it}|x_{it};\theta_1^0))^2+\hat{\pi}_1(\theta^0)f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=2}^{G}\pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it}),\\
&=\pi_1^0\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it})+\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=2}^{G}\pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it}),
\end{aligned}$$

as $N,T\to\infty$. In other words, we have that

$$\hat{\pi}_1(\theta^0)\xrightarrow{p}a\pi_1^0+b,$$

as $N,T\to\infty$, where

$$a=\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it}),$$

and

$$b=\int_{\mathcal{Y}}\frac{\hat{\pi}_1(\theta^0)f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=2}^{G}\pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G}\hat{\pi}_j(\theta^0)f_j(y_{it}|x_{it};\theta_j^0)}\upsilon(dy_{it}).$$

Therefore, $\hat{\pi}_1(\theta^0)$ will be consistently estimated if and only if $a=1$ and $b=0$. It is easy to see that constant part of the bias, b, will be go to zero if and only if all component densities are infinitely distant from each other at $\theta=\theta^0$, which means that the integral of the product of any two different densities will go to zero in the limit (i.e. $\int_{\mathcal{Y}}f_g(y_{it}|x_{it};\theta_g^0)\times f_j(y_{it}|x_{it};\theta_j^0)\upsilon(dy_{it})\to 0$

as $||\theta_g^0 - \theta_j^0|| \to \infty$ for any $j \in \mathbb{G}\backslash_g$, any $g \in \mathbb{G}$, and any $x_{it} \in \mathcal{X}$). Furthermore, we also have that

$$\begin{aligned} a &= \int_{\mathcal{Y}} \frac{\hat{\pi}_1(\theta^0)(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^G \hat{\pi}_j(\theta^0) f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}), \\ &\to \int_{\mathcal{Y}} \frac{\hat{\pi}_1(\theta^0)(f_1(y_{it}|x_{it};\theta_1^0))^2}{\hat{\pi}_1(\theta^0) f_1(y_{it}|x_{it};\theta_1^0)} \upsilon(dy_{it}), \\ &\to \int_{\mathcal{Y}} f_1(y_{it}|x_{it};\theta_1^0)\upsilon(dy_{it}) = 1, \end{aligned}$$

as all densities get infinitely distant from each other. The logic is similar for all other $\hat{\pi}_g(\theta^0)$. □

A.6 Corollary 3.2

Proof. The corollary is a direct consequence of Corollary 3.1 and Theorem 3.1 under Assumption 1(*ii*) where the component density $f_g(y_{it}|x_{it};\theta_g) > 0$ for each $g \in \mathbb{G}$ such that $\hat{\pi}(\theta^0)$ does not converge in probability to π^0 unless all component densities are infinitely distant from each other. □

A.7 Proposition 3.2

Proof. As in Theorem 3.1, the WLLN implies that

$$\begin{aligned} \hat{\pi}_1(\theta^0) &\xrightarrow{p} \mathbb{E}_0\left[\frac{f_1(y_{it}|x_{it};\theta_1^0)}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)}\right], \\ &= \pi_1^0 \int_{\mathcal{Y}} \frac{(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}) + \int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=2}^G \pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}), \end{aligned}$$

as $N, T \to \infty$. If each $\pi_g^0 = \frac{1}{G}$, then we have that

$$\begin{aligned} \hat{\pi}_1(\theta) &\xrightarrow{p} \frac{1}{G}\int_{\mathcal{Y}} \frac{(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}) + \frac{1}{G}\int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=2}^G f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}), \\ &= \frac{1}{G}\int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0)\sum_{g=1}^G f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^G f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}), \\ &= \frac{1}{G}\int_{\mathcal{Y}} f_1(y_{it}|x_{it};\theta_1^0)\upsilon(dy_{it}) = \frac{1}{G} = \pi_1^0, \end{aligned}$$

so the estimated mixing weight will converge to its true value as $N, T \to \infty$. If each $\pi_g^0 \neq \frac{1}{G}$, the former result does not hold anymore and we have that

$$\hat{\pi}_1(\theta^0) \xrightarrow{p} a\pi_1^0 + b,$$

where

$$a = \int_{\mathcal{Y}} \frac{(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^{G} f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}),$$

and

$$b = \int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0) \sum_{g=2}^{G} \pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G} f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}).$$

As in Theorem 3.1, we have that $a \to 1$ and $b \to 0$ as all component densities get infinitely distant from each other. □

A.8 Proposition 3.3

Proof. If all component densities get very similar to each other, this implies that

$$\sum_{j=1}^{G} f_j(y_{it}|x_{it};\theta_j^0) \approxeq G f_1(y_{it}|x_{it};\theta_1^0).$$

Using the same logic as in Proposition 3.2, we have that

$$\begin{aligned}
\hat{\pi}_1(\theta^0) &\xrightarrow{p} \pi_1^0 \int_{\mathcal{Y}} \frac{(f_1(y_{it}|x_{it};\theta_1^0))^2}{\sum_{j=1}^{G} f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}) + \int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0) \sum_{g=2}^{G} \pi_g^0 f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{j=1}^{G} f_j(y_{it}|x_{it};\theta_j^0)} \upsilon(dy_{it}), \\
&= \pi_1^0 \int_{\mathcal{Y}} \frac{(f_1(y_{it}|x_{it};\theta_1^0))^2}{G f_1(y_{it}|x_{it};\theta_1^0)} \upsilon(dy_{it}) + \int_{\mathcal{Y}} \frac{f_1(y_{it}|x_{it};\theta_1^0) f_1(y_{it}|x_{it};\theta_1^0) \sum_{g=2}^{G} \pi_g^0}{G f_1(y_{it}|x_{it};\theta_1^0)} \upsilon(dy_{it}), \\
&= \frac{\pi_1^0}{G} \int_{\mathcal{Y}} f_1(y_{it}|x_{it};\theta_1^0) \upsilon(dy_{it}) + \frac{(1-\pi_1^0)}{G} \int_{\mathcal{Y}} f_1(y_{it}|x_{it};\theta_1^0) \upsilon(dy_{it}), \\
&= \frac{\pi_1^0}{G} + \frac{1-\pi_1^0}{G} = \frac{1}{G},
\end{aligned}$$

as $N, T \to \infty$. Hence, removing the mixing weights from the RHS of eq.(3) is identical to including informative priors of $1/G$ on each π_g^0, which will "dominate" the likelihood if the component densities are too similar to each other. □

A.9 Lemma 3.3

Proof. Unbiasedness of all three classifiers are treated separately. For notational convenience, I drop the "0" subscript in the expected value $\mathbb{E}_0$ for all remaining proofs. All expectations are taken with respect to the true joint density $f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)$ unless stated otherwise.

1. **Joint density classifier** From Definition 2 and Definition 5, we can write that

$$\mathbb{E}[f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)] > \mathbb{E}[f_j(y_{it}, x_{it}|\theta_j^0, \xi_j^0)] \Rightarrow \arg\max_{g \in \mathbb{G}} \mathbb{E}[f_g(y_{it}, x_{it}|\theta_g^0, \xi_g^0)] = z_{it}^0,$$

for any value $j \neq z_{it}^0$ if $\theta_j^0 = \theta_g^0 \Leftrightarrow j = g$ or if $\xi_j^0 = \xi_g^0 \Leftrightarrow j = g$, which is satisfied if

$\boldsymbol{\mu}_j^0 = \boldsymbol{\mu}_g^0 \Leftrightarrow j = g$. Furthermore, we can write that

$$\arg\max_{g\in\mathbb{G}} \mathbb{E}[f_g(y_{it}, x_{it}|\theta_g^0, \xi_g^0)] \Leftrightarrow \arg\max_{g\in\mathbb{G}} \mathbb{E}\left[\log\left(\frac{f_g(y_{it}, x_{it}|\theta_g^0, \xi_g^0)}{f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)}\right)\right],$$

given that $f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0) > 0$ for any $(y_{it}, x_{it}) \in \mathcal{Y} \times \mathcal{X}$ and any $z_{it}^0 \in \mathbb{G}$ by Assumption 1(ii). Therefore, we can apply Jensen's inequality on the LHS and obtain the following

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{f_g(y_{it}, x_{it}|\theta_g^0, \xi_g^0)}{f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)}\right)\right] &\leq \log\left(\mathbb{E}\left[\frac{f_j(y_{it}, x_{it}|\theta_j^0, \xi_j^0)}{f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)}\right]\right), \\
&\leq \log\left(\int_{\mathcal{Y}}\int_{\mathcal{X}} \frac{f_j(y_{it}, x_{it}|\theta_j^0, \xi_j^0)}{f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)} f_{z_{it}^0}(y_{it}, x_{it}|\theta_{z_{it}^0}^0, \xi_{z_{it}^0}^0)\upsilon(dy_{it})\nu(dx_{it})\right), \\
&\leq \log\left(\int_{\mathcal{Y}} f_j(y_{it}|x_{it}; \theta_j^0)\upsilon(dy_{it}) \int_{\mathcal{X}} f_j(x_{it}|\xi_j^0)\nu(dx_{it})\right), \\
&\leq \log(1) = 0.
\end{aligned}$$

By Assumption 1(ii) and Assumption 2(υ), the upper bound of Jensen's inequality will be reached if and only if $j = z_{it}^0$, thus implying that $\arg\max_{g\in\mathbb{G}} \mathbb{E}[f_g(y_{it}, x_{it}|\theta_g^0, \xi_g^0)] = z_{it}^0$.

2. **Euclidean distance classifier** Following the same logic as above, we have that

$$\mathbb{E}[||x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0||^2] < \mathbb{E}[||x_{it} - \boldsymbol{\mu}_j^0||^2] \Rightarrow \arg\max_{g\in\mathbb{G}} \mathbb{E}[-||x_{it} - \boldsymbol{\mu}_g||^2] = z_{it}^0,$$

for any value $j \neq z_{it}^0$. From Appendix A.20, we also have that

$$\mathbb{E}[||x_{it} - \boldsymbol{\mu}_j||^2] = \sum_{l=1}^{p} \sigma_{z_{it}^0, ll}^2 + \sum_{l=1}^{p} (a_{jz_{it}^0, l})^2 > \sum_{l=1}^{p} \sigma_{z_{it}^0, ll}^2 = \mathbb{E}[||x_{it} - \boldsymbol{\mu}_{z_{it}^0}||^2],$$

given that at least one element in $\mathbf{a}_{jz_{it}^0} = \boldsymbol{\mu}_j^0 - \boldsymbol{\mu}_{z_{it}^0}^0$ is non-zero if $j \neq z_{it}^0$.

3. **Mahalanobis distance classifier** Following the same logic as above, we have that

$$\mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}_{z_{it}^0}^0, \Sigma_{z_{it}^0}^0)] < \mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}_j^0, \Sigma_j^0)] \Rightarrow \arg\max_{g\in\mathbb{G}} \mathbb{E}[-d^2(x_{it}, \boldsymbol{\mu}_g^0, \Sigma_g^0)] = z_{it}^0,$$

for any value $j \neq z_{it}^0$. From Appendix A.22, we also have that

$$\mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}_j^0, \Sigma_j^0)] = p + \sum_{l=1}^{p} (b_{jz_{it}^0, l})^2 > p = \mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}_{z_{it}^0}^0, \Sigma_{z_{it}^0}^0)],$$

if $\Sigma_j = \Sigma_{z_{it}^0}$ and given that at least one term in the sum $\sum_{l=1}^{p} (b_{jz_{it}^0, l})^2$ is strictly positive.

□

A.10 Lemma 3.4

Proof. From Definition 2 and Definition 5, we can write that

$$\mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}})] < \mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j)] \Rightarrow \arg\max_{g \in \mathbb{G}} \mathbb{E}[-d^2(x_{it}, \boldsymbol{\mu}_g, \Sigma_g)] = z^0_{it},$$

for any $j \neq z^0_{it}$. Using the results from Appendix A.21 and Appendix A.22, we have that

$$\mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}})] = p,$$

and that

$$\begin{aligned}\mathbb{E}[d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j)] &= p + \sum_{l=1}^{p}\sum_{m \leq l}(v_{jz^0_{it},lm})^2 + \sum_{l=1}^{p}(b_{jz^0_{it},l})^2 - 2\sum_{l=1}^{p} v_{jz^0_{it},ll}, \\ &= p + \sum_{l=1}^{p}\sum_{m \leq l}(v_{jz^0_{it},lm})^2 - 2\sum_{l=1}^{p} v_{jz^0_{it},ll},\end{aligned}$$

if $\boldsymbol{\mu}^0_j = \boldsymbol{\mu}^0_{z^0_{it}}$. This implies that the Mahalanobis distance classifier will be unbiased if

$$2\sum_{l=1}^{p} v_{jz^0_{it},ll} < \sum_{l=1}^{p}\sum_{m \leq l}(v_{jz^0_{it},lm})^2,$$

for any $j \neq z^0_{it}$. This inequality will be satisfied if p is sufficiently large given that the expression on the RHS is a sum of $p(p+1)/2$ positive terms while the expression on the LHS is a sum of p terms that can be either positive, negative, or null depending on the sign of the diagonal elements of $A_{jz^0_{it}} = W_{z^0_{it}} - W_j$. □

A.11 Theorem 3.2

Proof. From Definition 5(c), we have that

$$\mathbb{P}[\cup_{g=1}^{G}(z^M_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0) \neq z^0_{itg})] \Leftrightarrow \mathbb{P}[d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}}) \geq d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j)]$$

for at least one value $j \neq z^0_{it}$. Hence, we can also write

$$\begin{aligned}\mathbb{P}[\cup_{g=1}^{G}(z^M_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0) \neq z^0_{itg})] &= 1 - \mathbb{P}[\cap_{j \neq z^0_{it}}(d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}}) < d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j))], \\ &= 1 - \prod_{j \neq z^0_{it}} \mathbb{P}[d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}}) < d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j)], \\ &= 1 - \prod_{j \neq z^0_{it}} (1 - \mathbb{P}[d^2(x_{it}, \boldsymbol{\mu}^0_{z^0_{it}}, \Sigma^0_{z^0_{it}}) \geq d^2(x_{it}, \boldsymbol{\mu}^0_j, \Sigma^0_j)]),\end{aligned}$$

where the second equality comes from the fact that $\boldsymbol{\mu}^0_j$ and Σ^0_j are taken as given for any $j \in \mathbb{G}$.

Using the result from Appendix A.23, we have that

$$\mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)] \leq \frac{\mathbb{E}[d^2(x_{it}, z_{it}^0)] - \text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]}{\mathbb{E}[d^2(x_{it}, j)]},$$

where $d^2(x_{it}, j) = d^2(x_{it}, \boldsymbol{\mu}_j^0, \Sigma_j^0)$ for any $j \in \mathbb{G}$. Note that $\text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]$ is always negative since $\mathbb{P}[\mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]} = 0] = 1$ as $d^2(x_{it}, j) \to \infty$. Note also that it is independent of p given that the distribution of both $d^2(x_{it}, j) - \mathbb{E}[d^2(x_{it}, j)]$ and $\mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]} - \mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]$ are independent of p. To see this, we can write

$$\mathbb{P}[d^2(x_{it}, j) \geq a\mathbb{E}[d^2(x_{it}, j)]] = \mathbb{P}[d^2(x_{it}, j)a^{-1} - \mathbb{E}[d^2(x_{it}, j)] \geq 0] \leq \frac{\mathbb{E}[d^2(x_{it}, j)]}{a\mathbb{E}[d^2(x_{it}, j)]} = \frac{1}{a}$$

for any $a > 0$ using the standard Markov's inequality. Furthermore, we have that

$$\mathbb{P}[b\mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]} - \mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)] \geq 0] = \mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)],$$

for any $b \geq \mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]$. Consequently, there exists a finite value $M \in \mathbb{R}$ such that $|\text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]| \leq M$ as $p \to \infty$. Applying the result from Appendix A.22 and the fact that $\mathbb{E}[d^2(x_{it}, z_{it}^0)] = p$ to the above inequality, we have that

$$\mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)] \leq \frac{p - \text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]}{p + \sum_{l=1}^{p}[\sum_{m \leq l}(v_{jz_{it}^0, lm})^2 + (b_{jz_{it}^0, l})^2 - 2v_{jz_{it}^0, ll}]}.$$

As $p \to \infty$, the last term inside the bracket in the denominator gets dominated by the other term in the denominator since $v_{jz_{it}^0, ll}$ can be either negative, positive, or null. Therefore, there exists a constant $0 < k < \infty$ such that

$$p + \sum_{l=1}^{p}\left[\sum_{m \leq l}(v_{jz_{it}^0, lm})^2 + (b_{jz_{it}^0, l})^2 - 2v_{jz_{it}^0, ll}\right] \geq p + \frac{kp(p+1)}{2} + kp = p\left(\frac{2 + pk + 3k}{2}\right),$$

for any $p > p^*$, where p^* is a large positive discrete number. From this, we can also write

$$\begin{aligned}
\frac{p - \text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]}{p + \sum_{l=1}^{p}[\sum_{m \leq l}(v_{jz_{it}^0, lm})^2 + (b_{jz_{it}^0, l})^2 - 2v_{jz_{it}^0, ll}]} &\leq \frac{p}{p\left(\frac{2+pk+3k}{2}\right)} - \frac{\text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]}{p\left(\frac{2+pk+3k}{2}\right)}, \\
&= \frac{2}{2 + pk + 3k} - \frac{2\text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]}{p(2 + pk + 3k)}, \\
&= O(p^{-1}) + O_p(p^{-2}) = O_p(p^{-1}),
\end{aligned}$$

given that $\text{Cov}[d^2(x_{it}, j), \mathbb{1}_{[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)]}]$ is a random sequence in p, hence leading to

$$\mathbb{P}[d^2(x_{it}, z_{it}^0) \geq d^2(x_{it}, j)] = O_p(p^{-1}).$$

This implies that there exists a constant $0 < \tilde{k} < \infty$ such that

$$\mathbb{P}[\cup_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) \neq z_{itg}^0)] \leq 1 - (1 - \tilde{k}p^{-1})^{(G-1)},$$

for p sufficiently large. If G is a finite constant, then $\mathbb{P}[\cup_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) \neq z_{itg}^0)] = O_p(p^{-1})$ since

$$p(1 - (1 - \tilde{k}p^{-1})^{(G-1)}) = p - p\left(1 - \frac{(G-1)\tilde{k}}{p} + O(p^{-2})\right) \to (G-1)\tilde{k} \text{ as } p \to \infty.$$

□

A.12 Theorem 3.3

Proof. From Definition 4 and the proof of Theorem 3.2, we can write that

$$\begin{aligned}
\mathbb{P}[\cup_{i=1}^{N}\cup_{t=1}^{T}\cup_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) \neq z_{itg}^0)] &= 1 - \mathbb{P}[\cap_{i=1}^{N}\cap_{t=1}^{T}\cap_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{itg}^0)], \\
&= 1 - \prod_{i=1}^{N}\prod_{t=1}^{T}\mathbb{P}[\cap_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{itg}^0)], \\
&= 1 - \prod_{i=1}^{N}\prod_{t=1}^{T}(1 - \mathbb{P}[\cup_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) \neq z_{itg}^0)]), \\
&\leq 1 - (1 - \tilde{k}p^{-1})^{NT(G-1)},
\end{aligned}$$

for p sufficiently large. The second equality comes from the fact that all vectors x_{it} are independent from each other by Assumption 2(v). Relaxing this assumption to include serial correlation in x_{it} would lead to

$$\mathbb{P}[\cap_{i=1}^{N}\cap_{t=1}^{T}\cap_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{itg}^0)] = \prod_{i=1}^{N}\prod_{t=1}^{T}\mathbb{P}[\cap_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{itg}^0) | \cap_{j<t}\cap_{g=1}^{G}(z_{ij(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{ijg}^0)],$$

which is different from $\prod_{i=1}^{N}\prod_{t=1}^{T}\mathbb{P}[\cap_{g=1}^{G}(z_{it(g)}^{M}(\boldsymbol{\mu}^0, \Sigma^0) = z_{itg}^0)]$ if x_{it}. However, relaxing this independence assumption does not necessarily invalidate the conclusion of the theorem.

If we assume that $p = a(NT)^b$ and that $G = \tilde{a}(NT)^{\tilde{b}}$ with $(a, b, \tilde{a}, \tilde{b}) \in \mathbb{R}_{>0}^4$, then

$$\lim_{N,T\to\infty} 1 - (1 - \tilde{k}(a(NT)^b)^{-1})^{(\tilde{a}(NT)^{\tilde{b}}-1)NT} = \begin{cases} 1 & \text{if } 0 < b < \tilde{b} + 1, \\ 1 - \exp^{-\tilde{a}\tilde{k}/a} & \text{if } b = \tilde{b} + 1, \\ 0 & \text{if } b > \tilde{b} + 1, \end{cases}$$

which implies that $z_{itg}^{M}(\boldsymbol{\mu}^0, \Sigma^0)$ is a uniformly consistent classifier if the number of covariates p increases at a strictly higher-order rate than the number of groups relative to the sample size. For instance, if $\tilde{b} = 1$, then the number of groups is proportional to the sample size, and the number of covariates has to increase at a rate that is more than twice the growth rate in the sample size for $z_{itg}^{M}(\boldsymbol{\mu}^0, \Sigma^0)$ to be uniformly consistent. Note that if $\tilde{b} > 1$, then the average number of observations within each group goes to zero in the limit. In practice, it is important to impose $\tilde{b} \leq 1$ so that the

number of groups does not increase at a faster rate than the sample size. □

A.13 Corollary 3.3

Proof. This corollary is a special case of Theorem 3.3 where G is constant. If $\tilde{b} = 0$, then uniform consistency of $z^M_{itg}(\boldsymbol{\mu}^0, \Sigma^0)$ requires that p increases at a faster rate than the sample size. Note that if $\mathbb{P}[d^2(x_{it}, z^0_{it}) \geq d^2(x_{it}, j)] = O_p(p^{-\lambda})$ with $\lambda > 1$, then $z^M_{itg}(\boldsymbol{\mu}^0, \Sigma^0)$ would be a uniformly consistent classifier when $b > 1/\lambda$. This would allow the number of observations to increase at a higher rate than the number of covariates when G is constant. □

A.14 Corollary 3.4

Proof. From Definition 3, we can write that

$$\begin{aligned}\hat{E}_{NTp}(\theta^0, \xi^0) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G}\frac{\mathbb{1}[z^0_{itg} \neq z^M_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0)]}{2NT}, \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\mathbb{1}[z^0_{it} \neq z^M_{it}(\boldsymbol{\mu}^0, \Sigma^0)]}{NT} \xrightarrow{p} \mathbb{P}[z^0_{it} \neq z^M_{it}(\boldsymbol{\mu}^0, \Sigma^0)] \text{ as } N, T \to \infty.,\end{aligned}$$

which is equivalent to $\mathbb{P}[\cup_{g=1}^{G}(z^M_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0) \neq z^0_{itg})]$. By Theorem 3.2, we know that

$$\mathbb{P}[\cup_{g=1}^{G}(z^M_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0) \neq z^0_{itg})] \leq 1 - (1 - \tilde{k}p^{-1})^{(G-1)}.$$

Following a similar logic as in the proof of Theorem 3.3 and assuming that $G = ap^b$ with $(a, b) \in \mathbb{R}^2_{>0}$, we obtain that

$$\lim_{p\to\infty} 1 - (1 - \tilde{k}p^{-1})^{(ap^b - 1)} = 0 \quad \text{if } 0 < b < 1,$$

which means that the number of covariates has to grow at a strictly higher rate than the number of groups. □

A.15 Theorem 3.4

Proof. Using the same strategy as the proof of Theorem 3.2 and the results from Appendix A.20 and Appendix A.23, we can write that

$$\begin{aligned}\mathbb{P}[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 \geq ||x_{it} - \boldsymbol{\mu}^0_j||^2] &\leq \frac{\mathbb{E}[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2] - \text{Cov}[||x_{it} - \boldsymbol{\mu}^0_j||^2, \mathbb{1}_{[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 \geq ||x_{it} - \boldsymbol{\mu}^0_j||^2]}]}{\mathbb{E}[||x_{it} - \boldsymbol{\mu}^0_j||^2]}, \\ &\leq \frac{\sum_{l=1}^{p}\sigma^2_{z^0_{it},ll}}{\sum_{l=1}^{p}\sigma^2_{z^0_{it},ll} + \sum_{l=1}^{p}(a_{jz^0_{it},l})^2} - \frac{\text{Cov}[||x_{it} - \boldsymbol{\mu}^0_j||^2, \mathbb{1}_{[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 \geq ||x_{it} - \boldsymbol{\mu}^0_j||^2]}]}{\sum_{l=1}^{p}\sigma^2_{z^0_{it},ll} + \sum_{l=1}^{p}(a_{jz^0_{it},l})^2}, \\ &\leq \frac{1}{1 + \tilde{a}_{jz^0_{it}}} + O_p(p^{-1}) \xrightarrow{p} c_{j,it} \geq 0 \text{ as } p \to \infty,\end{aligned}$$

where $\tilde{a}_{jz^0_{it}} = \frac{||\boldsymbol{\mu}^0_j - \boldsymbol{\mu}^0_{z^0_{it}}||^2}{\text{tr}(\Sigma^0_{z^0_{it}})}$, and where $c_{j,it}$ is a positive random variable whose distribution will depend on the distribution of the true mean values $\boldsymbol{\mu}^0_j$ and $\boldsymbol{\mu}^0_{z^0_{it}}$, and on the distribution of the variance $\sigma^2_{z^0_{it},ll}$. Therefore, we can write that

$$\begin{aligned}
\mathbb{P}[\cup_{g=1}^{G}(z^E_{it(g)}(\boldsymbol{\mu}^0) \neq z^0_{itg})] &= \mathbb{P}[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 \geq ||x_{it} - \boldsymbol{\mu}^0_j||^2] \text{ for at least one } j \neq z^0_{it}, \\
&= 1 - \prod_{j \neq z^0_{it}} \mathbb{P}[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 < ||x_{it} - \boldsymbol{\mu}^0_j||^2], \\
&= 1 - \prod_{j \neq z^0_{it}} (1 - \mathbb{P}[||x_{it} - \boldsymbol{\mu}^0_{z^0_{it}}||^2 \geq ||x_{it} - \boldsymbol{\mu}^0_j||^2]), \\
&\leq 1 - \prod_{j \neq z^0_{it}} \left(1 - \frac{1}{1 + \tilde{a}_{jz^0_{it}}} - O_p(p^{-1})\right) \xrightarrow{p} \bar{c}_{it} \text{ as } p \to \infty,
\end{aligned}$$

where $\bar{c}_{it} \geq 0$. Therefore, we have that $\mathbb{P}[\cup_{g=1}^{G}(z^E_{it(g)}(\boldsymbol{\mu}^0, \Sigma^0) \neq z^0_{itg})] \xrightarrow{p} c_{it}$ where $c_{it} \geq 0$ is also a random variable whose value will vary between zero and $\bar{c}_{it}$ if $\bar{c}_{it} < 1$. □

A.16 Corollary 3.5

Proof. Akin to the proof of Corollary 3.4, we can write that

$$\hat{E}_{NTp}(\theta^0, \xi^0) = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\mathbb{1}[z^0_{it} \neq z^E_{it}(\boldsymbol{\mu}^0)]}{NT} \xrightarrow{p} \mathbb{P}[z^0_{it} \neq z^E_{it}(\boldsymbol{\mu}^0)] \text{ as } N, T \to \infty.$$

By Definition 5(b) and as shown in the proof of Theorem 3.4, we know that

$$\mathbb{P}[z^0_{it} \neq z^E_{it}(\boldsymbol{\mu}^0)] \leq 1 - \prod_{j \neq z^0_{it}} \left(1 - \frac{1}{1 + \tilde{a}_{jz^0_{it}}} - O_p(p^{-1})\right) \xrightarrow{p} \bar{c}_{it} \text{ as } p \to \infty.$$

Therefore, we have that $0 \leq \mathbb{P}[z^0_{it} \neq z^E_{it}(\boldsymbol{\mu}^0)] \leq \bar{c}_{it}$ as $p \to \infty$, where $\bar{c}_{it}$ will be equal to zero if and only if $\tilde{a}_{jg} \to \infty$ as $p \to \infty$ for all possible pairs $(g, j) \in \mathbb{G} \times \mathbb{G} \backslash g$. □

A.17 Theorem 3.5

Proof. Let's look at the CEM algorithm equipped with the joint density classifier. The first M-step of the algorithm can be written as follows :

$$\begin{aligned}
\hat{\theta}^{(1)}_{NT} &:= \arg\max_{\theta \in \Theta} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} z^D_{itg}(\theta^{(0)}, \xi^{(0)}) \log(f_g(y_{it}|x_{it}, \theta)), \\
\hat{\xi}^{(1)}_{NT} &:= \arg\max_{\xi \in \Xi} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} z^D_{itg}(\theta^{(0)}, \xi^{(0)}) \log(f_g(x_{it}|\xi)).
\end{aligned}$$

Note that the first equation can be rewritten as follows :

$$\hat{\theta}^{(1)}_{NT} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \mathbb{1}_{[z^D_{it(g)}(\theta^{(0)},\xi^{(0)})=z^0_{itg}]} z^D_{itg}(\theta^{(0)},\xi^{(0)}) \log(f_g(y_{it}|x_{it},\theta)) + \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \mathbb{1}_{[z^D_{it(g)}(\theta^{(0)},\xi^{(0)})\neq z^0_{itg}]} z^D_{itg}(\theta^{(0)},\xi^{(0)}) \log(f_g(y_{it}|x_{it},\theta)). \tag{16}$$

An analogous equation can be formulated for $\hat{\xi}^{(1)}_{NT}$. Given that $z^D_{itg}(\theta,\xi)$ is binary and that the joint density $f(y_{it},x_{it}|\theta,\xi)$ is continuous in both θ and ξ by definition of the joint density (and similarly for $z^M_{itg}(\theta,\xi)$), there exists a set $(\theta^*,\xi^*) \neq (\theta^0,\xi^0)$ in the neighborhood of (θ^0,ξ^0) such that $z^D_{itg}(\theta^0,\xi^0) = z^D_{itg}(\theta^*,\xi^*)$. If $\theta^{(0)} = \theta^*$ and if $\xi^{(0)} = \xi^*$, there exists a sample size $(NT)^*$ and a number of covariates p^* such that

$$z^D_{itg}(\hat{\theta}^{(1)}_{NT},\hat{\xi}^{(1)}_{NT}) = z^D_{itg}(\hat{\theta}^{*(1)}_{NT},\hat{\xi}^{*(1)}_{NT}),$$

for all sample sizes $(NT) > (NT)^*$ and all values of $p > p^*$, and where

$$\hat{\theta}^{*(1)}_{NT} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \mathbb{1}_{[z^D_{it(g)}(\theta^{(0)},\xi^{(0)})=z^0_{itg}]} z^D_{itg}(\theta^{(0)},\xi^{(0)}) \log(f_g(y_{it}|x_{it},\theta)),$$

$$\hat{\xi}^{*(1)}_{NT} = \arg\max_{\xi\in\Xi} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \mathbb{1}_{[z^D_{it(g)}(\theta^{(0)},\xi^{(0)})=z^0_{itg}]} z^D_{itg}(\theta^{(0)},\xi^{(0)}) \log(f_g(x_{it}|\xi)).$$

The equality $z^D_{itg}(\hat{\theta}^{(1)}_{NT},\hat{\xi}^{(1)}_{NT}) = z^D_{itg}(\hat{\theta}^{*(1)}_{NT},\hat{\xi}^{*(1)}_{NT})$ is valid given that the second term of eq.(16) becomes arbitrarily small for any $(NT) > (NT)^*$ and any $p > p^*$ by consistency of $z^D_{itg}(\theta,\xi)$. If $z^D_{itg}(\hat{\theta}^{(1)}_{NT},\hat{\xi}^{(1)}_{NT}) = z^D_{itg}(\hat{\theta}^{*(1)}_{NT},\hat{\xi}^{*(1)}_{NT}) = z^D_{itg}(\theta^{(0)},\xi^{(0)}) = z^D_{itg}(\theta^0,\xi^0)$, then the CEM algorithm has converged to a stationary point given that all estimated group memberships remained constant between two consecutive iterations. Therefore, we can write that

$$\begin{aligned}
\hat{\theta}^{(2)}_{NT} &= \arg\max_{\theta\in\Theta} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} z^D_{itg}(\theta^0,\xi^0)\log(f_g(y_{it}|x_{it},\theta)),\\
&\xrightarrow{p} \arg\max_{\theta\in\Theta} \sum_{g=1}^{G} \mathbb{E}\left[z^D_{itg}(\theta^0,\xi^0)\log(f_g(y_{it}|x_{it},\theta))\right] \text{ as } N,T,p\to\infty,\\
&= \arg\max_{\theta\in\Theta} \sum_{g=1}^{G} \mathbb{E}[z^D_{itg}(\theta^0,\xi^0)]\mathbb{E}[\log(f_g(y_{it}|x_{it},\theta))] + \sum_{g=1}^{G} Cov[z^D_{itg}(\theta^0,\xi^0),\log(f_g(y_{it}|x_{it},\theta))],\\
&= \arg\max_{\theta\in\Theta} \sum_{g=1}^{G} \mathbb{P}[z^D_{itg}(\theta^0,\xi^0)=1]\mathbb{E}[\log(f_g(y_{it}|x_{it},\theta))] + Cov[\sum_{g=1}^{G} z^D_{itg}(\theta^0,\xi^0),\log(f_g(y_{it}|x_{it},\theta))],\\
&= \arg\max_{\theta\in\Theta} \mathbb{E}[\log(f_{z^0_{it}}(y_{it}|x_{it},\theta))] = \theta^0,
\end{aligned}$$

where I use the fact that $z^D_{itg}(\theta^0, \xi^0) = 1$ w.p.a. 1 if and only if $g = z^0_{it}$, and $z^D_{itg}(\theta^0, \xi^0) = 0$ w.p.a. 1 if and only if $g \neq z^0_{it}$ by consistency of $z^D_{itg}(\theta^0, \xi^0)$. Furthermore, the covariance term is always equal to zero given that $\sum_{g=1}^{G} z^D_{itg}(\theta^0, \xi^0) = 1$ for any $(\theta^0, \pi^0) \in \Theta \times \Pi$ by construction. Note that the probability $\mathbb{P}[z^D_{itg}(\theta^0, \xi^0) = 1] \neq \{0, 1\}$ in finite samples even if $z^D_{itg}(\theta^0, \xi^0)$ is binary, thus leading to a finite-sample misclassification bias.

If $z^D_{itg}(\hat{\theta}^{(1)}_{NT}, \hat{\xi}^{(1)}_{NT}) = z^D_{itg}(\hat{\theta}^{*(1)}_{NT}, \hat{\xi}^{*(1)}_{NT}) \neq z^D_{itg}(\theta^{(0)}, \xi^{(0)}) = z^D_{itg}(\theta^0, \xi^0)$, then the CEM algorithm *has not* converged to a stationary point and other iterations of the algorithm are needed to reach a stationary point. Since the MCL function never decreases between two consecutive iterations of the CEM algorithm and its expected value is bounded from above by Assumption 2(i), there exists a discrete number $k < \infty$ such that $z^D_{itg}(\hat{\theta}^{(k)}_{NT}, \hat{\xi}^{(k)}_{NT}) = z^D_{itg}(\hat{\theta}^{(k+1)}_{NT}, \hat{\xi}^{(k+1)}_{NT})$, which implies that $\hat{\theta}^{(k+2)}_{NT} = \hat{\theta}^{(k+1)}_{NT}$ and that $\hat{\xi}^{(k+2)}_{NT} = \hat{\xi}^{(k+1)}_{NT}$, and the algorithm has converged to a stationary point such that $(\hat{\theta}^{(k+2)}_{NT}, \hat{\xi}^{(k+2)}_{NT}) \xrightarrow{p} (\theta^0, \xi^0)$ as $N, T \to \infty$.

If $\theta^{(0)}$ and $\xi^{(0)}$ are too distant from θ^0 and ξ^0 respectively, then the CEM algorithm may converge to a stationary point such that $\hat{\theta}^{(k+2)}_{NT} = \hat{\theta}^{(k+1)}_{NT} \xrightarrow{p} \bar{\theta} \neq \theta^0$ and such that $\hat{\xi}^{(k+2)}_{NT} = \hat{\xi}^{(k+1)}_{NT} \xrightarrow{p} \bar{\xi} \neq \xi^0$ as $N, T, p \to \infty$. In this case, other values of $\theta^{(0)}$ and $\xi^{(0)}$ have to be assessed to ensure consistency of the estimation procedure, as it is well known in the literature on the EM and the CEM algorithms (Frühwirth-Schnatter, 2006; McLachlan and Peel, 2000). □

A.18 Theorem 3.6

Proof. The derivation of the asymptotic distribution of the CEM algorithm is identical to the derivation of the asymptotic distribution of any ML estimator with cluster-robust variance but at the group level with the inclusion of the binary classifier $z_{itg}(\theta, \xi)$ in the definition of $s_{ig}(\theta)$. See Section 12.3 of Wooldridge (2010) for a general proof. The scaling factor $\sqrt{n_g(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})}$ comes from the fact that group membership is not restricted over time so that each unit i does not equally contribute to the estimation of the parameters $\hat{\theta}^{(k)}_g$. □

A.19 Example of inconsistency of θ when the mixing weights do not converge

Let Assumption 1 hold, and let the true mixture density be defined as follows

$$f(y_i|\theta^0, \pi^0) := \pi^0_1 \lambda^0_1 \exp(-\lambda^0_1 y_i) + \pi^0_2 \lambda^0_2 \exp(-\lambda^0_2 y_i),$$

with $\pi^0_1 = 1 - \pi^0_2$. Then, the ML estimates λ_1 and λ_2 will be asymptotically biased if π_1 does not converge to π^0_1.

Proof. Under Assumption 1(vi), we have that the first-order condition

$$\left.\frac{\partial \mathbb{E}[\log(f(y_i|\theta, \pi))]}{\partial \theta}\right|_{\theta=\theta^0} = 0,$$

is necessary (but insufficient) for θ to converge to θ^0 as the sample size tends to infinity. If we

assume that $\pi = \tilde{\pi} \neq \pi^0$, then we have that

$$\begin{aligned}
\left.\frac{\partial \mathbb{E}[\log(f(y_i|\theta,\tilde{\pi}))]}{\partial\theta}\right|_{\theta=\theta^0} &= \int_{\mathcal{Y}} \frac{f(y_i|\theta^0,\pi^0)}{f(y_i|\theta^0,\tilde{\pi})} \left.\frac{\partial f(y_i|\theta,\tilde{\pi})}{\partial\theta}\right|_{\theta=\theta^0} dy_i, \\
&= \int_{\mathcal{Y}} \frac{\pi_1^0\lambda_1^0\exp(-\lambda_1^0 y_i) + \pi_2^0\lambda_2^0\exp(-\lambda_2^0 y_i)}{\tilde{\pi}_1\lambda_1^0\exp(-\lambda_1^0 y_i) + \tilde{\pi}_2\lambda_2^0\exp(-\lambda_2^0 y_i)} [\tilde{\pi}_1\exp(-\lambda_1^0 y_i) + \tilde{\pi}_2\exp(-\lambda_2^0 y_i)] dy_i \\
&\quad - \int_{\mathcal{Y}} y_i(\pi_1^0\lambda_1^0\exp(-\lambda_1^0 y_i) + \pi_2^0\lambda_2^0\exp(-\lambda_2^0 y_i)) dy_i.
\end{aligned}$$

Reorganizing the terms then leads to

$$\left.\frac{\partial \mathbb{E}[\log(f(y_i|\theta,\tilde{\pi}))]}{\partial\theta}\right|_{\theta=\theta^0} = \int_{\mathcal{Y}} \frac{\pi_1^0\lambda_1^0 + \pi_2^0\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))}{\tilde{\pi}_1\lambda_1^0 + \tilde{\pi}_2\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))} [\tilde{\pi}_1\exp(-\lambda_1^0 y_i) + \tilde{\pi}_2\exp(-\lambda_2^0 y_i)] dy_i - \frac{\pi_1^0}{\lambda_1^0} - \frac{\pi_2^0}{\lambda_2^0},$$

given that $ay_i\exp(-ay_i)$ integrates to $\frac{1}{a}$ when y_i goes from 0 to $+\infty$ for any $a \in \mathbb{R}$. If $\tilde{\pi} = \pi^0$, then

$$\begin{aligned}
\left.\frac{\partial \mathbb{E}[\log(f(y_i|\theta,\pi^0))]}{\partial\theta}\right|_{\theta=\theta^0} &= \int_{\mathcal{Y}} [\pi_1^0\exp(-\lambda_1^0 y_i) + \pi_2^0\exp(-\lambda_2^0 y_i)] dy_i - \frac{\pi_1^0}{\lambda_1^0} - \frac{\pi_2^0}{\lambda_2^0}, \\
&= \pi_1^0 \int_{\mathcal{Y}} \exp(-\lambda_1^0 y_i) dy_i + \pi_2^0 \int_{\mathcal{Y}} \exp(-\lambda_2^0 y_i) dy_i - \frac{\pi_1^0}{\lambda_1^0} - \frac{\pi_2^0}{\lambda_2^0}, \\
&= \frac{\pi_1^0}{\lambda_1^0} + \frac{\pi_2^0}{\lambda_2^0} - \frac{\pi_1^0}{\lambda_1^0} - \frac{\pi_2^0}{\lambda_2^0} = 0,
\end{aligned}$$

which satisfies the FOC for consistency. If $\tilde{\pi} \neq \pi^0$, then we need to have

$$\int_{\mathcal{Y}} \frac{\pi_1^0\lambda_1^0 + \pi_2^0\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))}{\tilde{\pi}_1\lambda_1^0 + \tilde{\pi}_2\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))} [\tilde{\pi}_1\exp(-\lambda_1^0 y_i) + \tilde{\pi}_2\exp(-\lambda_2^0 y_i)] dy_i = \frac{\pi_1^0}{\lambda_1^0} + \frac{\pi_2^0}{\lambda_2^0},$$

to ensure that the true parameter values λ_1^0 and λ_2^0 are located at a maximum of the expected log likelihood function. By distributing the right parenthesis and reorganizing the terms, we get to this condition :

$$\int_{\mathcal{Y}} \frac{\tilde{\pi}_1\pi_1^0\lambda_1^0\exp(-\lambda_1^0 y_i) + (\tilde{\pi}_1\pi_2^0\lambda_2^0 + \tilde{\pi}_2\pi_1^0\lambda_1^0)\exp(-\lambda_2^0 y_i) + \tilde{\pi}_2\pi_2^0\lambda_2^0\exp(y_i(\lambda_1^0-2\lambda_2^0))}{\tilde{\pi}_1\lambda_1^0 + \tilde{\pi}_2\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))} dy_i = \frac{\pi_1^0}{\lambda_1^0} + \frac{\pi_2^0}{\lambda_2^0}.$$

We can integrate each term on the LHS with y_i going from 0 to ∞ using an appropriate change in variable. For instance, if we set $c_i = \tilde{\pi}_2\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0))$, we have that

$$\frac{dc_i}{dy_i} = (\lambda_1^0-\lambda_2^0)\tilde{\pi}_2\lambda_2^0\exp(y_i(\lambda_1^0-\lambda_2^0)) = (\lambda_1^0-\lambda_2^0)c_i,$$

which implies that

$$\exp(-\lambda_1^0 y_i) = \left(\frac{c_i}{\tilde{\pi}_2\lambda_2^0}\right)^{-\frac{\lambda_1^0}{\lambda_1^0-\lambda_2^0}}.$$

Hence, we get that the first term in the integral is equal to :

$$\begin{aligned}
\int_{\mathcal{Y}} \frac{\tilde{\pi}_1 \pi_1^0 \lambda_1^0 \exp(-\lambda_1^0 y_i)}{\tilde{\pi}_1 \lambda_1^0 + \tilde{\pi}_2 \lambda_2^0 \exp(y_i(\lambda_1^0 - \lambda_2^0))} dy_i &= \int_{\mathcal{C}} \frac{\tilde{\pi}_1 \pi_1^0 \lambda_1^0}{(\tilde{\pi}_1 \lambda_1^0 + c_i)(\lambda_1^0 - \lambda_2^0) c_i} \left(\frac{\tilde{\pi}_2 \lambda_2^0}{c_i} \right)^{\frac{\lambda_1^0}{\lambda_1^0 - \lambda_2^0}} dc_i, \\
&= \frac{\tilde{\pi}_1 \pi_1^0 \lambda_1^0 (\tilde{\pi}_2 \lambda_2^0)^{\frac{\lambda_1^0}{\lambda_1^0 - \lambda_2^0}}}{\lambda_1^0 - \lambda_2^0} \int_{\mathcal{C}} \left[(\tilde{\pi}_1 \lambda_1^0 + c_i) c_i^{\frac{\lambda_1^0}{\lambda_1^0 - \lambda_2^0} + 1} \right]^{-1} dc_i, \\
&= \frac{\tilde{\pi}_1 \pi_1^0 \lambda_1^0 (\tilde{\pi}_2 \lambda_2^0)^{\frac{\lambda_1^0}{\lambda_1^0 - \lambda_2^0}}}{\lambda_1^0 - \lambda_2^0} \left[\frac{(\lambda_1^0 - \lambda_2^0)(\tilde{\pi}_2 \lambda_2^0)^{\frac{-\lambda_1^0}{\lambda_1^0 - \lambda_2^0}} \, {}_2F_1(1, \frac{-\lambda_1^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0})}{\tilde{\pi}_1 (\lambda_1^0)^2} \right], \\
&= \frac{\pi_1^0}{\lambda_1^0} \times {}_2F_1 \left(1, \frac{-\lambda_1^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right),
\end{aligned}$$

where the third equality comes from the fact that $y_i \in [0, \infty] \Rightarrow c_i \in [\tilde{\pi}_2 \lambda_2^0, 0]$ if $(\lambda_1^0 - \lambda_2^0) < 0$, and where ${}_2F_1(a, b, c, z)$ refers to the hypergeometric function. It is defined for $|z| < 1$, which implies that

$$\tilde{\pi}_2 \lambda_2^0 < \tilde{\pi}_1 \lambda_1^0 \Rightarrow \tilde{\pi}_2 < \tilde{\pi}_1,$$

since we just assumed that $(\lambda_1^0 - \lambda_2^0) < 0$. Using the same strategy, we can show that the integral of the second and the third terms in the original integral are respectively equal to

$$\begin{aligned}
\int_{\mathcal{Y}} \frac{(\tilde{\pi}_1 \pi_2^0 \lambda_2^0 + \tilde{\pi}_2 \pi_1^0 \lambda_1^0) \exp(-\lambda_2^0 y_i)}{\tilde{\pi}_1 \lambda_1^0 + \tilde{\pi}_2 \lambda_2^0 \exp(y_i(\lambda_1^0 - \lambda_2^0))} dy_i &= \left(\frac{\pi_2^0}{\lambda_1^0} + \frac{\tilde{\pi}_2 \pi_1^0}{\tilde{\pi}_1 \lambda_2^0} \right) \times {}_2F_1 \left(1, \frac{-\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{\lambda_1^0 - 2\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right), \\
\int_{\mathcal{Y}} \frac{\tilde{\pi}_2 \pi_2^0 \lambda_2^0 \exp(y_i(\lambda_1^0 - 2\lambda_2^0))}{\tilde{\pi}_1 \lambda_1^0 + \tilde{\pi}_2 \lambda_2^0 \exp(y_i(\lambda_1^0 - \lambda_2^0))} dy_i &= \frac{\pi_2^0}{\lambda_2^0} \times \frac{\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0 (2 - \lambda_1^0 / \lambda_2^0)} \times {}_2F_1 \left(1, \frac{\lambda_1^0 - 2\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{2\lambda_1^0 - 3\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right).
\end{aligned}$$

For the first-order condition to be satisfied, it requires that

$$\begin{aligned}
&{}_2F_1 \left(1, \frac{-\lambda_1^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right) = 1, \\
&\left(\frac{\pi_2^0}{\lambda_1^0} + \frac{\tilde{\pi}_2 \pi_1^0}{\tilde{\pi}_1 \lambda_2^0} \right) \times {}_2F_1 \left(1, \frac{-\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{\lambda_1^0 - 2\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right) = 0, \\
&\frac{\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0 (2 - \lambda_1^0 / \lambda_2^0)} \times {}_2F_1 \left(1, \frac{\lambda_1^0 - 2\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{2\lambda_1^0 - 3\lambda_2^0}{\lambda_1^0 - \lambda_2^0}, \frac{-\tilde{\pi}_2 \lambda_2^0}{\tilde{\pi}_1 \lambda_1^0} \right) = 1,
\end{aligned}$$

since no hypergeometric function contains π_1^0 nor π_2^0 as arguments. However, the second equality can never be true given that

$$\frac{\pi_2^0}{\lambda_1^0} + \frac{\tilde{\pi}_2 \pi_1^0}{\tilde{\pi}_1 \lambda_2^0} = 0 \Rightarrow \frac{\lambda_2^0}{\lambda_1^0} = -\frac{\tilde{\pi}_2 \pi_1^0}{\tilde{\pi}_1 \pi_2^0},$$

which is only possible if both ratios are equal to zero, a contradiction since all elements on each

side of the equation are strictly positive by assumption. Moreover, we have that

$$ {}_2F_1(a,b,c,z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{j=0}^{\infty} \frac{\Gamma(a+j)\Gamma(b+j)}{\Gamma(c+j)j!} z^j, $$

where $\Gamma(\cdot)$ denotes the Gamma function, and which is never equal to zero for any value a, b, c, and z such that ${}_2F_1(a,b,c,z)$ is well-defined, although it becomes arbitrarily close to zero as $\tilde{\pi}_2 \to 0$. □

A.20 Lemma S.1

Let Assumptions 1(ii)-(iii) and 2 hold. Then

$$ \mathbb{E}[||x_{it} - \boldsymbol{\mu}_j^0||^2] = \sum_{l=1}^{p} \sigma^2_{z_{it}^0,ll} + \sum_{l=1}^{p} (a_{jz_{it}^0,l})^2, $$

where $a_{jg,l}$ is the l^{th} element of the p-sized column-vector $\boldsymbol{a}_{jg} := \boldsymbol{\mu}_j^0 - \boldsymbol{\mu}_g^0$ for any pair $(j,g) \in \mathbb{G} \times \mathbb{G}$.

Proof. Given that $\mathbb{E}[||x_{it} - \boldsymbol{\mu}_j^0||^2]$ is a scalar, we have that

$$ \begin{aligned} \mathbb{E}[||x_{it} - \boldsymbol{\mu}_j^0||^2] &= \text{tr}(\mathbb{E}[||x_{it} - \boldsymbol{\mu}_j^0||^2]), \\ &= \text{tr}(\mathbb{E}[(x_{it} - \boldsymbol{\mu}_j^0)^\top (x_{it} - \boldsymbol{\mu}_j^0)]), \\ &= \text{tr}(\mathbb{E}[(x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0 - \boldsymbol{a}_{jz_{it}^0})(x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0 - \boldsymbol{a}_{jz_{it}^0})^\top]), \\ &= \text{tr}(\Sigma_{z_{it}^0}^0) - 2 \times \text{tr}(\mathbb{E}[(x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0)\boldsymbol{a}_{jz_{it}^0}^\top]) + \text{tr}(\boldsymbol{a}_{jz_{it}^0}\boldsymbol{a}_{jz_{it}^0}^\top), \\ &= \sum_{l=1}^{p} \sigma^2_{z_{it}^0,ll} + \sum_{l=1}^{p} (a_{jz_{it}^0,l})^2, \end{aligned} $$

given that $\mathbb{E}[x_{it}] = \boldsymbol{\mu}_{z_{it}^0}^0$ and that $\boldsymbol{a}_{jz_{it}^0}$ is non-random, and where $tr(\cdot)$ is the trace operator. □

A.21 Lemma S.2

Let Assumptions 1(ii)-(iii) and 2 hold. Then

$$ tr(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) = p - 2\sum_{l=1}^{p} v_{jz_{it}^0,ll} + \sum_{l=1}^{p}\sum_{m\le l} (v_{jz_{it}^0,lm})^2, $$

where $v_{jg,lm}$ corresponds to the entry located at the l^{th} row and at the m^{th} column of $V_{jg} = W_g^{-1}A_{jg}$, with W_g and A_{jg} both corresponding to $p \times p$ lower triangular matrices for any pair $(j,g) \in \mathbb{G} \times \mathbb{G}$.

Proof. Given that Σ_j^0 is a $p \times p$ symmetric, positive-definite matrix, there exists a lower triangular matrix W_j such that $\{\Sigma_j^0\}^{-1} = W_j W_j^\top$ for each $j \in \mathbb{G}$. If we define the lower triangular matrix

$A_{jz_{it}^0} := W_{z_{it}^0} - W_j$ for any $j \neq z_{it}^0$, we have that

$$\begin{aligned}
\text{tr}(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) &= \text{tr}(W_j W_j^\top \Sigma_{z_{it}^0}^0), \\
&= \text{tr}((W_{z_{it}^0} - A_{jz_{it}^0})(W_{z_{it}^0} - A_{jz_{it}^0})^\top \Sigma_{z_{it}^0}^0), \\
&= p + \text{tr}(A_{jz_{it}^0} A_{jz_{it}^0}^\top \Sigma_{z_{it}^0}^0 - A_{jz_{it}^0} W_{z_{it}^0}^\top \Sigma_{z_{it}^0}^0 - W_{z_{it}^0} A_{jz_{it}^0}^\top \Sigma_{z_{it}^0}^0), \\
&= p + \text{tr}(A_{jz_{it}^0} A_{jz_{it}^0}^\top \{W_{z_{it}^0}^\top\}^{-1} W_{z_{it}^0}^{-1} - A_{jz_{it}^0} W_{z_{it}^0}^\top \{W_{z_{it}^0}^\top\}^{-1} W_{z_{it}^0}^{-1} - W_{z_{it}^0} A_{jz_{it}^0}^\top \{W_{z_{it}^0}^\top\}^{-1} W_{z_{it}^0}^{-1}), \\
&= p + \text{tr}(A_{jz_{it}^0}^\top \{W_{z_{it}^0}^\top\}^{-1} W_{z_{it}^0}^{-1} A_{jz_{it}^0}) - \text{tr}(A_{jz_{it}^0} W_{z_{it}^0}^{-1}) - \text{tr}(A_{jz_{it}^0}^\top \{W_{z_{it}^0}^\top\}^{-1}), \\
&= p + \text{tr}(\{W_{z_{it}^0}^{-1} A_{jz_{it}^0}\}^\top W_{z_{it}^0}^{-1} A_{jz_{it}^0}) - 2\text{tr}(W_{z_{it}^0}^{-1} A_{jz_{it}^0}), \\
&= p + \text{tr}(V_{jz_{it}^0}^\top V_{jz_{it}^0}) - 2\text{tr}(V_{jz_{it}^0}), \\
&= p - 2\sum_{l=1}^{p} v_{jz_{it}^0,ll} + \sum_{l=1}^{p}\sum_{m\leq l}(v_{jz_{it}^0,lm})^2
\end{aligned}$$

given that each element of $V_{jz_{it}^0} = W_{z_{it}^0}^{-1} A_{jz_{it}^0}$ above the main diagonal is equal to zero. □

A.22 Lemma S.3

Let Assumptions 1(ii)-(iii) and 2 hold. Then

$$\mathbb{E}[(x_{it} - \boldsymbol{\mu}_j^0)\{\Sigma_j^0\}^{-1}(x_{it} - \boldsymbol{\mu}_j^0)^\top] = p - 2\sum_{l=1}^{p} v_{jz_{it}^0,ll} + \sum_{l=1}^{p}\sum_{m\leq l}(v_{jz_{it}^0,lm})^2 + \sum_{l=1}^{p}(b_{jz_{it}^0,l})^2$$

where $v_{jg,lm}$ is defined as in Appendix A.21 and where $b_{jg,l}$ corresponds to the l^{th} element of the p-sized row-vector $\mathbf{b}_{jg} = \mathbf{a}_{jg}^\top W_j$ for any pair $(j,g) \in \mathbb{G} \times \mathbb{G}$, with $\mathbf{a}_{jg}$ defined as in Appendix A.20 and where W_j is a $p \times p$ lower triangular matrix such that $\{\Sigma_j^0\}^{-1} = W_j W_j^\top$ for any $j \in \mathbb{G}$.

Proof. From the proof in Appendix A.21, we can write that

$$\begin{aligned}
\mathbb{E}[(x_{it} - \boldsymbol{\mu}_j^0)\{\Sigma_j^0\}^{-1}(x_{it} - \boldsymbol{\mu}_j^0)^\top] &= \text{tr}(\mathbb{E}[(x_{it} - \boldsymbol{\mu}_j^0)^\top \{\Sigma_j^0\}^{-1}(x_{it} - \boldsymbol{\mu}_j^0)]), \\
&= \text{tr}(\{\Sigma_j^0\}^{-1}\mathbb{E}[(x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0 - \boldsymbol{a}_{jz_{it}^0})(x_{it} - \boldsymbol{\mu}_{z_{it}^0}^0 - \boldsymbol{a}_{jz_{it}^0})^\top]), \\
&= \text{tr}(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) + \text{tr}(\{\Sigma_j^0\}^{-1}\mathbb{E}[2(\boldsymbol{\mu}_{z_{it}^0}^0 - x_{it})\boldsymbol{a}_{jz_{it}^0}^\top + \boldsymbol{a}_{jz_{it}^0}\boldsymbol{a}_{jz_{it}^0}^\top]), \\
&= \text{tr}(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) + \text{tr}(W_j W_j^\top \boldsymbol{a}_{jz_{it}^0}\boldsymbol{a}_{jz_{it}^0}^\top), \\
&= \text{tr}(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) + \sum_{l=1}^{p}(b_{jz_{it}^0,l})^2,
\end{aligned}$$

From the same proof, we also know that

$$\text{tr}(\{\Sigma_j^0\}^{-1}\Sigma_{z_{it}^0}^0) = p - 2\sum_{l=1}^{p} v_{jz_{it}^0,ll} + \sum_{l=1}^{p}\sum_{m\leq l}(v_{jz_{it}^0,lm})^2,$$

where $v_{jz_{it}^0,lm}$ is defined as in Appendix A.21. Combining the two results completes the proof. □

A.23 Lemma S.4

Let $f(x)$ and $g(x)$ be two distinct, univariate and positive functions of the p-variate random variable X, where $f : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ and $g : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ such that the two functions share the same support and the same image. Let also $f(x) \neq g(x)$ for at least one $x \in \mathcal{X}$ with $E[f(x)] < \infty$ and $E[g(x)] < \infty$ for any $|x|_1 < \infty$, where $|\cdot|_1$ denotes the L^1 norm. Then

$$\mathbb{P}[f(x) \geq g(x)] \leq \frac{\mathbb{E}[f(x)] - Cov[g(x), \mathbb{1}_{[f(x)\geq g(x)]}]}{\mathbb{E}[g(x)]},$$

provided that $\mathbb{E}[g(x)] \neq 0$.

Proof. This lemma is a generalization of Markov's inequality to univariate functions of X on both sides of the inequality that is contained within the probability. The standard Markov's inequality for functions of X states that

$$\mathbb{P}[f(x) \geq a] \leq \frac{\mathbb{E}[f(x)]}{a},$$

for any constant $a > 0$ and any function $f(x)$, which is a special case of the generalized inequality when $g(x) = a$. The argument for the generalized Markov's inequality goes as follows. Let the indicator function $\mathbb{1}_{[f(x)\geq g(x)]}$ be equal to one if and only if $f(x) \geq g(x)$ and zero otherwise. Therefore, we have that

$$f(x) \geq g(x)\mathbb{1}_{[f(x)\geq g(x)]},$$

which is always satisfied. Hence, we can take expectation on both sides of the inequality and it will never be reversed since

$$\begin{aligned}
\mathbb{E}[g(x)\mathbb{1}_{[f(x)\geq g(x)]}] &= \int_{\mathcal{X}} g(x)\mathbb{1}_{[f(x)\geq g(x)]} p_X(x)dx, \\
&= \int_{f(x)\geq g(x)} g(x)p_X(x)dx, \\
&\leq \int_{f(x)\geq g(x)} f(x)p_X(x)dx \leq \int_{\mathcal{X}} f(x)p_X(x)dx = \mathbb{E}[f(x)],
\end{aligned}$$

where $p_X(x)$ corresponds to the true generating density of X. The last inequality comes from the fact that $f(x) \geq 0$ for any $x \in \mathcal{X}$. Note also that equality occurs when $g(x) = f(x)$ for all $x \in \mathcal{X}\backslash_{x^*}$, where x^* is a set of measure zero in $\mathcal{X}$. Therefore we can write that

$$\begin{aligned}
\mathbb{E}[f(x)] &\geq \mathbb{E}[g(x)\mathbb{1}_{[f(x)\geq g(x)]}], \\
&\geq \text{Cov}[g(x), \mathbb{1}_{[f(x)\geq g(x)]}] + \mathbb{E}[g(x)]\mathbb{E}[\mathbb{1}_{[f(x)\geq g(x)]}], \\
&\geq \text{Cov}[g(x), \mathbb{1}_{[f(x)\geq g(x)]}] + \mathbb{E}[g(x)]\mathbb{P}[f(x) \geq g(x)].
\end{aligned}$$

Subtracting $\text{Cov}[g(x), \mathbb{1}_{[f(x)\geq g(x)]}]$ and then dividing by $\mathbb{E}[g(x)]$ on each side of the inequality leads to the generalized Markov's inequality. □

B Details of the estimation procedure for the second simulation exercise

In practice, both the EM and the CEM algorithms correspond to an iterative weighted generalized least squares (IWGLS) procedure where the weights depend on the chosen algorithm. It is a form of generalized least squares (GLS) procedure given that the variance of the unit-random effects $\sigma^2_{\alpha,j}$ is taken into account through the use of the following equation

$$\hat{\tilde{\beta}}_g^{(k+1)} = (\sum_{i=1}^{N} \{\tilde{\mathbf{X}}_{ig}^{(k)}\}^\top \{\hat{\Omega}_g^{(k)}\}^{-1} \tilde{\mathbf{X}}_{ig}^{(k)})^{-1} (\sum_{i=1}^{N} \{\tilde{\mathbf{X}}_{ig}^{(k)}\}^\top \{\hat{\Omega}_g^{(k)}\}^{-1} \tilde{\mathbf{y}}_{ig}^{(k)}), \tag{17}$$

where $\tilde{\mathbf{X}}_{ig}^{(k)} = (w_{i1g}^{(k)} X_{i1}^\top, ..., w_{iTg}^{(k)} X_{iT}^\top)^\top$, $\tilde{\mathbf{y}}_{ig}^{(k)} = (w_{i1g}^{(k)} y_{i1}, ..., w_{iTg}^{(k)} y_{iT})^\top$, with

$$w_{itg}^{(k)} = \begin{cases} z_{itg}^D(\hat{\tilde{\beta}}^{(k)}, \hat{\Omega}^{(k-1)}, \hat{\xi}^{(k-1)}) \text{ if the CEM algorithm is used,} \\ \tau_{itg}(\hat{\tilde{\beta}}^{(k)}, \hat{\Omega}^{(k-1)}, \hat{\pi}^{(k-1)}) \text{ if the EM algorithm is used,} \end{cases} \tag{18}$$

where $\hat{\Omega}^{(k)} = (\hat{\Omega}_1^{(k)}, ..., \hat{\Omega}_G^{(k)})$ refers to the set of variance-covariance matrices of the outcome for each group, and where $\hat{\pi}^{(k)} = (\hat{\pi}_1^{(k)}, ..., \hat{\pi}_G^{(k)})$ is estimated using eq.(7) and eq.(8). The elements on the main diagonal of $\hat{\Omega}_g^{(k)}$ are estimated using

$$\hat{\omega}_{\alpha+\epsilon,g}^{2(k)} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (w_{itg}^{(k)} \hat{\epsilon}_{itg}^{(k)})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} w_{itg}^{(k)} - 2p - T}, \tag{19}$$

which corrects for the finite-sample bias, while the off-diagonal elements are estimated using

$$\hat{\omega}_{\alpha,g}^{2(k)} = \frac{1}{N_g} \sum_{i=1}^{N} \left(\frac{\sum_{t=1}^{T} w_{itg}^{(k)} \hat{\epsilon}_{itg}^{(k)}}{\sum_{t=1}^{T} w_{itg}^{(k)}} - \frac{1}{N_g} \sum_{j=1}^{N} \left(\frac{\sum_{t=1}^{T} w_{jtg}^{(k)} \hat{\epsilon}_{jtg}^{(k)}}{\sum_{t=1}^{T} w_{jtg}^{(k)}} \right) \right)^2, \tag{20}$$

where $\hat{\epsilon}_{itg}^{(k)} = y_{it} - X_{it} \hat{\tilde{\beta}}_g^{(k)}$, $N_g = \sum_{i=1}^{N} \mathbb{1}[\sum_{t=1}^{T} w_{itg}^{(k)} \neq 0]$, and where $w_{itg}^{(k)}$ is defined as above. When using the CEM algorithm, one has to make sure that $\sum_{t=1}^{T} w_{itg}^{(k)} \neq 0$ for any $i \in \{1, ..., N\}$, which never occurs with the EM algorithm. Finally, $\hat{\xi}^{(k)} = (\hat{\boldsymbol{\mu}}^{(k)}, \hat{\Sigma}^{(k)})$ is computed using the following equations

$$\hat{\boldsymbol{\mu}}_g^{(k)} = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{w_{itg}^{(k)} x_{it}}{\sum_{j=1}^{N} \sum_{l=1}^{T} w_{jlg}^{(k)}}, \qquad \hat{\Sigma}_g^{(k)} = \frac{\sum_{i=1}^{N} w_{itg}^{(k-1)} (x_{it} - \hat{\boldsymbol{\mu}}_g^{(k)})(x_{it} - \hat{\boldsymbol{\mu}}_g^{(k)})^\top}{(\sum_{j=1}^{N} \sum_{t=1}^{T} w_{jtg}^{(k)}) - p}.$$

Note that the estimators presented in eq.(17) and eq.(19) do not correspond to the estimators typically used in the literature on the EM algorithm and finite mixtures (see, for instance, Celeux

(2019)). Those more "typical" estimators are represented by the following equations

$$\hat{\tilde{\beta}}_g^{(k+1)} = (\sum_{i=1}^{N} \{\tilde{\mathbf{X}}_{ig}^{(k)}\}^\top \{\hat{\Omega}_g^{(k)}\}^{-1} \mathbf{X}_i)^{-1} (\sum_{i=1}^{N} \{\tilde{\mathbf{X}}_{ig}^{(k)}\}^\top \{\hat{\Omega}_g^{(k)}\}^{-1} \mathbf{y}_i),$$

and

$$\hat{\omega}_{\alpha+\epsilon,g}^{2(k)} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} w_{itg}^{(k)} (\hat{\epsilon}_{itg}^{(k)})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} w_{itg}^{(k)} - 2p - T},$$

where $\mathbf{X}_i = (X_{i1}^\top, ..., X_{iT}^\top)^\top$ and where $\mathbf{y}_i = (y_{i1}, ..., y_{iT})^\top$. However, those two estimators yielded very poor performance in the context of the iterative GLS approach, which explains why they were discarded for the second simulation exercise.

C Additional Simulation Results

C.1 First Simulation Exercise

$\boldsymbol{\mu}^0$	$E(\theta^0, \pi^0)$ (%)	N	$l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$	RMSE, EM	$l^{MC}(\hat{\theta}, \hat{\pi}) - l^C(\theta^0, \pi^0)$	RMSE, CEM
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(-0.25, 0.25)	**40.1**	2,000	2.614	2.58488	982.1	0.60388
		10,000	0.816	2.57130	4840.7	0.58816
		20,000	0.031	2.47074	9621.9	0.57935
		50,000	1.425	2.41382	24246.1	0.58406
		200,000	-3.344	0.76522	96596.7	0.58118
		1,000,000	-12.122	0.09716	483152.0	0.57993
(-0.5, 0.5)	**30.9**	2,000	3.189	1.05061	875.9	0.48139
		10,000	-7.132	0.91988	4224.2	0.43543
		20,000	3.388	0.72230	8349.5	0.43733
		50,000	1.964	0.09274	21240.7	0.43996
		200,000	2.207	0.11510	84543.9	0.43986
		1,000,000	-0.187	0.09766	421934.5	0.44062
(-1, 1)	**15.9**	2,000	1.660	0.15334	566.4	0.36339
		10,000	2.248	0.00785	2561.1	0.37335
		20,000	5.436	0.06951	4933.1	0.38264
		50,000	1.485	0.00616	12777.9	0.38766
		200,000	1.769	0.02440	51345.8	0.38348
		1,000,000	0.884	0.01423	254605.4	0.38570
(-2, 2)	**2.3**	2,000	2.620	0.02663	108.5	0.18752
		10,000	3.828	0.00676	447.8	0.16991
		20,000	4.196	0.01332	739.4	0.14588
		50,000	2.830	0.00904	1941.6	0.13660
		200,000	1.624	0.00247	7993.7	0.14377
		1,000,000	0.685	0.00085	39168.4	0.13756

Table 5: Root mean square errors (RMSEs) of the estimated mean values and differences in log likelihood values when $G = 2$ and $\pi^0 = (0.25, 0.75)$; the true variances are all equal to one.

Algorithm (1)	$\boldsymbol{\mu^0}$ (2)	$\hat{\pi}_1$ (3)	$\hat{\pi}_2$ (4)	$\hat{\mu}_1$ (5)	$\hat{\mu}_2$ (6)	$\hat{\sigma}_1$ (7)	$\hat{\sigma}_2$ (8)
EM	**(-0.25, 0.25)**	0.00000	1.00000	-0.19285	0.12505	3.16313	1.02349
	(-0.5, 0.5)	0.19334	0.80666	-0.63267	0.46162	0.97664	1.00657
	(-1, 1)	0.25499	0.74501	-0.98112	1.00699	1.00801	0.99772
	(-2, 2)	0.24995	0.75005	-2.00118	2.00020	1.00143	1.00024
CEM	**(-0.25, 0.25)**	0.49900	0.50100	-0.69324	0.94005	0.61924	0.61468
	(-0.5, 0.5)	0.49328	0.50672	-0.63225	1.10893	0.67170	0.64087
	(-1, 1)	0.44343	0.55657	-0.69554	1.45259	0.86439	0.71105
	(-2, 2)	0.27751	0.72249	-1.82512	2.08522	1.09129	0.91621

Table 6: Estimated values for each scenario when $G = 2$, $\pi^0 = (0.25, 0.75)$, and $N = 1,000,000$.

$\boldsymbol{\mu}^0$	$E(\theta^0, \pi^0)$ (%)	N	$l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$	RMSE, EM	$l^{MC}(\hat{\theta}, \hat{\pi}) - l^C(\theta^0, \pi^0)$	RMSE, CEM
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(-0.25, 0, 0.25)	**60.0**	3,000	7.062	0.82353	2653.3	0.75456
		15,000	7.985	0.29892	13070.8	0.70468
		30,000	2.764	0.29072	26068.5	0.70865
		75,000	1.607	0.51585	65618.5	0.70856
		300,000	0.990	0.52223	261714.4	0.70774
		1,500,000	1.686	0.51465	1309189.5	0.70765
(-0.5, 0, 0.5)	**53.5**	3,000	4.823	1.05944	2490.3	0.58624
		15,000	6.376	0.26268	12273.4	0.55455
		30,000	4.895	0.37070	24317.5	0.55477
		75,000	3.152	0.49155	61451.0	0.55747
		300,000	1.762	0.42158	245027.0	0.55629
		1,500,000	1.057	0.41918	1226337.5	0.55690
(-1, 0, 1)	**41.1**	3,000	8.089	0.87639	2014.0	0.37785
		15,000	5.904	0.27492	9905.0	0.33807
		30,000	2.704	0.30480	19608.5	0.34539
		75,000	4.140	0.40140	49342.0	0.33863
		300,000	0.401	0.39097	196468.9	0.33955
		1,500,000	-10.171	0.37673	982396.2	0.33981
(-2, 0, 2)	**21.2**	3,000	1.728	0.11972	1060.2	0.11688
		15,000	6.249	0.20071	5079.3	0.11967
		30,000	6.388	0.05679	10043.3	0.11239
		75,000	3.624	0.07909	25627.2	0.10628
		300,000	3.213	0.04284	100468.7	0.10024
		1,500,000	1.063	0.00650	503814.6	0.09927
(-4, 0, 4)	**3.0**	3,000	3.193	0.02519	172.5	0.02161
		15,000	8.119	0.02359	672.3	0.01736
		30,000	4.563	0.01481	1414.9	0.02137
		75,000	4.624	0.00558	3715.3	0.00696
		300,000	2.414	0.00083	14440.6	0.00729
		1,500,000	2.229	0.00048	71910.0	0.00537

Table 7: Root mean square errors (RMSEs) of the estimated mean values and differences in log-likelihood values when $G = 3$ and $\pi^0 = (0.33, 0.33, 0.33)$; the true variances are all equal to one.

Algorithm (1)	$\boldsymbol{\mu^0}$ (2)	$\hat{\pi}_1$ (3)	$\hat{\pi}_2$ (4)	$\hat{\pi}_3$ (5)	$\hat{\mu}_1$ (6)	$\hat{\mu}_2$ (7)	$\hat{\mu}_3$ (8)	$\hat{\sigma}_1$ (9)	$\hat{\sigma}_2$ (10)	$\hat{\sigma}_3$ (11)
	(-0.25, 0, 0.25)	0.01078	0.42347	0.56575	-1.10053	-0.26507	0.21926	0.89502	0.97883	0.99398
	(-0.5, 0, 0.5)	0.00730	0.30079	0.69191	-0.92592	-0.52649	0.23927	0.91388	0.99304	1.03159
EM	**(-1, 0, 1)**	0.49942	0.15715	0.34342	-0.75353	0.57987	0.83027	1.04830	1.05685	1.03924
	(-2, 0, 2)	0.33473	0.32860	0.33667	-1.99669	-0.00699	1.99180	1.00167	0.99196	1.00376
	(-4, 0, 4)	0.33340	0.33319	0.33342	-3.99988	0.00026	3.99921	1.00077	1.00009	1.00243
	(-0.25, 0, 0.25)	0.32979	0.33561	0.33461	-1.12162	-0.00643	1.11171	0.53858	0.25226	0.54084
	(-0.5, 0, 0.5)	0.33200	0.33495	0.33304	-1.18319	-0.00163	1.18093	0.56821	0.26772	0.56922
CEM	**(-1, 0, 1)**	0.33573	0.32810	0.33617	-1.41648	-0.00147	1.41588	0.65189	0.32674	0.65265
	(-2, 0, 2)	0.34597	0.30693	0.34710	-2.12326	-0.00417	2.11981	0.81654	0.52904	0.81851
	(-4, 0, 4)	0.33501	0.32989	0.33510	-4.00703	-0.00028	4.00609	0.97402	0.91287	0.97574

Table 8: Estimated values for each scenario when $G = 3$, $\pi^0 = (0.33, 0.33, 0.33)$, and $N = 1,500,000$.

$\boldsymbol{\mu}^0$	$E(\theta^0, \pi^0)$ (%)	N	$l(\hat{\theta}, \hat{\pi}) - l(\theta^0, \pi^0)$	RMSE, EM	$l^{MC}(\hat{\theta}, \hat{\pi}) - l^C(\theta^0, \pi^0)$	RMSE, CEM
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(-0.25, 0, 0.25)	**60.0**	3,000	3.657	2.52930	2684.8	0.76083
		15,000	1.828	2.51983	13065.7	0.71051
		30,000	0.880	2.01044	26111.1	0.72044
		75,000	-0.336	1.50215	65795.0	0.70565
		300,000	-1.053	1.42663	262800.7	0.71203
		1,500,000	0.389	1.08786	1314054.2	0.70950
(-0.5, 0, 0.5)	**53.5**	3,000	2.924	1.02572	2563.2	0.59841
		15,000	2.556	0.10369	12437.3	0.57450
		30,000	1.006	0.22664	24647.7	0.57466
		75,000	0.504	0.24050	62337.4	0.57338
		300,000	1.496	0.15410	248502.9	0.56899
		1,500,000	1.828	0.39178	1243552.9	0.56916
(-1, 0, 1)	**41.1**	3,000	3.642	0.85642	2108.2	0.54906
		15,000	6.196	0.76961	10329.7	0.40603
		30,000	3.261	0.71102	20296.1	0.40375
		75,000	1.718	0.57232	51538.1	0.43594
		300,000	3.031	0.66091	205069.9	0.42943
		1,500,000	0.724	0.57985	1025338.2	0.43783
(-2, 0, 2)	**21.2**	3,000	2.995	0.29103	1177.3	0.55276
		15,000	6.864	0.20622	5601.3	0.59551
		30,000	8.495	0.16203	11069.8	0.59692
		75,000	3.917	0.09940	28538.7	0.51499
		300,000	2.727	0.03457	113209.0	0.56665
		1,500,000	2.601	0.01603	563246.0	0.57190
(-4, 0, 4)	**3.0**	3,000	3.343	0.01737	183.0	0.07314
		15,000	7.144	0.01578	717.6	0.06637
		30,000	6.789	0.01481	1397.2	0.06635
		75,000	2.383	0.00902	3701.2	0.05595
		300,000	2.203	0.00211	14878.7	0.06305
		1,500,000	2.615	0.00099	73359.0	0.06297

Table 9: Root mean square errors (RMSEs) of the estimated mean values and differences in log-likelihood values when $G = 3$ and $\pi^0 = (0.167, 0.33, 0.50)$; the true variances are all equal to one.

Algorithm (1)	$\boldsymbol{\mu^0}$ (2)	$\hat{\pi}_1$ (3)	$\hat{\pi}_2$ (4)	$\hat{\pi}_3$ (5)	$\hat{\mu}_1$ (6)	$\hat{\mu}_2$ (7)	$\hat{\mu}_3$ (8)	$\hat{\sigma}_1$ (9)	$\hat{\sigma}_2$ (10)	$\hat{\sigma}_3$ (11)
	(-0.25, 0, 0.25)	0.00025	0.99974	0.00000	-2.09923	0.08381	0.60173	0.45722	1.01764	3.42487
	(-0.5, 0, 0.5)	0.26489	0.73511	0.00001	-0.40353	0.37203	1.05926	1.00303	1.01502	2.70881
EM	**(-1, 0, 1)**	0.00484	0.46835	0.52681	-1.94502	-0.33621	0.94938	0.68833	1.11344	1.01268
	(-2, 0, 2)	0.17078	0.32996	0.49926	-1.97632	0.01442	2.00168	1.00944	0.99699	1.00143
	(-4, 0, 4)	0.16660	0.33350	0.49989	-4.00138	-0.00040	4.00096	1.00137	1.00095	1.00137
	(-0.25, 0, 0.25)	0.33150	0.33576	0.33274	-1.03147	0.08228	1.19483	0.53883	0.25142	0.53733
	(-0.5, 0, 0.5)	0.33104	0.33514	0.33382	-1.00657	0.16769	1.32890	0.56975	0.26364	0.55726
CEM	**(-1, 0, 1)**	0.32736	0.33341	0.33923	-1.06787	0.34915	1.66976	0.68134	0.31255	0.60695
	(-2, 0, 2)	0.31474	0.31603	0.36923	-1.49399	0.70851	2.47243	0.97691	0.47570	0.69843
	(-4, 0, 4)	0.17548	0.32881	0.49571	-3.92188	0.06899	4.03216	1.03303	0.90627	0.95769

Table 10: Estimated values for each scenario when $G = 3$, $\pi^0 = (0.167, 0.33, 0.50)$, and $N = 1,500,000$.

C.2 Second Simulation Exercise

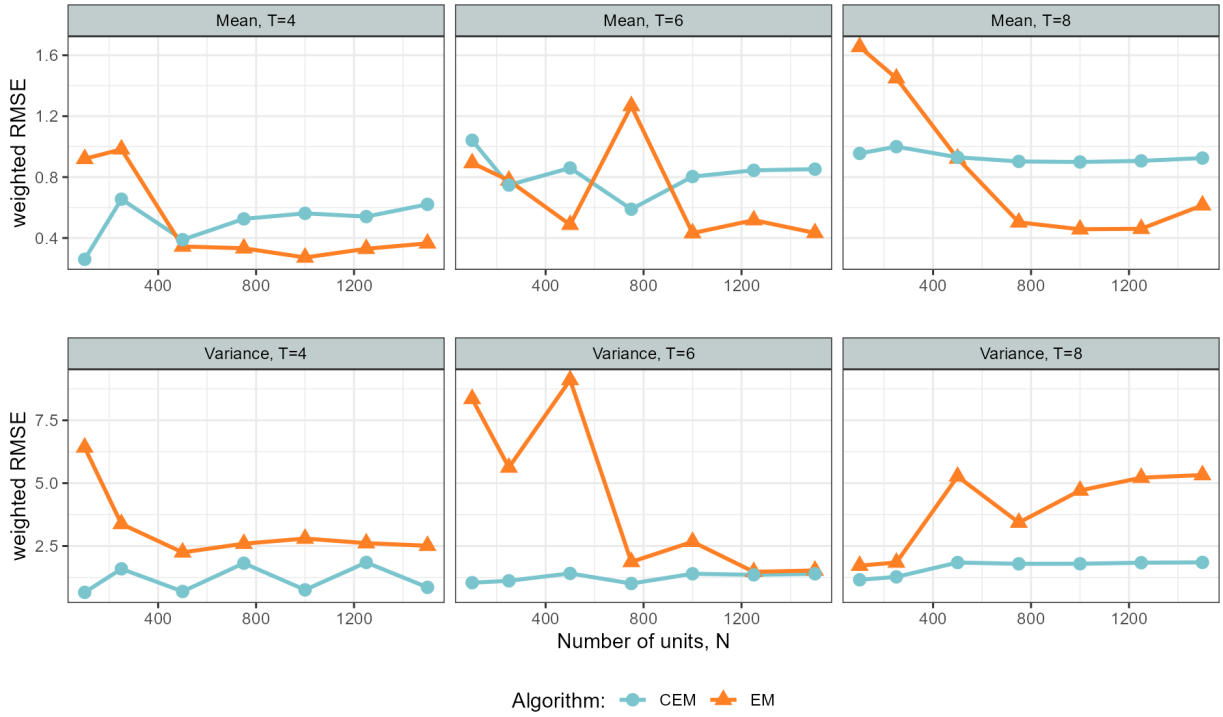


Figure 7: Evolution of the estimation error when $G = 3$ and when $\hat{E}_{NTp}(\theta^0, \xi^0) = [4.52\%, 6.87\%]$ with $p = 3$. The estimates selected to compute the weighted RMSEs are the ones that maximize the log likelihood function associated with each algorithm. The y-axis stands as the weighted RMSE for each total number of periods T, and each type of parameter (mean and variance). The "true" mixing weights all lie between 0.234 and 0.504.

N	$\hat{E}(\cdot)$	$T=4$		$T=6$		$T=8$	
		EM (%)	CEM (%)	EM (%)	CEM (%)	EM (%)	CEM (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
100	$\hat{E}(\theta^0, \xi^0)$	4.75	4.75	6.50	6.50	5.63	5.63
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	30.50	5.50	37.83	36.17	41.50	24.88
250	$\hat{E}(\theta^0, \xi^0)$	5.00	5.00	6.87	6.87	6.10	6.10
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	34.60	74.30	35.27	31.73	38.95	31.15
500	$\hat{E}(\theta^0, \xi^0)$	4.55	4.55	6.13	6.13	6.38	6.38
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	22.30	17.75	35.70	64.23	37.55	66.03
750	$\hat{E}(\theta^0, \xi^0)$	4.57	4.57	6.31	6.31	6.15	6.15
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	20.43	74.57	38.53	28.96	38.83	67.30
1000	$\hat{E}(\theta^0, \xi^0)$	4.52	4.52	5.82	5.82	6.06	6.06
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	21.65	29.18	35.42	64.57	38.73	66.43
1250	$\hat{E}(\theta^0, \xi^0)$	4.92	4.92	5.91	5.91	6.14	6.14
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	20.56	74.10	24.96	64.79	38.26	66.86
1500	$\hat{E}(\theta^0, \xi^0)$	6.14	6.14	6.47	6.47	6.39	6.39
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	20.05	57.27	25.43	64.74	38.98	67.28

Table 11: Misclassification rates evaluated at the true parameter values and evaluated at the parameter values that maximize the log likelihood function for each algorithm when $\hat{E}_{NTp}(\theta^0, \xi^0) = [4.52\%, 6.87\%]$. The NTp subscripts are dropped for brevity. The misclassification rates of the EM algorithm are computed with the maximum posterior probability. CEM = CEM.

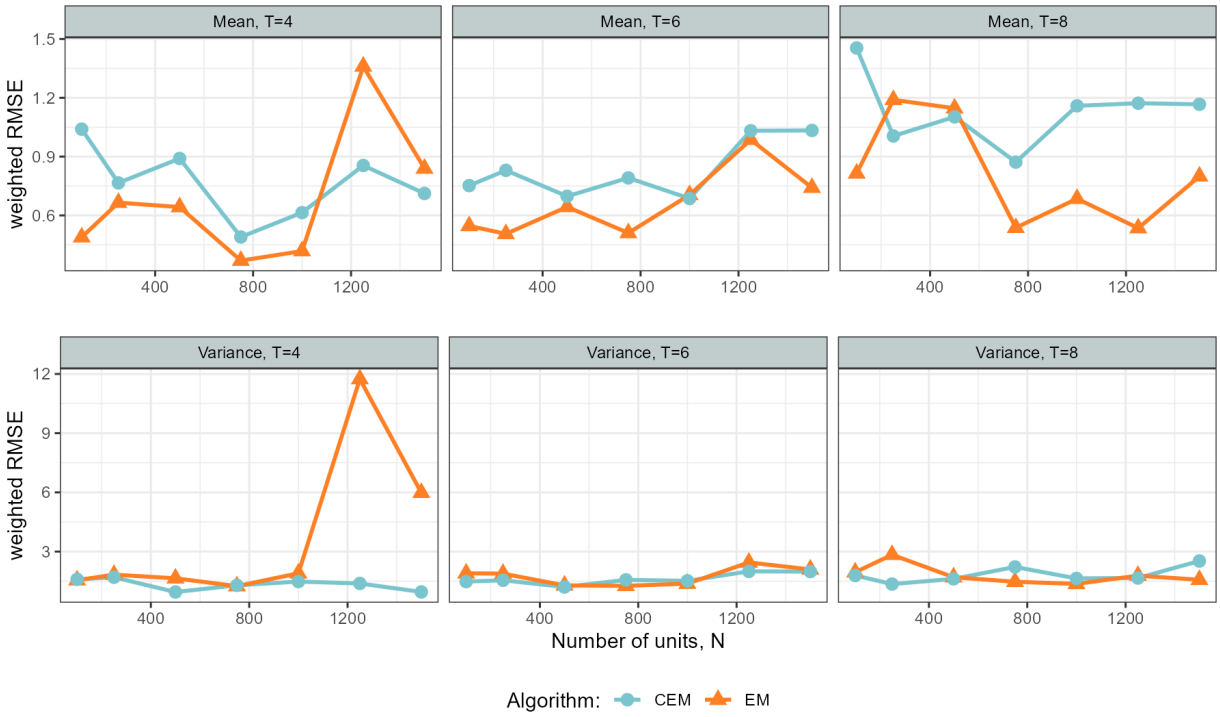


Figure 8: Evolution of the estimation error when $G = 3$ and when $\hat{E}_{NTp}(\theta^0, \xi^0) = [13.00\%, 16.20\%]$ with $p = 2$. The estimates selected to compute the weighted RMSEs are the ones that maximize the log likelihood function associated with each algorithm. The y-axis stands as the weighted RMSE for each total number of periods T, and each type of parameter (mean and variance). The "true" mixing weights all lie between 0.234 and 0.504.

N	$\hat{E}(\cdot)$	$T=4$		$T=6$		$T=8$	
		EM (%)	CEM (%)	EM (%)	CEM (%)	EM (%)	CEM (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
100	$\hat{E}(\theta^0, \xi^0)$	13.00	13.00	14.33	14.33	14.63	14.63
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	46.75	39.75	69.00	38.50	68.38	43.00
250	$\hat{E}(\theta^0, \xi^0)$	16.20	16.20	15.53	15.53	15.80	15.80
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	44.80	73.80	61.20	40.67	62.25	39.35
500	$\hat{E}(\theta^0, \xi^0)$	13.95	13.95	13.67	13.67	13.73	13.73
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	45.20	75.05	48.47	39.27	53.03	46.95
750	$\hat{E}(\theta^0, \xi^0)$	14.37	14.37	14.04	14.04	14.02	14.02
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	43.63	34.03	47.87	41.62	50.22	72.75
1000	$\hat{E}(\theta^0, \xi^0)$	14.63	14.63	14.90	14.90	14.40	14.40
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	44.15	78.33	48.15	40.62	51.73	49.28
1250	$\hat{E}(\theta^0, \xi^0)$	13.50	13.50	13.65	13.65	13.45	13.45
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	46.20	51.92	50.11	84.95	54.88	52.33
1500	$\hat{E}(\theta^0, \xi^0)$	14.12	14.12	13.81	13.81	13.58	13.58
	$\hat{E}(\hat{\theta}^{(k)}, \hat{\xi}^{(k)})$	46.95	31.05	51.12	84.68	52.31	84.82

Table 12: Misclassification rates evaluated at the true parameter values and evaluated at the parameter values that maximize the log likelihood function for each algorithm when $\hat{E}_{NTp}(\theta^0, \xi^0) = [13.00\%, 16.20\%]$. The NTp subscripts are dropped for brevity. The misclassification rates of the EM algorithm are computed with the maximum posterior probability. CEM = CEM.

D Covariates included in the empirical analysis

Covariate	Binary Part	Continuous Part
Time-varying ERA	X	X
Time-varying COCI		X
Time-averaged ERA	X	X
Time-averaged Charlson	X	X
Time-averaged COCI	X	X
Gender	X	X
Time-fixed effects	X	X
Unit-random effects	X	X
Intercept	X	X

Table 13: Covariates included in all component densities for each part of the model. COCI = Continuity of care indicator, ERA = Elder's risk assessment.

E Definition of the ERA and the Charlson Indices

Conditions	Weight
Congestive heart failure	1
Myocardial infarction	1
Peripheral vascular disease	1
Chronic pulmonary disease	1
Cerebrovascular disease	1
Dementia	1
Connective tissue disease	1
Ulcer disease	1
Mild liver disease	1
Diabetes	1
Depression	1
Use of warfarin	1
Hypertension	1
Hemiplegia	2
Moderate or severe renal disease	2
Diabetes with end organ damage	2
Any tumor	2
Leukemia	2
Lymphoma	2
Skin ulcers/cellulitis	2
Moderate or severe liver disease	3
Metastatic cancer	6
AIDS	6

(a) Charlson index

Parameters	Score
Social support index	
1st and 2nd quintiles	-1
3rd quintile and null value	0
4th and 5th quintiles	1
Age	
65-69	0
70-79	1
80-89	3
≥90	7
Number of days in hospital during the two previous years	
1 to 5 days	5
≥ 6 days	11
History of comorbidities	
Diabetes	2
CAD/MI/CHF	3
Stroke	2
COPD	5
Cancer	1
Dementia	3

(b) ERA index

Figure 9: The scoring systems of the modified ERA index and the modified Charlson index. CAD = Coronary artery disease; MI = Myocardial infarction; CHF = Chronic heart failure; COPD = Chronic obstructive pulmonary disorder; AIDS = Acquired immune deficiency syndrome. Underlined items are removed from the index to avoid collinearity.

F Additional Empirical Results

F.1 Estimates Obtained from the Best Model, CEM Algorithm

Coefficients	Group/Component				
	1	2	3	4	5
			Binary Part		
ERA	-5.413***	19.666***	-14.607***	4.341*	174.416**
	(0.904)	(5.046)	(1.755)	(1.785)	(55.578)
Time-averaged Charlson	-46.186***	31.002***	52.267***	35.537***	-758.674***
	(2.099)	(4.890)	(2.355)	(4.457)	(65.068)
Time-averaged COCI	-2.129***	329.400***	-6.345***	-2.077*	1789.412***
	(0.279)	(29.245)	(0.600)	(0.867)	(122.124)
Time-averaged ERA	1.179	47.679***	8.680***	0.449	659.342***
	(0.923)	(5.181)	(1.553)	(2.298)	(110.359)
Male	1.157	277.050***	-9.908***	5.760	381.782***
	(1.355)	(27.497)	(1.693)	(3.222)	(22.582)
Number of observations	2142	1392	2075	1774	1863
			Continuous Part		
ERA	-0.340***	-0.003	-0.669***	0.085*	-0.119***
	(0.050)	(0.028)	(0.066)	(0.037)	(0.034)
COCI	0.608***	0.391	0.349***	-0.103***	-0.085***
	(0.043)	(2.4E06)	(0.039)	(0.018)	(0.013)
Time-averaged Charlson	0.102	0.005	0.014	0.043*	-0.116***
	(0.070)	(0.021)	(0.081)	(0.021)	(0.028)
Time-averaged COCI	-0.171***	0.013	-0.128***	-0.032	-0.005
	(0.017)	(0.012)	(0.027)	(0.020)	(0.011)
Time-averaged ERA	0.549***	0.043	0.809***	-0.062	0.170***
	(0.052)	(0.031)	(0.063)	(0.043)	(0.035)
Male	0.251***	-0.012	0.078	-0.051	-0.015
	(0.066)	(0.043)	(0.079)	(0.061)	(0.038)
Number of observations	1801	1330	1097	1720	1828

Table 14: Additional estimates associated to the optimal set of estimates obtained from the CEM algorithm. Cluster-robust standard errors are shown in parentheses. The number of observations refers to the sum of the estimated group memberships within each group. * = p-value$<$ 0.05, ** = p-value$<$ 0.01, *** = p-value$<$ 0.001.

F.2 Estimates Obtained from the Best Model, EM Algorithm

Coefficients	Group/Component 1	2	3	4
Binary Part				
ERA	-110.673***	-922.474	-1.195	54144.590***
	(29.091)	(476.582)	(0.705)	(1830.669)
Time-averaged Charlson	-420.397***	205.818	-42.752***	487254.501***
	(14.880)	(360.249)	(2.808)	(20966.975)
Time-averaged COCI	-2329.817***	-1363.916***	-1.363***	22241.703***
	(78.736)	(118.584)	(0.291)	(751.892)
Time-averaged ERA	705.402***	3824.340***	2.434**	-63320.044***
	(34.534)	(582.868)	(0.815)	(2140.879)
Male	-232.246***	2166.655***	-1.726	-44511.269***
	(28.839)	(644.333)	(1.056)	(1504.569)
Number of observations	2452.8	2784.1	1956.5	2052.5
Continuous Part				
ERA	-0.124***	-0.011	-0.469***	-0.272***
	(0.030)	(0.031)	(0.076)	(0.043)
COCI	-0.292***	-0.006	0.153***	0.098***
	(0.027)	(0.006)	(0.017)	(0.018)
Time-averaged Charlson	-0.132***	0.060**	-15.542***	-11.622
	(0.025)	(0.021)	(1.573)	(51.432)
Time-averaged COCI	-0.073***	-0.026*	-0.021	0.089***
	(0.017)	(0.012)	(0.025)	(0.023)
Time-averaged ERA	0.230***	0.039	0.541***	0.372***
	(0.033)	(0.034)	(0.085)	(0.055)
Male	0.150**	0.000	0.014	0.186*
	(0.049)	(0.041)	(0.090)	(0.092)
Number of observations	2402.0	2782.9	540.6	2050.6

Table 15: Additional estimates associated to the optimal set of estimates obtained from the EM algorithm. Cluster-robust standard errors are shown in parentheses. The number of observations refers to the sum of the estimated posterior probabilities within each group. * = p-value$<$ 0.05, ** = p-value$<$ 0.01, *** = p-value$<$ 0.001.

F.3 Estimated Time-Fixed Effects, EM Algorithm

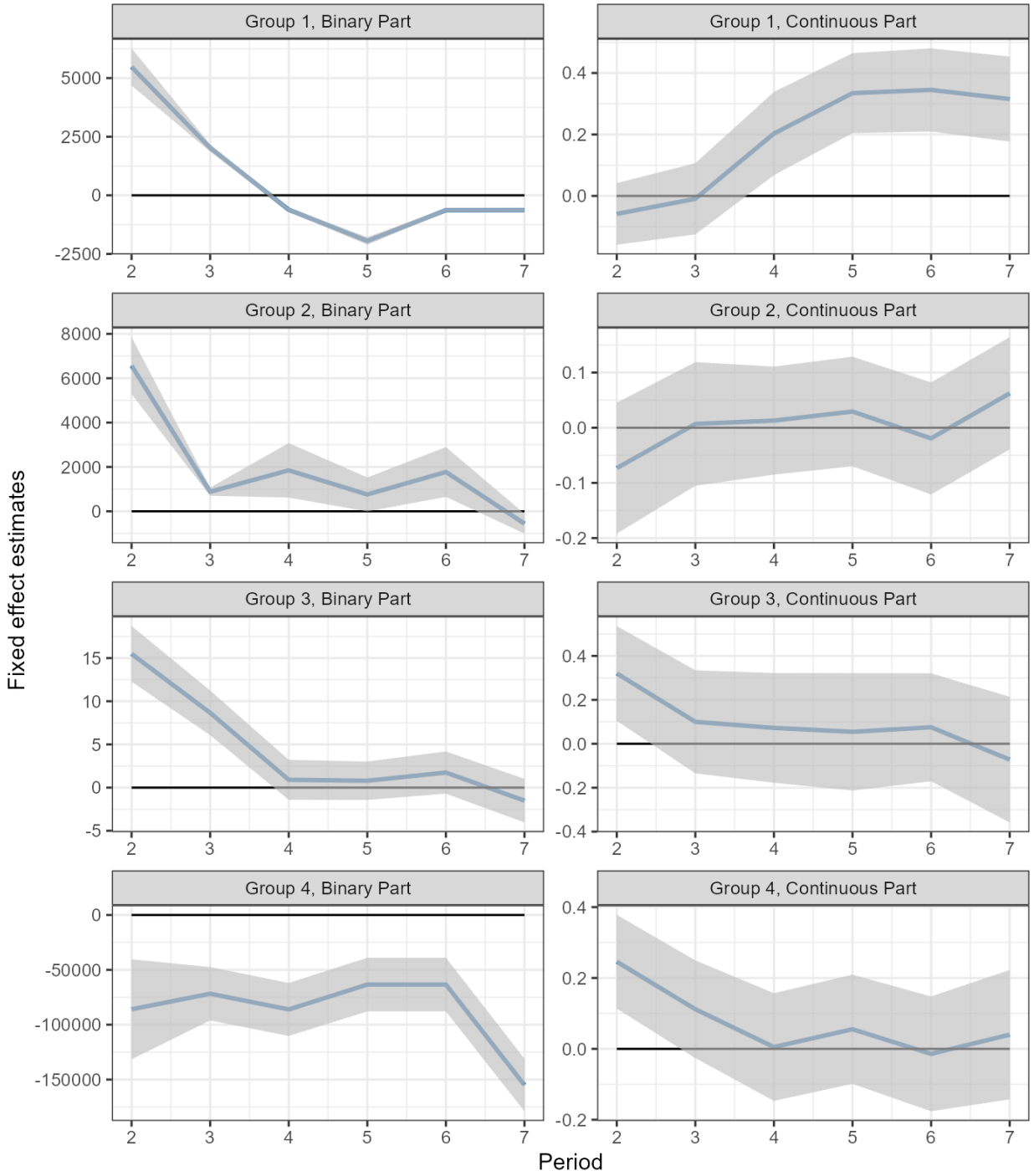


Figure 10: Time-fixed effects associated to the optimal set of estimates obtained from the EM algorithm. The value of the first time-fixed effect is equal to zero and is set as the reference value. The initial visit to the ED occurs at the end of the second period. The shaded areas correspond to the 95% cluster-robust confidence interval and do not account for uncertainty in group memberships.