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

Raphaël Langevin

McGill University

Stata Conference, July 21, 2023

Motivation

- (Unobserved) heterogeneity is practically everywhere in social sciences.
  - Let's assume that all datasets are incomplete up to some point.
  - Can lead to severe omitted-variable bias if the missing information is correlated with the observed information.
  - Perfect randomization is often impractical and expected values (β, ATE, ATET, etc.) of coefficients "mask" the heterogeneity in the distributions.
- One can try to solve the problem by grouping/clustering homogeneous units altogether.
  - Homogeneity is conditional on both observed and unobserved information.
  - This helps to recover meaningful estimates and/or to get a sense of the distribution of the coefficient(s) of interest.
  - What should we do when the true grouping pattern (e.g. the cohort in heterogeneous DID) is unobserved?
  - People do change! In panel data analysis, the true grouping pattern might change over time.

Motivation - The Frailty example

- Frailty is defined as a state of vulnerability in elders.
- Individuals who share the same frailty level will also react similarly to health adverse events and new diagnoses.
- Frailty is usually unobserved in both clinical and administrative health data.

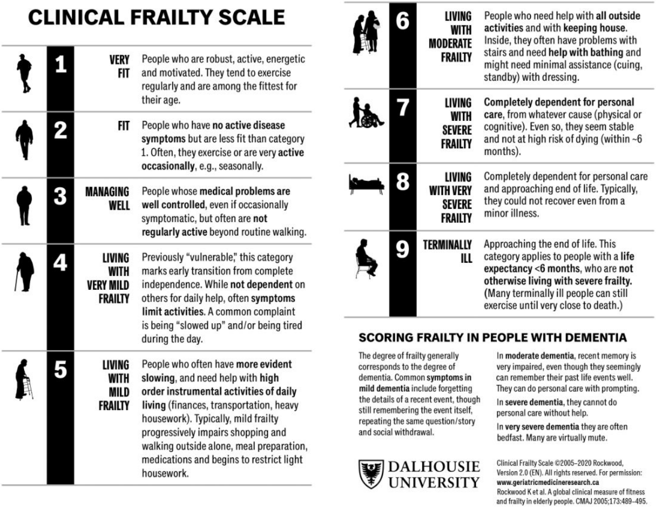


Figure 1: Clinical Frailty Scale (CFS) from Dalhousie University [Rockwood and Mitnitski, 2007]. The scale goes from 1 (Very fit) to 9 (Terminally ill).

Motivation

- Finite mixtures and latent class analysis have been extensively used to account for such unobserved heterogeneity in applied work. But not without major issues.
    - The objective function is usually multimodal, so you need to try multiple initial parameter values even in the simplest cases (abstract from that for now).
    - Estimates can be very imprecise and unstable.
    - Contrary to the common belief, I show that consistency of MLE of finite mixtures is never guaranteed in practice.
- I show how we can get consistent estimates of all parameters in the mixture by maximizing a different objective function than the objective used in both the **fmm** and **gsem** commands.
- There is no Stata command yet, but it would be easy to add an additional subcommand to the **cluster** command to integrate such a consistent estimation procedure.

General framework

- Let's define the following mixture density :

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

  - $i = \{1,..,N\}, t = \{1,..,T\}$,
  - y_{it} is a univariate outcome (discrete or continuous),
  - x_{it} is a p-sized vector of strictly exogenous covariates,
  - $f_g(\cdot|\cdot;\theta_g)$ is the density of the g^{th} component in the mixture,
  - there is $G < \infty \in \mathbb{N}^+$ groups of observations, where G is known, but the true group membership is unknown,
  - $\pi = (\pi_1, ..., \pi_G) \in \Pi$ is a vector of mixing weights to estimate, with $\pi_g \in (0,1)$ for each $g \in \{1, ..., G\} = \mathbb{G}$, and with $\sum_{g=1}^{G} \pi_g = 1$,
  - $\theta = (\theta_1, ..., \theta_G) \in \Theta \subset \mathbb{R}^{p \times G}$ contains all the parameters for each $f_g(\cdot)$.

- The *mixture log likelihood function* is defined as follows :

$$L(\theta,\pi) := \sum_{i=1}^{N} \sum_{t=1}^{T} \log(\sum_{g=1}^{G} \pi_g f_g(y_{it}|x_{it};\theta_g)) \qquad (2)$$

General framework

- For each dataset $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{NT\times(p+1)}$, there exists a set of true parameters, denoted by (θ^0, π^0), such that

$$f(y_{it}|x_{it};\theta^0,\pi^0) = \sum_{g=1}^{G} \pi_g^0 f_g(y_{it}|x_{it};\theta_g^0). \tag{3}$$

- Let's define the true grouping variable as follows

$$z_{itg}^0 := \begin{cases} 1 & \text{if and only if } y_{it} \text{ is generated by } f_g(\cdot|x_{it};\theta_g^0), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

- The g^{th} true mixing weight, π_g^0, is such that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{z_{itg}^0}{NT} \xrightarrow{p} \mathbb{E}[z_g^0] = \pi_g^0, \tag{5}$$

as N and T both tend to infinity.

Literature review

- This general setup is very flexible and has been namely used in :
  - Health economics to recover unobserved types of patients [Deb and Trivedi, 1997, 2002, Conway and Deb, 2005] and model the tail distribution of healthcare expenditures [Jones et al., 2015, 2016, Kasteridis et al., 2022];
  - Labour economics to model duration of unemployment spells and career decisions of young men [Heckman and Singer, 1984, Keane and Wolpin, 1997];
  - Econometric theory where the mixture density is estimated non-parametrically [Kasahara and Shimotsu, 2009, Compiani and Kitamura, 2016];
  - The March 2023 issue of the Stata journal [Jenkins and Rios-Avila, 2023].

- In the parametric case, the maximization of $L(\theta, \pi)$ with respect to θ and π is (almost) always carried out by the expectation-maximization (EM) algorithm [Dempster et al., 1977].

- Any algorithm that can globally maximize $L(\theta, \pi)$ with respect to θ and π is assumed to yield consistent estimates due to the strong consistency property of maximum lilkelihood estimation (MLE) [Wald, 1949, Redner and Walker, 1984, Chen, 2017].

Contributions to the literature

- "Perhaps the most troublesome implication of [the obtained results] is that, if the component densities are poorly separated, then impractically large sample sizes might be required in order to expect even moderately precise maximum-likelihood estimates." Redner and Walker [1984].

- I show that globally maximizing the mixture log likelihood function as shown in (2) does not yield consistent estimates under mild regularity conditions.

- I show that maximizing the *max-component log likelihood* function will lead to consistent estimators of all parameters in the mixture (including the mixing weights) under certain assumptions.
    - The K-means and the classification EM (CEM) algorithms both maximize this objective function.
    - K-means and CEM yield consistent estimates if group membership is assumed to be constant over time for all units [Bonhomme and Manresa, 2015].
    - Some authors have tried to relax this assumption, but never completely [Lumsdaine et al., 2023, Okui and Wang, 2021].
    - It is possible to get consistent estimates with unrestricted group membership (as claimed in the classical setup) if the G joint densities of the covariates and the outcome are asymptotically non-overlapping.

Regularity conditions

Assumption 1

1. (Generic identifiability) $f(\mathbf{y}|\mathbf{x};\theta,\pi) = f(\mathbf{y}|\mathbf{x};\theta',\pi') \Leftrightarrow \theta = \theta'$ and $\pi = \pi'$ for any dataset $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{NT\times(p+1)}$, up to any "label switching", and assuming that G is known (see Section 1.3 of Frühwirth-Schnatter [2006]).
2. (Boundedness) $\mathbb{E}_0[\log f(y_{it}|x_{it};\theta,\pi)] < \infty$ for any $\theta \in \Theta$ and any $\pi \in \Pi$.
3. (Common support) $f_g(y_{it}|x_{it};\theta_g) > 0$ for all $g \in G$ and all $\theta_g \in \Theta$, where all components' densities share the same support.
4. (Continuous differentiability) $f_g(y_{it}|x_{it};\theta_g)$ is continuously differentiable with respect to θ_g for all $g \in \mathbb{G}$.

Approximate MLE of π

- The "approximate" MLE of π_g, denoted by $\pi_g(\theta)$, is defined as follows

$$\pi_g(\theta) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{itg}(\theta,\pi) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\pi_g f_g(y_{it}|x_{it};\theta_g)}{\sum_{l=1}^{G}\pi_l f_l(y_{it}|x_{it};\theta_l)}, \quad (6)$$

where $\sum_{g=1}^{G}\pi_g(\theta) = 1$ by construction [Redner and Walker, 1984]. Proof

(In)consistency of MLE

- Under continuous differentiability of the objective function with respect to θ, consistent estimation of θ requires that

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

where $\mathbb{E}_0$ is the expected value with respect to the true mixture density.

- This is equivalent to

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

- If $\pi(\theta^0) \xrightarrow{p} \pi^0$, then the above condition reduces to

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

which is always true if the limits of the integral is not a function of θ.

(In)consistency of MLE

- If $\pi(\theta^0)$ does not converge to π^0, then MLE is inconsistent by construction for the mixing weights.

- If we don't care about π^0, we still need to show that

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

  holds independently of the value to which $\pi(\theta^0)$ will converge. This is not easy to show and will not hold in most cases. Development of a two-component mixture

- It is important to note that convergence of $\pi(\theta^0)$ to π^0 is not automatic (if we don't want to rely on any kind of circular argument). Example of a circular argument

- Therefore, we have to find a way to guarantee that $\pi(\theta^0) \xrightarrow{p} \pi^0$ as $N, T \to \infty$. The problem is similar to the incidental parameter problem in non-linear fixed effects models.

(In)consistency of MLE

- Let's recall that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{z_{itg}^0}{NT} \xrightarrow{p} \mathbb{E}[z_g^0], = \pi_g^0.$$

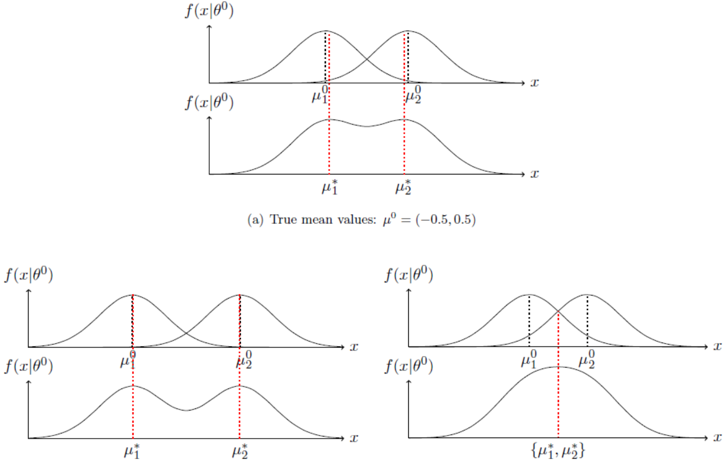
- Therefore, if

$$\tau_{itg}(\theta^0, \pi(\theta^0)) = \frac{\pi_g(\theta^0) f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{l=1}^{G}\pi_l(\theta^0) f_l(y_{it}|x_{it};\theta_l^0)} = z_{itg}^0,$$

  for all values of $(y_{it}, x_{it}) \in \mathcal{Y}|\mathcal{X}$ and all $g \in \mathbb{G}$, then we will have that $\pi(\theta^0) \xrightarrow{p} \pi^0$ as N and T tend to infinity.

- This will happen if and only if all component's densities are infinitely distant to each other (i.e. $f_g(y_{it}|x_{it};\theta_g^0) \approxeq 0$ for any $g \neq z_{it}^0$ and for all values of $(y_{it}, x_{it}) \in \mathcal{Y}|\mathcal{X}_g$, where $\cap_{g=1}^{G}\mathcal{Y}|\mathcal{X}_g = \emptyset$). Proof of convergence

- It is very easy to see it graphically.



(a) True mean values: $\mu^0 = (-0.5, 0.5)$

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

(c) True mean values: $\mu^0 = (-0.35, 0.35)$

Figure 1: Various mixtures of two normal densities with equal mixing weights and equal variances. The upper graph in each panel shows the two normal densities when they are identified separately, whereas the lower graph of each panel shows the observed mixture density (lower graphs are rescaled to improve comparability). 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.

(In)efficiency of MLE

- It is never sure that maximizing the standard mixture log likelihood will converge to the true parameter values unless $\pi(\theta^0) = \pi^0$.
- If the estimation procedure acts like if the true group membership were known for all observations, then this procedure would share the so-called *oracle property* and would be asymptotically efficient [Su et al., 2016].
- The EM algorithm can never share this property since the assignment of each observation to each group/component is probabilistic ($\tau_{itg}(\theta, \pi) > 0$).
- I show how the K-means and the CEM algorithms can share the oracle property without restricting group membership over time.

The standard CEM algorithm

- The "standard" CEM algorithm maximizes the *max-component log likelihood* function :

$$L^{MC}(\theta) := \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} z_{itg}(\theta) \log f_g(y_{it}|x_{it};\theta_g), \tag{7}$$

where

$$z_{itg}(\theta) := \begin{cases} 1 & \text{if } g = \arg\max_{l \in \mathbb{G}} f_l(y_{it}, x_{it}|\theta_l), \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

- Compare this to the standard mixture log likelihood (eq.(2)) :

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

- The mixing weights become a by-product of the estimation procedure :

$$\pi_g^{MC}(\theta) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{itg}(\theta) \xrightarrow{p} \mathbb{E}[z_g(\theta)]. \tag{9}$$

The consistent CEM algorithm

- The standard CEM algorithm is known to be inconsistent, as the K-means [Bryant and Williamson, 1978, Bryant, 1991, Celeux and Govaert, 1992].

- To make the algorithm consistent, let's use

$$z_{itg}(\theta, \check{\theta}, p) := \begin{cases} 1 & \text{if } g = \arg\max_{l \in \mathbb{G}} f_l(y_{it}, x_{it} | \theta_l, \check{\theta}_l) \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

instead of $z_{itg}(\theta)$, where

$$f_l(y_{it}, x_{it} | \theta_l, \check{\theta}_l) := f_l(y_{it} | x_{it}; \theta_l) \prod_{j=1}^{p} f_l(x_{itj} | \check{\theta}_{lj}), \tag{11}$$

and where $f_l(x_{itj} | \check{\theta}_{lj})$ is the l^{th} component's density of the j^{th} covariate in the vector x_{it}, with $\check{\theta} = (\check{\theta}_1, ..., \check{\theta}_G)$, and $\check{\theta}_l = (\check{\theta}_{l1}, ..., \check{\theta}_{lp})^\top$.

- The algorithm then alternates between an expectation/assignment step (E-step) and a conditional maximization step (M-step), just as does the EM algorithm.

The consistent CEM algorithm

Assumption 2

- The j^{th} element in x_{it}, denoted by x_{itj}, is distributed according to some true density $f_g(x_{itj}|\check{\theta}_g^0) \equiv f_g(x_{itj}|\check{\theta}^0)$ if and only if $z_{itg}^0 = 1$. Let's also define the following ratio for x_{itj} :

$$\chi_{itj}(\check{\theta}^0) := \arg\max_{g \in \mathbb{G}} \frac{f_g(x_{itj}|\check{\theta}^0)}{f_{z_{it}^0}(x_{itj}|\check{\theta}^0)},$$

  where $z_{it}^0 = g$ if and only if $z_{itg}^0 = 1$. Then we assume that

$$\mathbb{P}\left[\lim_{p\to\infty}\left(\frac{f_{z_{it}^0}(y_{it}|x_{it};\theta^0)}{f_l(y_{it}|x_{it};\theta^0)} \prod_{j:\chi_{itj}(\check{\theta}^0)\neq l} \frac{f_{z_{it}^0}(x_{itj}|\check{\theta}^0)}{f_l(x_{itj}|\check{\theta}^0)} > \prod_{j:\chi_{itj}(\check{\theta}^0)= l} \frac{f_l(x_{itj}|\check{\theta}^0)}{f_{z_{it}^0}(x_{itj}|\check{\theta}^0)}\right)\right] = 1,$$

  for any $l \in \mathbb{G}\backslash_{z_{it}^0}$ and all values of $i \in \{1, ..., N\}$ and all $t \in \{1, ..., T\}$.
- $\text{plim}_{N,T\to\infty}\, n_g^0 = \infty$ for all $g \in \mathbb{G}$, where $n_g^0 = \sum_{i=1}^N \sum_{t=1}^T z_{itg}^0$.

The consistent CEM algorithm

Theorem 3.2

Let Assumptions 1 and 2 hold. Let's also define $z_{itg}(\theta, \check{\theta}, p)$ as in eq.(10). Then $z_{itg}(\theta^0, \check{\theta}^0, p) \xrightarrow{p} z^0_{itg}$ for all values of $(i, t) \in \mathcal{Y}|\mathcal{X}$ and all $g \in \mathbb{G}$ as p tends to infinity. Proof of Theorem 3.2

- Under Assumption 2, all observations in the sample will be correctly classified at the true parameter values if the number of covariates is sufficiently large.

- Standard asymptotics and inference from MLE will be applicable component-wise if Assumption 2 hold as $N, T \to \infty$ and if there is no "cross-group" dependence.

- The rate at which N, T, and p tend to infinity can remain undetermined, as long as the classification error rate goes to zero in the limit when evaluated at the true parameter values [Dzemski and Okui, 2021].

- The second part of Assumption 2 says that the number of groups, G, cannot grow faster than the number of observations within each group.

Monte Carlo simulations

- Two simulation exercises were performed.
  - The first one shows that MLE of finite mixtures leads to inconsistent estimates.
  - The second one compares the finite-sample performance of the EM algorithm and the consistent CEM algorithm.

- The first data-generating process (DGP) is described by :

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

where $\boldsymbol{\mu}^0 = (\mu^0_1, ..., \mu^0_G)$ refers to the vector of true mean value and where z_i^0 is the true i^{th} group membership, with $\epsilon_i \sim N(0,1)$.

- The second DGP is described by :

$$y_{it} = x_{it}^\top \beta_{z_{it}^0} + \bar{x}_i^\top \gamma_{z_{it}^0} + \delta_{tz_{it}^0} + \alpha_{iz_{it}^0} + \epsilon_{it},$$
$$x_{itj} = \mu^0_{jz_{it}^0} + \nu_{it},$$

where δ_t and α_i are time-fixed and unit-random effects, respectively. All parameters vary across groups, except for $\epsilon_{it} \sim N(0,1)$ and $\nu_{it} \sim N(0,1)$. The true categorical group membership z_{it}^0 follows an AR(1) process.

Results - First DGP, $G^0 = 2$, $\pi^0 = (0.5, 0.5)$

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	**(-0.125, 0.125)**	0.00115	0.99885	-0.19774	0.00027	1.10658	1.00789
	(-0.25, 0.25)	0.97877	0.02123	-0.01772	0.81930	1.02642	0.90533
	(-0.5, 0.5)	0.60267	0.39733	-0.39864	0.60477	1.01862	0.98370
	(-1, 1)	0.49848	0.50152	-1.00223	0.99623	1.00036	1.00262
	(-2, 2)	0.50007	0.49993	-1.99996	2.00065	1.00071	1.00056
CEM	**(-0.125, 0.125)**	0.50002	0.49998	-0.80421	0.80437	0.60777	0.60754
	(-0.25, 0.25)	0.49973	0.50027	-0.82321	0.82241	0.62135	0.62142
	(-0.5, 0.5)	0.50057	0.49943	-0.89461	0.89674	0.67019	0.66946
	(-1, 1)	0.49902	0.50098	-1.16912	1.16463	0.79896	0.80082
	(-2, 2)	0.49990	0.50010	-2.01769	2.01695	0.96537	0.96637

Table 1 : Estimated values for each scenario of true mean values with $G^0 = 2$, $\pi^0 = (0.5, 0.5)$, equal unit standard errors, and $N = 1,000,000$. $\boldsymbol{\mu}^0$ = true mean values; $\hat{\pi}$ = estimated mixing weights; $\hat{\boldsymbol{\mu}}$ = estimated mean values; $\hat{\sigma}$ = estimated standard errors; CEM stands for the standard CEM algorithm.

- The results confirm the insights given by Figure 2. All biases decrease as the distance between the mean values increases.

Results - First DGP, $G^0 = 2$, $\pi^0 = (0.5, 0.5)$

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	**(-0.125, 0.125)**	0.00115	0.99885	-0.19774	0.00027	1.10658	1.00789
	(-0.25, 0.25)	0.97877	0.02123	-0.01772	0.81930	1.02642	0.90533
	(-0.5, 0.5)	0.60267	0.39733	-0.39864	0.60477	1.01862	0.98370
	(-1, 1)	0.49848	0.50152	-1.00223	0.99623	1.00036	1.00262
	(-2, 2)	0.50007	0.49993	-1.99996	2.00065	1.00071	1.00056
CEM	**(-0.125, 0.125)**	0.50002	0.49998	-0.80421	0.80437	0.60777	0.60754
	(-0.25, 0.25)	0.49973	0.50027	-0.82321	0.82241	0.62135	0.62142
	(-0.5, 0.5)	0.50057	0.49943	-0.89461	0.89674	0.67019	0.66946
	(-1, 1)	0.49902	0.50098	-1.16912	1.16463	0.79896	0.80082
	(-2, 2)	0.49990	0.50010	-2.01769	2.01695	0.96537	0.96637

Table 1 : Estimated values for each scenario of true mean values with $G^0 = 2$, $\pi^0 = (0.5, 0.5)$, equal unit standard errors, and $N = 1,000,000$. $\boldsymbol{\mu}^0$ = true mean values; $\hat{\pi}$ = estimated mixing weights; $\hat{\boldsymbol{\mu}}$ = estimated mean values; $\hat{\sigma}$ = estimated standard errors; CEM stands for the standard CEM algorithm.

- The results confirm the insights given by Figure 2. All biases decrease as the distance between the mean values increases.

Results - First DGP, $G^0 = 2$, $\pi^0 = (0.5, 0.5)$

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	**(-0.125, 0.125)**	0.00115	0.99885	-0.19774	0.00027	1.10658	1.00789
	(-0.25, 0.25)	0.97877	0.02123	-0.01772	0.81930	1.02642	0.90533
	(-0.5, 0.5)	0.60267	0.39733	-0.39864	0.60477	1.01862	0.98370
	(-1, 1)	0.49848	0.50152	-1.00223	0.99623	1.00036	1.00262
	(-2, 2)	0.50007	0.49993	-1.99996	2.00065	1.00071	1.00056
CEM	**(-0.125, 0.125)**	0.50002	0.49998	-0.80421	0.80437	0.60777	0.60754
	(-0.25, 0.25)	0.49973	0.50027	-0.82321	0.82241	0.62135	0.62142
	(-0.5, 0.5)	0.50057	0.49943	-0.89461	0.89674	0.67019	0.66946
	(-1, 1)	0.49902	0.50098	-1.16912	1.16463	0.79896	0.80082
	(-2, 2)	0.49990	0.50010	-2.01769	2.01695	0.96537	0.96637

Table 1 : Estimated values for each scenario of true mean values with $G^0 = 2$, $\pi^0 = (0.5, 0.5)$, equal unit standard errors, and $N = 1,000,000$. $\boldsymbol{\mu}^0$ = true mean values; $\hat{\pi}$ = estimated mixing weights; $\hat{\boldsymbol{\mu}}$ = estimated mean values; $\hat{\sigma}$ = estimated standard errors; CEM stands for the standard CEM algorithm.

- The results confirm the insights given by Figure 2. All biases decrease as the distance between the mean values increases.

Results - First DGP, $G^0 = 3$, $\pi^0 = (0.167, 0.33, 0.5)$

μ^0	$E(\theta^0)$ (%)	N	$L(\hat{\mu}) - L(\mu^0)$	RMSE, EM	$L^{MC}(\hat{\mu}) - L^C(\mu^0)$	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

Table 2 : Root mean square errors (RMSEs) of the estimated mean values and differences in log likelihood with $G^0 = 3$, $\pi^0 = (0.167, 0.33, 0.5)$, equal unit standard errors, and $N = 1,500,000$. $E(\theta^0)$ = error classification rate at the true parameter values, $L(\hat{\mu}) - L(\mu^0)$ = distance between the log likelihood value evaluated at the estimated mean values and the log likelihood value evaluated at the true mean values; CEM stands for the standard CEM algorithm.

Results - First DGP, $G^0 = 3$, $\pi^0 = (0.167, 0.33, 0.5)$

Algorithm (1)	μ^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

Table 3 : Estimated values for each scenario of true mean values with $G^0 = 3$, $\pi^0 = (0.167, 0.33, 0.5)$, equal unit standard errors, and $N = 1,500,000$. $\boldsymbol{\mu}^0$ = true mean values; $\hat{\pi}$ = estimated mixing weights; $\hat{\boldsymbol{\mu}}$ = estimated mean values; $\hat{\sigma}$ = estimated standard errors.

- The results show that convergence to the true values is not necessarily happening when searching for the values that maximizes the standard mixture likelihood. The conclusion is similar when maximizing the standard max-component log likelihood function.

Results - Second DGP, $G^0 = 3$, $0.25 < \pi^0 < 0.4$

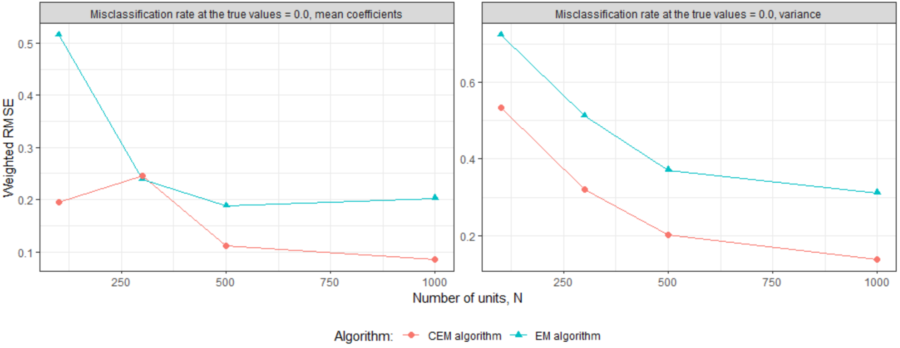


Figure 3: Weighted RMSEs as a function of N when $G^0 = 3$, the classification error rate at the true parameter values is equal to zero, and when looking at the highest log likelihood value only among all sets of initial values; The true mixing weights vary between 0.25 and 0.4 for each component and for each value of N; CEM stands for the consistent CEM algorithm.

- The consistent CEM algorithm correctly classifies all observations for all values of N in this setup, but not the EM algorithm.

Results - Second DGP, $G^0 = 3$, $0.25 < \pi^0 < 0.4$

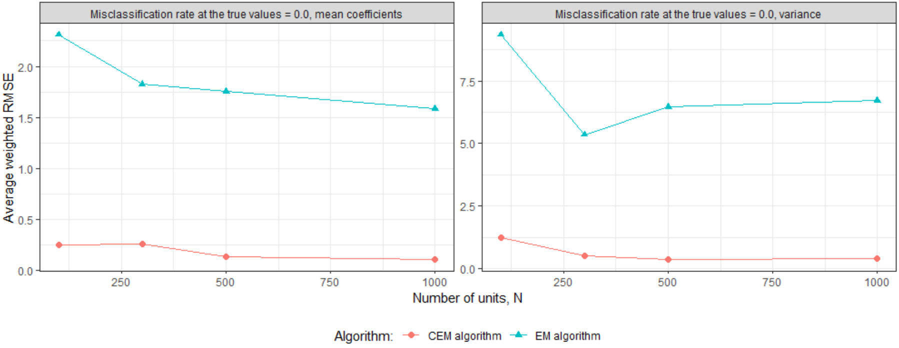


Figure 4: Average weighted RMSEs as a function of N when $G^0 = 3$, the classification error rate at the true parameter values is equal to zero, and when looking at the weighted RMSE averaged over all sets of initial values; The true mixing weights vary between 0.25 and 0.4 for each component and for each value of N; CEM stands for the consistent CEM algorithm.

- The consistent CEM algorithm yields results that are much less sensitive to the choice of initial parameter values than the EM algorithm in this setup.

Simulation results, $G^0 = 3$, $0.25 < \pi^0 < 0.4$

$E_{NT}(\theta^0)$ (%) (1)	N (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\xi}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\xi}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**100**	EM	6.60	0.5156	0.7241	2.3091	9.3371
		CEM	0.00	0.1956	0.5348	0.2441	1.2619
	300	EM	8.20	0.2391	0.5129	1.8244	5.3499
		CEM	0.00	0.2456	0.3210	0.2578	0.5261
	500	EM	6.48	0.1884	0.3715	1.7515	6.4694
		CEM	0.00	0.1123	0.2026	0.1300	0.3643
	1000	EM	6.60	0.2028	0.3122	1.5842	6.7064
		CEM	0.00	0.0848	0.1371	0.1021	0.4103
[4.1, 4.6]	**100**	EM	28.80	1.6232	0.9484	1.9916	1.4811
		CEM	5.60	0.1667	0.5840	0.7967	2.3871
	300	EM	30.33	1.4384	0.9696	1.7024	1.1605
		CEM	4.93	0.1757	0.3483	0.5302	1.5488
	500	EM	29.88	1.8206	1.0551	1.6749	1.1976
		CEM	4.80	0.1224	0.2467	0.4510	1.2071
	1000	EM	30.28	0.5142	1.0401	1.7240	1.1915
		CEM	4.78	0.0976	0.1811	0.4435	1.3069

Table 4: Simulation results when $G^0 = 3$, $T = 5$, and when the model is correctly specified; $E_{NT}(\theta) =$ Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\pi^0 = (0.422, 0.276, 0.302)$ for $N = 100$, $\pi^0 = (0.453, 0.267, 0.280)$ for $N = 300$, $\pi^0 = (0.442, 0.267, 0.291)$ for $N = 500$, and $\pi^0 = (0.439, 0.273, 0.288)$ for $N = 1000$; $\hat{\theta}^*$, $\hat{\xi}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Simulation results, $G^0 = 3$, $0.25 < \pi^0 < 0.4$

$E_{NT}(\theta^0)$ (%) (1)	N (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\boldsymbol{\xi}}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\boldsymbol{\xi}}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**100**	EM	6.60	0.5156	0.7241	2.3091	9.3371
		CEM	0.00	0.1956	0.5348	0.2441	1.2619
	300	EM	8.20	0.2391	0.5129	1.8244	5.3499
		CEM	0.00	0.2456	0.3210	0.2578	0.5261
	500	EM	6.48	0.1884	0.3715	1.7515	6.4694
		CEM	0.00	0.1123	0.2026	0.1300	0.3643
	1000	EM	6.60	0.2028	0.3122	1.5842	6.7064
		CEM	0.00	0.0848	0.1371	0.1021	0.4103
[4.1, 4.6]	**100**	EM	28.80	1.6232	0.9484	1.9916	1.4811
		CEM	5.60	0.1667	0.5840	0.7967	2.3871
	300	EM	30.33	1.4384	0.9696	1.7024	1.1605
		CEM	4.93	0.1757	0.3483	0.5302	1.5488
	500	EM	29.88	1.8206	1.0551	1.6749	1.1976
		CEM	4.80	0.1224	0.2467	0.4510	1.2071
	1000	EM	30.28	0.5142	1.0401	1.7240	1.1915
		CEM	4.78	0.0976	0.1811	0.4435	1.3069

Table 4: Simulation results when $G^0 = 3$, $T = 5$, and when the model is correctly specified; $E_{NT}(\theta) =$ Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\pi^0 = (0.422, 0.276, 0.302)$ for $N = 100$, $\pi^0 = (0.453, 0.267, 0.280)$ for $N = 300$, $\pi^0 = (0.442, 0.267, 0.291)$ for $N = 500$, and $\pi^0 = (0.439, 0.273, 0.288)$ for $N = 1000$; $\hat{\theta}^*$, $\hat{\boldsymbol{\xi}}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Simulation results, $G^0 = 3$, $0.25 < \pi^0 < 0.4$

$E_{NT}(\theta^0)$ (%) (1)	N (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\xi}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\xi}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**100**	EM	6.60	0.5156	0.7241	2.3091	9.3371
		CEM	0.00	0.1956	0.5348	0.2441	1.2619
	300	EM	8.20	0.2391	0.5129	1.8244	5.3499
		CEM	0.00	0.2456	0.3210	0.2578	0.5261
	500	EM	6.48	0.1884	0.3715	1.7515	6.4694
		CEM	0.00	0.1123	0.2026	0.1300	0.3643
	1000	EM	6.60	0.2028	0.3122	1.5842	6.7064
		CEM	0.00	0.0848	0.1371	0.1021	0.4103
[4.1, 4.6]	**100**	EM	28.80	1.6232	0.9484	1.9916	1.4811
		CEM	5.60	0.1667	0.5840	0.7967	2.3871
	300	EM	30.33	1.4384	0.9696	1.7024	1.1605
		CEM	4.93	0.1757	0.3483	0.5302	1.5488
	500	EM	29.88	1.8206	1.0551	1.6749	1.1976
		CEM	4.80	0.1224	0.2467	0.4510	1.2071
	1000	EM	30.28	0.5142	1.0401	1.7240	1.1915
		CEM	4.78	0.0976	0.1811	0.4435	1.3069

Table 4: Simulation results when $G^0 = 3$, $T = 5$, and when the model is correctly specified; $E_{NT}(\theta) =$ Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\pi^0 = (0.422, 0.276, 0.302)$ for $N = 100$, $\pi^0 = (0.453, 0.267, 0.280)$ for $N = 300$, $\pi^0 = (0.442, 0.267, 0.291)$ for $N = 500$, and $\pi^0 = (0.439, 0.273, 0.288)$ for $N = 1000$; $\hat{\theta}^*$, $\hat{\xi}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Simulation results, $G^0 = 3$, $0.088 < \pi^0 < 0.54$

$E_{NT}(\theta^0)$ (%) (1)	G (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\boldsymbol{\xi}}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\boldsymbol{\xi}}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**2**	EM	12.76	1.3986	55.1772	1.8632	58.7199
		CEM	8.92	1.1394	52.4770	1.1406	52.4023
	3	EM	4.44	0.7944	0.4050	1.4405	32.9976
		CEM	0.24	1.0632	0.3133	0.8677	12.0512
	4	EM	12.64	0.7851	0.4139	1.3928	16.9940
		CEM	53.48	0.4510	0.4734	1.0465	3.9966
4.4	**2**	EM	18.88	0.1477	4.0783	0.4165	4.4970
		CEM	9.12	0.2316	4.7860	0.2949	4.7717
	3	EM	20.28	0.3773	1.7291	0.6324	3.4111
		CEM	6.40	0.1867	0.4277	0.4889	2.3391
	4	EM	20.72	0.6745	4.2007	0.7649	2.6814
		CEM	54.60	0.7046	2.3045	0.7153	1.6646

Table 5: Simulation results when $G^0 = 3$, $N = 500$, $T = 5$, and when the model is both correctly and incorrectly specified in terms of G; $\pi^0 = (0.089, 0.535, 0.376)$ for all scenarios; $E_{NT}(\theta) =$ Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\hat{\theta}^*$, $\hat{\boldsymbol{\xi}}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Simulation results, $G^0 = 3$, $0.088 < \pi^0 < 0.54$

$E_{NT}(\theta^0)$ (%) (1)	G (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\xi}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\xi}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**2**	EM	12.76	1.3986	55.1772	1.8632	58.7199
		CEM	8.92	1.1394	52.4770	1.1406	52.4023
	3	EM	4.44	0.7944	0.4050	1.4405	32.9976
		CEM	0.24	1.0632	0.3133	0.8677	12.0512
	4	EM	12.64	0.7851	0.4139	1.3928	16.9940
		CEM	53.48	0.4510	0.4734	1.0465	3.9966
4.4	**2**	EM	18.88	0.1477	4.0783	0.4165	4.4970
		CEM	9.12	0.2316	4.7860	0.2949	4.7717
	3	EM	20.28	0.3773	1.7291	0.6324	3.4111
		CEM	6.40	0.1867	0.4277	0.4889	2.3391
	4	EM	20.72	0.6745	4.2007	0.7649	2.6814
		CEM	54.60	0.7046	2.3045	0.7153	1.6646

Table 5: Simulation results when $G^0 = 3$, $N = 500$, $T = 5$, and when the model is both correctly and incorrectly specified in terms of G; $\pi^0 = (0.089, 0.535, 0.376)$ for all scenarios; $E_{NT}(\theta) =$ Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\hat{\theta}^*$, $\hat{\boldsymbol{\xi}}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Simulation results, $G^0 = 3$, $0.088 < \pi^0 < 0.54$

$E_{NT}(\theta^0)$ (%) (1)	G (2)	Algorithm (3)	$E_{NT}(\hat{\theta}^*)$ (%) (4)	RMSE$_w$ $\hat{\xi}^*$ (5)	RMSE$_w$ $\hat{\sigma}^{2,*}$ (6)	Average RMSE$_w$, $\hat{\xi}$ (7)	Average RMSE$_w$, $\hat{\sigma}^2$ (8)
0.0	**2**	EM	12.76	1.3986	55.1772	1.8632	58.7199
		CEM	8.92	1.1394	52.4770	1.1406	52.4023
	3	EM	4.44	0.7944	0.4050	1.4405	32.9976
		CEM	0.24	1.0632	0.3133	0.8677	12.0512
	4	EM	12.64	0.7851	0.4139	1.3928	16.9940
		CEM	53.48	0.4510	0.4734	1.0465	3.9966
4.4	**2**	EM	18.88	0.1477	4.0783	0.4165	4.4970
		CEM	9.12	0.2316	4.7860	0.2949	4.7717
	3	EM	20.28	0.3773	1.7291	0.6324	3.4111
		CEM	6.40	0.1867	0.4277	0.4889	2.3391
	4	EM	20.72	0.6745	4.2007	0.7649	2.6814
		CEM	54.60	0.7046	2.3045	0.7153	1.6646

Table 5: Simulation results when $G^0 = 3$, $N = 500$, $T = 5$, and when the model is both correctly and incorrectly specified in terms of G; $\pi^0 = (0.089, 0.535, 0.376)$ for all scenarios; $E_{NT}(\theta)$ = Classification error rate evaluated at θ; RMSE$_w$ = Weighted root mean square error; $\hat{\theta}^*$, $\hat{\boldsymbol{\xi}}^*$, and $\hat{\sigma}^{2,*}$ correspond respectively to the whole set of estimated parameter values, the mean coefficient estimates, and the variance estimates that are associated with the highest log likelihood value. CEM stands for the consistent CEM algorithm.

Empirical application

- The goal is to model the healthcare expenditure (HCE) of a cohort of non-institutionalized elders using administrative data from the province of Québec, Canada.
  - I use a finite mixture of two-part models.
  - $N = 1,330$, $T = 7$, and all periods are three-months long.
  - The covariates include a comorbidity indicator, a elder's risk indicator (i.e. a poor proxy of frailty), continuity of care, and gender.

- The density of the outcome conditional on the covariates and θ is defined generally as a two-part process by :

$$f_g(y_{it}|x_{it};\theta) = \mathbb{P}[y_{it} = 0|x^b_{it};\theta_g]^{(1-d_{it})} \left[\mathbb{P}[y_{it} > 0|x^b_{it};\theta_g] f_g(y_{it}|y_{it} > 0, x_{it};\theta)\right]^{d_{it}},$$

  where x^b_{it} is the vector of covariates used in the binary part, and where d_{it} is equal to 1 if $y_{it} > 0$ and zero otherwise.
  - The binary part is a Probit model while the continuous part is log-normal, both using a Mundlak specification (as in the second simulation exercise).

- Selection of the initial parameter values and the number of groups is performed using BIC and cross-validation.

Results - Empirical application

G (1)	Algorithm (2)	Goodness-of-fit measure (3)	BIC ranking among initial values (1=lowest) (4) 1	2	3	4	5
1	-	BIC	14.6194	-	-	-	-
		$RMSE_{CV}$	2.0530	-	-	-	-
2	EM	BIC	12.5896	12.5899	12.5944	12.5997	12.6014
		$RMSE_{CV}$	1.6833	1.6512	2.2448	1.6489	2.9498
	CEM	BIC	12.6934	12.6965	12.6969	12.6988	12.7039
		$RMSE_{CV}$	1.6475	2.7029	1.3670	1.6711	1.6606
3	EM	BIC	10.6818	10.6869	10.6935	10.6993	10.7012
		$RMSE_{CV}$	1.4512	1.3019	1.3655	NA	1.4610
	CEM	BIC	11.5007	11.5017	11.5179	11.5183	11.5187
		$RMSE_{CV}$	1.8012	2.1767	1.5641	2.0958	NA
4	EM	BIC	9.6097	9.7520	9.8066	9.8331	9.8414
		$RMSE_{CV}$	1.5777	1.2052	3.3790	1.6355	1.2521
	CEM	BIC	9.5445	9.5485	9.5486	9.5908	9.6484
		$RMSE_{CV}$	1.1832*	1.8243	2.3783	2.3110	2.3756
5	EM	BIC	8.7560*	9.0540	9.2138	9.3315	9.3605
		$RMSE_{CV}$	2.7742	4.2452	3.5989	1.2022*	1.3531
	CEM	BIC	9.0361*	9.1735	9.3390	9.3458	9.3693
		$RMSE_{CV}$	1.2165	NA	NA	NA	NA

Table 6 : BIC values and root mean squared errors obtained by grouped 10-fold cross-validation for each one of the five smallest BIC values obtained by random initialization; BIC = Bayesian information criterion, $RMSE_{CV}$ = Cross-validated root mean squared error (on the log outcome).

Results - Empirical application

G	Algorithm	Goodness-of-fit measure	BIC ranking among initial values (1=lowest) (4)				
(1)	(2)	(3)	1	2	3	4	5
1	-	BIC	14.6194	-	-	-	-
		$RMSE_{CV}$	2.0530	-	-	-	-
2	EM	BIC	12.5896	12.5899	12.5944	12.5997	12.6014
		$RMSE_{CV}$	1.6833	1.6512	2.2448	1.6489	2.9498
	CEM	BIC	12.6934	12.6965	12.6969	12.6988	12.7039
		$RMSE_{CV}$	1.6475	2.7029	1.3670	1.6711	1.6606
3	EM	BIC	10.6818	10.6869	10.6935	10.6993	10.7012
		$RMSE_{CV}$	1.4512	1.3019	1.3655	NA	1.4610
	CEM	BIC	11.5007	11.5017	11.5179	11.5183	11.5187
		$RMSE_{CV}$	1.8012	2.1767	1.5641	2.0958	NA
4	EM	BIC	9.6097	9.7520	9.8066	9.8331	9.8414
		$RMSE_{CV}$	1.5777	1.2052	3.3790	1.6355	1.2521
	CEM	BIC	9.5445	9.5485	9.5486	9.5908	9.6484
		$RMSE_{CV}$	1.1832*	1.8243	2.3783	2.3110	2.3756
5	EM	BIC	8.7560*	9.0540	9.2138	9.3315	9.3605
		$RMSE_{CV}$	2.7742	4.2452	3.5989	1.2022*	1.3531
	CEM	BIC	9.0361*	9.1735	9.3390	9.3458	9.3693
		$RMSE_{CV}$	1.2165	NA	NA	NA	NA

Table 6 : BIC values and root mean squared errors obtained by grouped 10-fold cross-validation for each one of the five smallest BIC values obtained by random initialization; BIC = Bayesian information criterion, $RMSE_{CV}$ = Cross-validated root mean squared error (on the log outcome).

Results per group

Group number (1)	Estimated mixing weights (2)	Moment (3)	Male (4)	ERA (5)	Time-averaged ERA (6)	COCI (7)	Time-averaged COCI (8)	Time-averaged Charlson (9)
Global	1.000	Mean	0.3770	2.0534	2.0534	3.1861	3.9952	1.3493
		Variance	0.2349	2.9624	2.6056	9.8073	4.3050	1.3419
1	0.1438	Mean	0.3737	1.8345	1.8006	10.0000	5.8652	1.1747
		Variance	0.2342	2.4974	2.1076	0.0000	2.9505	1.0185
2	0.2883	Mean	0.4085	3.8076	3.6534	2.1609	3.0130	2.1931
		Variance	0.2417	2.4567	2.1792	2.2577	2.2839	2.1028
3	0.3340	Mean	0.3488	1.1287	1.2804	2.7342	3.8148	1.0170
		Variance	0.2272	0.6422	0.8164	2.9500	2.9308	0.4885
4	0.2338	Mean	0.3918	1.4066	1.4009	0.9982	4.4319	0.9307
		Variance	0.2384	1.6886	1.4736	0.0001	5.6768	0.5798

Table 7 : Descriptive statistics of the observations contained within each group created by the preferred specification when using the CEM algorithm.

Results per group

Group number (1)	Estimated mixing weights (2)	Moment (3)	Male (4)	ERA (5)	Time-averaged ERA (6)	COCI (7)	Time-averaged COCI (8)	Time-averaged Charlson (9)
Global	1.000	Mean	0.3770	2.0534	2.0534	3.1861	3.9952	1.3493
		Variance	0.2349	2.9624	2.6056	9.8073	4.3050	1.3419
1	0.1438	Mean	0.3737	1.8345	1.8006	10.0000	5.8652	1.1747
		Variance	0.2342	2.4974	2.1076	0.0000	2.9505	1.0185
2	0.2883	Mean	0.4085	3.8076	3.6534	2.1609	3.0130	2.1931
		Variance	0.2417	2.4567	2.1792	2.2577	2.2839	2.1028
3	0.3340	Mean	0.3488	1.1287	1.2804	2.7342	3.8148	1.0170
		Variance	0.2272	0.6422	0.8164	2.9500	2.9308	0.4885
4	0.2338	Mean	0.3918	1.4066	1.4009	0.9982	4.4319	0.9307
		Variance	0.2384	1.6886	1.4736	0.0001	5.6768	0.5798

Table 7 : Descriptive statistics of the observations contained within each group created by the preferred specification when using the CEM algorithm.

Results - Empirical application

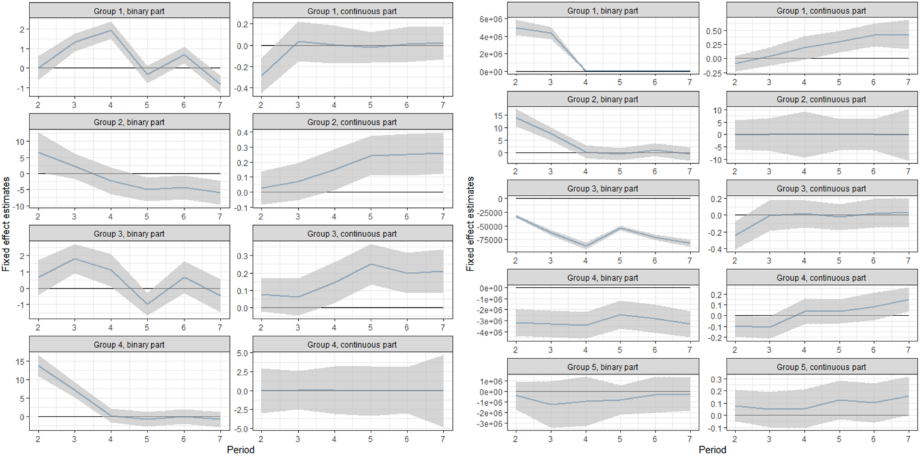


Figure 5: Estimates of the time-fixed effects from the preferred specification with the CEM (left graphs) and the EM (right graphs) algorithms for each group and each part of the model; The shaded areas correspond to the cluster(unit)-robust 95% confidence interval and do not account for uncertainty in group membership.

Results - Empirical application

Coefficients	Group/Component			
	1	2	3	4
	Binary part			
Time-varying ERA	0.1425	-1.8204**	-1.6829***	0.4866
	(0.1124)	(0.6171)	(0.3422)	(0.5837)
Time-averaged Charlson	0.1325	-2.2330***	-0.4526	1.5104***
	(0.0843)	(0.4238)	(0.4150)	(0.4358)
Time-averaged COCI	0.7436***	-3.2721***	2.8710***	-0.8521***
	(0.0530)	(0.4048)	(0.3252)	(0.1633)
Time-averaged ERA	-0.5766***	-0.6218	2.6422***	1.0780
	(0.1254)	(0.6902)	(0.3809)	(0.6360)
Male	-0.4018*	-1.5790	-3.5767***	-0.6795
	(0.1646)	(1.2512)	(0.3580)	(0.6821)
N	1330	2666	3088	2162

Table 8 : Additional estimates of the binary part of the preferred specification obtained with the CEM algorithm; Fully robust standard errors are shown in parenthesis; * = p-value $<$ 0.05, ** = p-value $<$ 0.01, *** = p-value $<$ 0.001.

Results - Empirical application

Coefficients	Group/Component			
	1	2	3	4
	Continuous part			
Time-varying ERA	0.0459	0.0061	-0.4117***	0.0102
	(0.0504)	(0.0312)	(0.0400)	(0.8758)
Time-varying COCI	0.2813	-0.1484***	-0.0548***	-44.7446
	(68455.1668)	(0.0138)	(0.0115)	(25.4370)
Time-averaged Charlson	0.0031	0.0569**	0.1041***	-0.0248
	(0.0350)	(0.0208)	(0.0305)	(0.6384)
Time-averaged COCI	0.0141	-0.0634**	-0.0080	-0.0533
	(0.0164)	(0.0196)	(0.0133)	(0.2976)
Time-averaged ERA	0.0495	0.1108**	0.5744***	-0.0074
	(0.0535)	(0.0361)	(0.0368)	(1.0346)
Male	-0.0434	0.0790	-0.0196	0.1092
	(0.0620)	(0.0615)	(0.0470)	(1.1032)
N	1330	2548	3088	810

Table 9 : Additional estimates of the continuous part of the preferred specification obtained with the CEM algorithm; Fully robust standard errors are shown in parenthesis; * = p-value$<$ 0.05, ** = p-value$<$ 0.01, *** = p-value$<$ 0.001.

Results - Empirical application

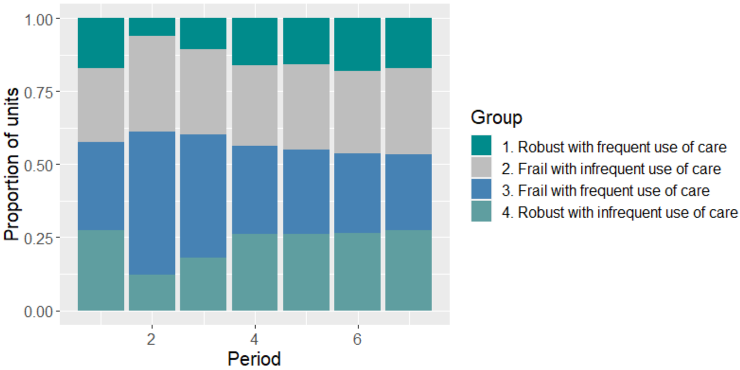


Figure 6: Proportions of the total number of observations in each group at each period (from the preferred specification with the CEM algorithm). The total number of observations at each period is equal to {1,330;1,330;1,330;1,326;1,317;1,308;1,305}.

- The dynamic behaviour of the group membership is modeled in the second step. The first step consistently estimates the group membership [Bonhomme et al., 2019, 2022].

Results - Empirical application

Group at t	Group at $t+1$			
	1	2	3	4
1	0.2284	0.1831	0.3330	0.2554
2	0.1044	0.7587	0.0386	0.0983
3	0.1339	0.0626	0.5708	0.2328
4	0.1657	0.1219	0.3584	0.3540

Figure 7: Transition matrix estimated from the grouping variable based on the preferred specification estimated with the CEM algorithm.

- Transitions into "frailty" (groups 2 and 3) are more likely than transitions out of "frailty" groups.
- Using exclusively group membership at period t to predict group membership at period $t + 1$ correctly classifies 52.2% of all observations.
- Using a dynamic multinomial logit model with all other covariates increases this percentage to 61.6% (with only one lag).

Conclusion

- Simulation results confirm that maximizing the standard likelihood of a mixture density leads to inconsistent estimates if the components' densities are not infinitely distant from each other.
- Simulation results also show that the consistent CEM algorithm produces less biased and more stable estimates than the EM algorithm in finite samples.
- Estimation results using healthcare expenditures show that the consistent CEM algorithm yields more credible estimates with smaller out-of-sample prediction errors than the EM algorithm.
- A two-step procedure is warranted to model the dynamics of the latent variable under conditional independence of the outcome from past groupings.
- The computational burden is an issue : more reliable and faster algorithms need to be developed to reach the global maximum of the objective function.
- All specifications in the first step are static. Introducing lagged dependent variables and feedback effects are left for further research.

Acknowledgements

- Thanks to Erin C. Strumpf, Saraswata Chaudhuri, Philippe Goulet Coulombe, William MacCausland, and Olivier Gagnon for helpful comments on early drafts of this paper.
- Thanks to the participants of NY Camp Econometrics XVII and of the 18^{th} CIREQ PhD Student Conference in Montréal for advice and comments.
- Thanks to Marie-Josée Sirois and the Centre d'excellence sur le vieillissement de Québec (CEVQ) for their support and for providing the data for the empirical application.
- This work was supported by the SSHRC grant 767-2020-2809.

SSHRC CRSH
CRSH SSHRC

UNIVERSITÉ McGill

References I

Stéphane Bonhomme and Elena Manresa. Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184, 2015. ISSN 0012-9682. URL http://www.jstor.org/stable/43616962. Publisher: [Wiley, The Econometric Society].

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. A Distributional Framework for Matched Employer Employee Data. *Econometrica*, 87(3): 699–739, 2019. ISSN 1468-0262. doi: 10.3982/ECTA15722. URL http://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15722. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15722.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing Unobserved Heterogeneity. *Econometrica*, 90(2):625–643, 2022. ISSN 1468-0262. doi: 10.3982/ECTA15238. URL http://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15238. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15238.

Peter Bryant and John A. Williamson. Asymptotic Behaviour of Classification Maximum Likelihood Estimates. *Biometrika*, 65(2):273–281, 1978. ISSN 0006-3444. doi: 10.2307/2335205. URL http://www.jstor.org/stable/2335205. Publisher: [Oxford University Press, Biometrika Trust].

Peter G. Bryant. Large-sample results for optimization-based clustering methods. *Journal of Classification*, 8(1):31–44, January 1991. ISSN 0176-4268, 1432-1343. doi: 10.1007/BF02616246. URL https://link.springer.com/10.1007/BF02616246.

Gilles Celeux. EM Methods for Finite Mixtures. In Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert, editors, *Handbook of Mixture Analysis*, pages 21–39. Chapman and Hall/CRC, Boca Raton, Florida : CRC Press, [2019], 1 edition, January 2019. ISBN 978-0-429-05591-1. doi: 10.1201/9780429055911-2. URL https://www.taylorfrancis.com/books/9780429508240/chapters/10.1201/9780429055911-2.

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

References III

Jiahua Chen. Consistency of the MLE under Mixture Models. *Statistical Science*, 32(1), February 2017. ISSN 0883-4237. doi: 10.1214/16-STS578. URL https://projecteuclid.org/journals/statistical-science/volume-32/issue-1/Consistency-of-the-MLE-under-Mixture-Models/10.1214/16-STS578.full.

Giovanni Compiani and Yuichi Kitamura. Using mixtures in econometric models: a brief review and some new results. *The Econometrics Journal*, 19(3):C95–C127, 2016. ISSN 1368-423X. doi: 10.1111/ectj.12068. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12068. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ectj.12068.

Karen Smith Conway and Partha Deb. Is prenatal care really ineffective? Or, is the 'devil' in the distribution? *Journal of Health Economics*, 24(3):489–513, May 2005. ISSN 01676296. doi: 10.1016/j.jhealeco.2004.09.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629605000068.

Partha Deb and Pravin K. Trivedi. Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics*, 12(3,):313–336, 1997. URL http://www.jstor.org/stable/2285252.

References IV

Partha Deb and Pravin K. Trivedi. The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, 21(4):601–625, July 2002. ISSN 01676296. doi: 10.1016/S0167-6296(02)00008-5. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629602000085.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL http://www.jstor.org/stable/2984875. Publisher: [Royal Statistical Society, Wiley].

Andreas Dzemski and Ryo Okui. Convergence rate of estimators of clustered panel models with misclassification. *Economics Letters*, 203:109844, June 2021. ISSN 01651765. doi: 10.1016/j.econlet.2021.109844. URL https://linkinghub.elsevier.com/retrieve/pii/S016517652100121X.

Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer series in statistics. Springer, New York, 2006. ISBN 978-0-387-32909-3. OCLC: ocm71262594.

Alexander Gepperth and Benedikt Pfülb. Gradient-Based Training of Gaussian Mixture Models for High-Dimensional Streaming Data. *Neural Processing Letters*, 53(6):4331–4348, December 2021. ISSN 1370-4621, 1573-773X. doi: 10.1007/s11063-021-10599-3. URL https://link.springer.com/10.1007/s11063-021-10599-3.

J. Heckman and B. Singer. A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, 52(2): 271–320, 1984. ISSN 0012-9682. doi: 10.2307/1911491. URL https://www.jstor.org/stable/1911491. Publisher: [Wiley, Econometric Society].

Yingyao Hu and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, November 2012. ISSN 03044076. doi: 10.1016/j.jeconom.2012.05.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407612001479.

References VI

Stephen P. Jenkins and Fernando Rios-Avila. Finite mixture models for linked survey and administrative data: Estimation and postestimation. *The Stata Journal: Promoting communications on statistics and Stata*, 23(1):53–85, March 2023. ISSN 1536-867X, 1536-8734. doi: 10.1177/1536867X231161976. URL http://journals.sagepub.com/doi/10.1177/1536867X231161976.

Andrew M. Jones, James Lomas, and Nigel Rice. Healthcare Cost Regressions: Going Beyond the Mean to Estimate the Full Distribution. *Health Economics*, 24(9):1192–1212, 2015. ISSN 1099-1050. doi: 10.1002/hec.3178. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3178. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.3178.

Andrew M. Jones, James Lomas, Peter T. Moore, and Nigel Rice. A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(4):951–974, 2016. ISSN 1467-985X. doi: 10.1111/rssa.12141. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12141. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12141.

References VII

Hiroyuki Kasahara and Katsumi Shimotsu. Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica*, 77(1):135–175, 2009. ISSN 1468-0262. doi: 10.3982/ECTA6763. URL http://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6763. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6763.

Panagiotis Kasteridis, Nigel Rice, and Rita Santos. Heterogeneity in end of life health care expenditure trajectory profiles. *Journal of Economic Behavior & Organization*, 204:221–251, December 2022. ISSN 01672681. doi: 10.1016/j.jebo.2022.10.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0167268122003675.

Michael P. Keane and Kenneth I. Wolpin. The Career Decisions of Young Men. *Journal of Political Economy*, 105(3):473–522, June 1997. ISSN 0022-3808, 1537-534X. doi: 10.1086/262080. URL https://www.journals.uchicago.edu/doi/10.1086/262080.

Robin L. Lumsdaine, Ryo Okui, and Wendun Wang. Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics*, 233(1):45–65, 2023. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2022.01.001. URL https://www.sciencedirect.com/science/article/pii/S0304407622000033.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite Mixture Models. page 26, 2019.

Ryo Okui and Wendun Wang. Heterogeneous structural breaks in panel data models. *Journal of Econometrics*, 220(2):447–473, February 2021. ISSN 03044076. doi: 10.1016/j.jeconom.2020.04.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407620301287.

Richard A. Redner and Homer F. Walker. Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review*, 26(2):195–239, 1984. ISSN 0036-1445. URL http://www.jstor.org/stable/2030064. Publisher: Society for Industrial and Applied Mathematics.

K. Rockwood and A. Mitnitski. Frailty in Relation to the Accumulation of Deficits. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62(7):722–727, July 2007. ISSN 1079-5006, 1758-535X. doi: 10.1093/gerona/62.7.722. URL https://academic.oup.com/biomedgerontology/article-lookup/doi/10.1093/gerona/62.7.722.

Liangjun Su, Zhentao Shi, and Peter C. B. Phillips. Identifying Latent Structures in Panel Data. *Econometrica*, 84(6):2215–2264, 2016. ISSN 1468-0262. doi: 10.3982/ECTA12560. URL http://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12560. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12560.

Kentaro Tanaka. 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, 2009. ISSN 0303-6898. URL http://www.jstor.org/stable/41000314. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

Abraham Wald. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949. ISSN 0003-4851. URL http://www.jstor.org/stable/2236315. Publisher: Institute of Mathematical Statistics.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983. ISSN 0090-5364. URL http://www.jstor.org/stable/2240463. Publisher: Institute of Mathematical Statistics.

- Let's define the g^{th} *posterior probability* of the it^{th} observation as follows :

$$\tau_{itg}(\theta, \pi) := \frac{\pi_g f_g(y_{it}|x_{it}; \theta_g)}{f(y_{it}|x_{it}; \theta, \pi)} = \frac{\pi_g f_g(y_{it}|x_{it}; \theta_g)}{\sum_{l=1}^{G} \pi_l f_l(y_{it}|x_{it}; \theta_l)}, \tag{12}$$

- The probability $\tau_{itg}(\theta, \pi)$ represents the probability that the it^{th} observation has arisen from the g^{th} group's/component's density.

- This comes from a direct application of Bayes' rule on the unobserved grouping variable z^0_{itg}.

- Recall that

$$L(\theta, \pi) = \sum_{i=1}^{N} \sum_{t=1}^{T} \log(\sum_{g=1}^{G} \pi_g f_g(y_{it}|x_{it}; \theta_g)).$$

"Approximate" MLE of π

- The mixture log likelihood function can be rewritten as [Dempster et al., 1977, Celeux, 2019] :

$$L(\theta,\pi) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \tau_{itg} \log(\pi_g f_g(y_{it}|x_{it};\theta_g)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G} \tau_{itg} \log \tau_{itg},$$

where $\tau_{itg} \equiv \tau_{itg}(\theta,\pi)$, as defined above by eq.(12).

- If τ_{itg} is taken as given (the "approximation"), then we have that

$$\frac{\partial L(\theta,\pi)}{\partial \pi_g} = \frac{\partial}{\partial \pi_g} \sum_{g=1}^{G} \log \pi_g \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{itg}.$$

- By the properties of the cross-entropy function, the RHS is maximized when $\pi_g = \alpha \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{itg}$ where α is a normalizing constant. Imposing the unit constraint $\sum_{g=1}^{G} \pi_g = 1$ directly leads to $\alpha = \frac{1}{NT}$. Back to Assumption 1

Development of the two-component mixture

- Let's look at the following two-component mixture density with no covariates.

$$f(y_i|\theta,\pi(\theta)) := \pi_1(\theta)f_1(y_i|\theta_1) + \pi_2(\theta)f_2(y_i|\theta_2).$$

- Note that $\pi_2(\theta) = 1 - \pi_1(\theta)$ and that $\pi_2'(\theta) = -\pi_1'(\theta)$, where the prime notation stands as the derivative with respect to θ. Hence, we get that the first-order consistency condition can be written as follows

$$\mathbb{E}_0\left[\frac{\pi_1'(\theta)(f_1(y_i|\theta_1^0) - f_2(y_i|\theta_2^0))}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Bigg|_{\theta=\theta^0} +$$
$$\mathbb{E}_0\left[\frac{\pi_1(\theta^0)(f_1'(y_i|\theta_1) - f_2'(y_i|\theta_2))}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Bigg|_{\theta=\theta^0} + \mathbb{E}_0\left[\frac{f_2'(y_i|\theta_2)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Bigg|_{\theta=\theta^0} = 0.$$

- For simplicity, let's define the asymptotic mixing weights as follows

$$\pi_1(\theta^0) := \mathbb{E}_0\left[\frac{f_1(y_i|\theta_1^0)}{f_1(y_i|\theta_1^0) + f_2(y_i|\theta_2^0)}\right] > 0,$$

which leads to the following derivative :

$$\pi_1'(\theta)\big|_{\theta=\theta^0} := \mathbb{E}_0\left[\frac{f_1'(y_i|\theta_1)f_2(y_i|\theta_2^0) - f_1(y_i|\theta_1^0)f_2'(y_i|\theta_2)}{(f_1(y_i|\theta_1^0) + f_2(y_i|\theta_2^0))^2}\right]\Bigg|_{\theta=\theta^0} \lesseqgtr 0.$$

Development of the two-component mixture

- If we look only at the first term in the condition, we have that it is equivalent to

$$\mathbb{E}_0\left[\frac{f_1(y_i|\theta_1^0)-f_2(y_i|\theta_2^0)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\mathbb{E}_0\left[\frac{f_1'(y_i|\theta_1)f_2(y_i|\theta_2^0)-f_1(y_i|\theta_1^0)f_2'(y_i|\theta_2)}{(f_1(y_i|\theta_1^0)+f_2(y_i|\theta_2^0))^2}\right]\Big|_{\theta=\theta^0},$$

while the second term is equivalent to

$$\mathbb{E}_0\left[\frac{f_1(y_i|\theta_1^0)}{f_1(y_i|\theta_1^0)+f_2(y_i|\theta_2^0)}\right]\mathbb{E}_0\left[\frac{f_1'(y_i|\theta_1)-f_2'(y_i|\theta_2)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Big|_{\theta=\theta^0}.$$

- Hence, the condition becomes

$$\mathbb{E}_0\left[\frac{f_1(y_i|\theta_1^0)-f_2(y_i|\theta_2^0)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\mathbb{E}_0\left[\frac{f_1'(y_i|\theta_1)f_2(y_i|\theta_2^0)-f_1(y_i|\theta_1^0)f_2'(y_i|\theta_2)}{(f_1(y_i|\theta_1^0)+f_2(y_i|\theta_2^0))^2}\right]\Big|_{\theta=\theta^0}+$$
$$\mathbb{E}_0\left[\frac{f_1'(y_i|\theta_1)-f_2'(y_i|\theta_2)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Big|_{\theta=\theta^0}\mathbb{E}_0\left[\frac{f_1(y_i|\theta_1^0)}{f_1(y_i|\theta_1^0)+f_2(y_i|\theta_2^0)}\right]+$$
$$\mathbb{E}_0\left[\frac{f_2'(y_i|\theta_2)}{f(y_i|\theta^0,\pi(\theta^0))}\right]\Big|_{\theta=\theta^0}=0$$

- If $\pi(\theta^0)\neq\pi^0$, the last two terms will not be equal to zero unless the two densities $f_1(\cdot)$ and $f_2(\cdot)$ are equal or infinitely distant to each other. Back to MLE

Example of a circular argument

- Using the definition of $\pi(\theta)$ from equation (6) and the WLLN, we have that

$$\pi_g(\theta^0) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{l=1}^{G}\pi_l(\theta^0)f_l(y_{it}|x_{it};\theta_l^0)} \xrightarrow{p} \mathbb{E}\left[\frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{l=1}^{G}\pi_l(\theta^0)f_l(y_{it}|x_{it};\theta_l^0)}\right]$$

as N and T tend to infinity.

- If it is true that

$$\mathbb{E}\left[\frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{l=1}^{G}\pi_l(\theta^0)f_l(y_{it}|x_{it};\theta_l^0)}\right] = \pi_g^0,$$

for all $g \in \mathbb{G}$, then we do get that $\pi(\theta^0) \xrightarrow{p} \pi^0$ for all $g \in \mathbb{G}$.

- The above equation is equivalent to

$$\int_{\mathcal{Y}} \frac{\pi_g^0(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\sum_{l=1}^{G}\pi_l(\theta^0)f_l(y_{it}|x_{it};\theta_l^0)}\sum_{l=1}^{G}\pi_l^0 f_l(y_{it}|x_{it};\theta_l^0)\upsilon(dy_{it}) = \pi_g^0,$$

which will be true if $\pi_g(\theta^0) = \pi_g^0$ for all $g \in \mathbb{G}$. This leads to a circular reasoning that does not prove anything. Back to MLE

Proof of convergence of $\pi(\theta^0)$ to π^0

- Let's impose that $f_g(y_{it}|x_{it};\theta_g^0) = 0$ for any $g \neq z_{it}^0$ and for all values of $(y_{it}, x_{it}) \in \mathcal{Y}|\mathcal{X}_g$, where $\cap_{g=1}^G \mathcal{Y}_g = \emptyset$.

- In this case, we can write that

$$\mathbb{E}\left[\frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{f(y_{it}|x_{it};\theta^0,\pi(\theta^0))}\right] = \int_{\mathcal{Y}} \frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{f(y_{it}|x_{it};\theta^0,\pi(\theta^0))} f(y_{it}|x_{it};\theta^0,\pi^0)\upsilon(dy_{it}),$$
$$= \int_{\mathcal{Y}_g} \frac{\pi_g(\theta^0)f_g(y_{it}|x_{it};\theta_g^0)}{\pi_{z_{it}^0}(\theta^0)f_{z_{it}^0}(y_{it}|x_{it};\theta_{z_{it}^0}^0)} \pi_{z_{it}^0}^0 f_{z_{it}^0}(y_{it}|x_{it};\theta_{z_{it}^0}^0)\upsilon(dy_{it}),$$
$$= \int_{\mathcal{Y}_g} \pi_{z_{it}^0}^0 f_{z_{it}^0}(y_{it}|x_{it};\theta_{z_{it}^0}^0)\upsilon(dy_{it}) = \pi_{z_{it}^0}^0,$$

where the second equality comes from the fact that $f_g(y_{it}|x_{it};\theta_g^0) = 0$ for any $g \neq z_{it}^0$.

- Therefore, $\pi_g^0 \xrightarrow{p} \pi_g^0$ will be true if $f_g(y_{it}|x_{it};\theta_g^0) = 0$ for any $g \neq z_{it}^0$, which will happen if and only if all component's densities are infinitely distant from each other under Assumption 1. Back to MLE

Proof of Theorem 3.2

- Let's define the categorical assignment variable, $z_{it}(\theta^0, p)$, as follows :

$$z_{it}(\theta^0, p) \equiv z_{it}(\theta^0, \check{\theta}^0, p) := \arg\max_{g \in \mathbb{G}} \frac{f_g(y_{it}|x_{it};\theta^0)}{f_{z^0_{it}}(y_{it}|x_{it};\theta^0)} \prod_{j=1}^{p} \frac{f_g(x_{itj}|\check{\theta}^0)}{f_{z^0_{it}}(x_{itj}|\check{\theta}^0)},$$

then we have that

$$\lim_{p\to\infty} \frac{f_l(y_{it}|x_{it};\theta^0)}{f_{z^0_{it}}(y_{it}|x_{it};\theta^0)} \prod_{j=1}^{p} \frac{f_l(x_{itj}|\check{\theta}^0)}{f_{z^0_{it}}(x_{itj}|\check{\theta}^0)} = \lim_{p\to\infty} \frac{f_l(y_{it}|x_{it};\theta^0)}{f_{z^0_{it}}(y_{it}|x_{it};\theta^0)} \prod_{j:\chi_{itj}(\check{\theta}^0)\neq l} \frac{f_l(x_{itj}|\check{\theta}^0)}{f_{z^0_{it}}(x_{itj}|\check{\theta}^0)} \times \prod_{j:\chi_{itj}(\check{\theta}^0)= l} \frac{f_l(x_{itj}|\check{\theta}^0)}{f_{z^0_{it}}(x_{itj}|\check{\theta}^0)},$$

$$\lim_{p\to\infty} \frac{f_l(y_{it}|x_{it};\theta^0)}{f_{z^0_{it}}(y_{it}|x_{it};\theta^0)} \prod_{j=1}^{p} \frac{f_l(x_{itj}|\check{\theta}^0)}{f_{z^0_{it}}(x_{itj}|\check{\theta}^0)} < 1,$$

for any $l \in \mathbb{G}\backslash_{z^0_{it}}$ and all (i,t) pairs as a direct consequence of Assumption 2. This leads to $z_{itg}(\theta^0, p) \xrightarrow{a.s.} z^0_{itg}$ for all (i,t) pairs and all $g \in \mathbb{G}$ as p tends to infinity, thus implying $z_{itg}(\theta^0, p) \xrightarrow{p} z^0_{itg}$ for all (i,t) pairs and all $g \in \mathbb{G}$ as p tends to infinity. Back to Theorem 3.2