Fast Mini-Batch EM and MCEM Algorithms

Belhal Karimi, Marc Lavielle, Eric Moulines
CMAP, Ecole Polytechnique, Universite Paris-Saclay, 91128 Palaiseau, France
belhal.karimi@polytechnique.edu

March 23, 2018

Abstract

The EM algorithm is one of the most popular algorithm for inference in latent data models. For large datasets, an incremental variant has been proposed in which the E-step is computed for only a mini-batch of observations. In this paper, we propose and analyse the Monte Carlo version of the incremental EM (MCIEM) in which the conditional expectation is evaluated by a Markov Chain Monte Carlo (MCMC). Various applications are presented in this contribution showing convergence of the estimated parameters and the evolution of the convergence rates with respect to the mini-batch size.

1 Introduction

Many problems in computational statistics reduce to maximising a function, defined on a feasible set Θ, of the following form:

$$g(\theta) \triangleq \int_{\mathsf{Z}} f(z, \theta)\mu(\mathrm{d}z) \tag{1}$$

with f a strictly positive function μ almost everywhere. In the incomplete data framework, the function g is the incomplete data likelihood, z is the missing data vector and f is the complete data likelihood, that is the likelihood of the observations and the missing data, with respect to the measure μ. When the direct derivation of this expression is hard, several methods use the complete model to iteratively find the quantity of interest. The EM algorithm has been the object of considerable interest since its presentation by Dempster, Laird and Rubin in 1977. It has been relatively effective in context of maximum likelihood estimation for incomplete models. This algorithm is monotonic in likelihood making it a stable tool to work with. Many improvements have been provided since the birth of this algorithm. In particular, in (R. Neal, 2007), the authors demonstrate, in the context of mixture models, an incremental EM variant where a single observation is considered at each iteration that converges twice as fast as standard EM in a mixture estimation problem. In terms of efficiency of computation, (Cappe and Moulines, 2009; Cappe, 2009) introduced an online version where the whole dataset is not analyzed at each iteration but a growing batch of it only. In 1994, (Hudson and Larkin, 1994) introduced the ordered subsets EM (OS-EM) algorithm for accelerated image reconstruction using emission tomography data. This EM variant suggests considering ordered subsets of the data at each iteration. A block sequential variant used for Maximum A Posteriori estimation, and its convergence properties for the Poisson model, requiring fewer assumptions, was given in (Pierro and Yamagishi, 2001). Since then, block-iterative EM methods have been very popular in the medical imaging community for tomographic image reconstruction (Ahn and Fessler, 2003; Erdogan and Fessler, 1999; H. Ing-Tsung, 2002; Kole and Beekman, 2005) due to their remarkably fast convergence rates. Incremental EM methods are considerably efficient for solving problems with large datasets including clustering and segmenting audio signals (A. Bietti and Cont, 2015), learning and interpreting data acquired from sensors on planetary robots (X. R. Wang and Upcroft, 2005), identifying human action on videos (J. Xu, 2009) or analysing energy consumption in hospitals (S. Ng, 2005). Convergence properties of the incremental variant of the EM algorithm have been provided in (A. Gunawardana, 2005) using Zangwill convergence theorem. Yet, those properties are given in the context of discrete variables and when the sampling scheme is the same from a pass over the data to another. In this article, we are presenting a mini-batch variant of the EM algorithm and providing its convergence properties using the MISO (Minimisation by Incremental Surrogate Optimisation) framework introduced in (Mairal, 2015). When the quantity computed at the E-step involves infeasible computations, new methods have been developed in order to by-pass the issue. The Monte Carlo EM algorithm (Wei and Tanner, 1990) has been proposed in the context of mixture problem and involves splitting the E-step in a first step where the latent variables are simulated and then a Monte Carlo integration of the expectation of the complete log likelihood. The MCEM algorithm has been successfully applied in non linear mixed effects model of plant growth (C. Baey and Cournede, 2016) or to do inference for joint modelling of time to event data coming from clinical trials in (Chakraborty and Das, 2010). This algorithm has been vastly studied in (Levine and Casella, 2001) or (McLachlan and Krishnan, 2007) and its convergence properties have been derived, initially, in (Biscarat, 1994; Chan and Ledolter, 1995) and more recently in (Neath, 2012) and (Fort and Moulines, 2003). In this contribution, we adapt the MISO framework to the stochastic version of the mini-batch EM algorithm in order to prove its convergence.

The paper is composed of two main parts corresponding to the convergence properties of the mini-batch EM (MBEM) and the mini-batch MCEM (MBMCEM). Each section provides the executed algorithm and the convergence theorem. Finally, we investigate, through a simulation study, how fast these algorithms are.

2 Convergence of the mini-batch EM algorithm

2.1 Model assumptions and notations

M 1. *The parameter set Θ is a closed convex subset of $\mathbb{R}^p$.*

Let N be an integer and for $i \in [\![1, N]\!]$, Z_i be a subset of $\mathbb{R}^{m_i}$, μ_i be a σ-finite measure on the Borel σ-algebra $\mathcal{Z}_i = \mathcal{B}(\mathsf{Z}_i)$ and $\{f_i(z_i, \theta), \theta \in \Theta\}$ be a family of positive μ_i-integrable Borel functions on Z_i. Set $z = (z_i \in \mathsf{Z}_i, 1 \leq i \leq N) \in \mathsf{Z}$ where $\mathsf{Z} = \bigtimes_{n=1}^{N} \mathsf{Z}_i$ and μ the product of the measures $(\mu_i, 1 \leq i \leq N)$.
Define, for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$:

$$\begin{aligned} g_i(\theta) &\triangleq \int_{\mathsf{Z}_i} f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \\ p_i(z_i, \theta) &\triangleq \begin{cases} \frac{f_i(z_i,\theta)}{g_i(\theta)} & \text{if } g_i(\theta) \neq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

Note that $p_i(z_i, \theta)$ defines a probability density function with respect to μ_i. Thus $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ is a family of probability density. We denote by $\{\mathbb{P}_{i,\theta}; \theta \in \Theta\}$ the associated family of probability measures. For all $\theta \in \Theta$, we set

$$\begin{aligned} f(z, \theta) &= \prod_{i=1}^{N} f_i(z_i, \theta) \\ g(\theta) &= \prod_{i=1}^{N} g_i(\theta) \\ p(z, \theta) &= \prod_{i=1}^{N} p_i(z_i, \theta) \end{aligned} \tag{3}$$

Remark 1. *An example of this setting is the incomplete data framework. In this case, we consider N independent observations $(y_i \in \mathsf{Y}_i, 1 \leq i \leq N)$ where Y_i is a subset of $\mathbb{R}^{\ell_i}$ and missing data $(z_i \in \mathsf{Z}_i, 1 \leq i \leq N)$. In this framework, we get*

- *$f_i(z_i, \theta)$ is the complete data likelihood that is the likelihood of the observed data y_i augmented with the missing data z_i.*

- *$g_i(\theta)$ is the incomplete data likelihood that is the likelihood of the observed data y_i.*

- *$p_i(z_i, \theta)$ is the posterior distribution of the missing data z_i given the observed data y_i.*

Our objective is to maximise the function $\theta \to \log g(\theta)$ or equivalently to minimize the objective function $\ell : \Theta \mapsto \mathbb{R}$ defined as:

$$\ell(\theta) \triangleq -\log g(\theta) = \sum_{i=1}^{N} \ell_i(\theta) \tag{4}$$

where $\ell_i(\theta) \triangleq -\log g_i(\theta)$. The EM algorithm is an iterative optimisation algorithm that minimizes the function $\theta \to \ell(\theta)$ when its direct minimisation is difficult. Denote by θ^{k-1} the current fit of the parameter at iteration k. The k-th step of the EM algorithm might be decomposed into two steps. The E-step consists in computing the surrogate function defined for all $\theta \in \Theta$ as :

$$\begin{aligned} Q(\theta, \theta^{k-1}) &\triangleq -\int_{\mathsf{Z}} p(z, \theta^{k-1}) \log f(z, \theta) \mu(\mathrm{d}z) \\ &= -\sum_{i=1}^{N} \int_{\mathsf{Z}_i} p_i(z_i, \theta^{k-1}) \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) = \sum_{i=1}^{N} Q_i(\theta, \theta^{k-1}) \end{aligned} \tag{5}$$

where:

$$Q_i(\theta, \theta^{k-1}) \triangleq -\int_{\mathsf{Z}_i} p_i(z_i, \theta^{k-1}) \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \tag{6}$$

In the M-step, the value of θ minimizing $Q(\theta, \theta^{k-1})$ is calculated. This yields the new parameter estimate θ^k. These two steps are repeated until convergence. The essence of the EM algorithm is that decreasing $Q(\theta, \theta^{k-1})$ forces a decrease of the function $\theta \to \ell(\theta)$ (Dempster and Rubin, 1977). The mini-batch version of the EM algorithm is described as follows:

```
Algorithm 1 mini-batch EM algorithm
Initialization: given an initial parameter estimate $\theta^0$, for all $i \in [\![1, N]\!]$ compute a surrogate function $\vartheta \to R_i^0(\vartheta) = Q_i(\vartheta, \theta^0)$ defined by (6).
Iteration k: given the current estimate $\theta^{k-1}$:
    1. Pick a set $I_k$ uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
    2. For all $i \in I_k$, compute $\vartheta \to Q_i(\vartheta, \theta^{k-1})$ defined by (6).
    3. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{N} R_i^k(\vartheta)$ where $R_i^k(\vartheta)$ are defined recursively as follows:
           $R_i^k(\vartheta) = \begin{cases} Q_i(\vartheta, \theta^{k-1}) & \text{if } i \in I_k \\ R_i^{k-1}(\vartheta) & \text{otherwise} \end{cases}$   (7)
```

We remark that, for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$:

$$R_i^k(\theta) = Q_i(\theta, \theta^{\tau_{i,k}}) \tag{8}$$

where for all $i \in [\![1, N]\!]$, $\tau_{i,0} = 0$ and $k \geq 1$ the indices $\tau_{i,k}$ are defined recursively as follows:

$$\tau_{i,k} = \begin{cases} k-1 & \text{if } i \in I_k \\ \tau_{i,k-1} & \text{otherwise} \end{cases} \tag{9}$$

As noted in (A. Gunawardana, 2005) and (R. Neal, 2007), there is no guarantee, unlike the EM algorithm, that the objective function $\theta \to \ell(\theta)$ decreases at each iteration. We also remark that we recover the full EM algorithm when the mini-batch size p is set to be equal to N. Let $\mathcal{T}(\Theta)$ be a neighborhood of Θ. To study the convergence of the MBEM algorithm we consider the following assumptions:

M 2. *For all $i \in [\![1, N]\!]$, assume that:*

a. *For all $\theta \in \Theta$ and $z_i \in \mathsf{Z}_i$, $f_i(z_i, \theta)$ is strictly positive, the function $\theta \to f_i(z_i, \theta)$ is two-times differentiable on $\mathcal{T}(\Theta)$ for μ_i almost every z_i and for all $\vartheta \in \Theta$:*

$$\int_{\mathsf{Z}_i} |\nabla f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty \quad \textit{and} \quad \int_{\mathsf{Z}_i} p_i(z_i, \vartheta) |\log f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty \tag{10}$$

b. *For all $\theta \in \Theta$, there exist $\delta > 0$ and a measurable function $\psi_\theta : \mathsf{Z}_i \to \mathbb{R}$ such that $\sup_{\|\vartheta - \theta\| \leq \delta} |\nabla^2 f_i(z_i, \vartheta)| \leq \psi_\theta(z_i)$ for μ_i almost every z_i with $\int_{\mathsf{Z}_i} \psi_\theta(z_i) \mu_i(\mathrm{d}z_i) < \infty$.*

c. *There exist a measurable function $\phi_i : \mathsf{Z}_i \to \mathbb{R}$ and $L_i < \infty$ such that $\sup_{\theta \in \Theta} |\nabla^2 \log f_i(z_i, \theta)| \leq \phi_i(z_i)$ for μ_i almost every z_i with $\sup_{\theta \in \Theta} \int_{\mathsf{Z}_i} p_i(z_i, \theta) \phi_i(z_i) \mu_i(\mathrm{d}z_i) \leq L_i$.*

d. *For all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, $\sup_{\theta \in \Theta} |\nabla^2 l_i(\theta)| < \infty$.*

It is easily checked that these assumptions imply for all $i \in [\![1, N]\!]$ that:

1. The function $\theta \to g_i(\theta)$ is continuously differentiable on $\mathcal{T}(\Theta)$ and the Fisher identity (Fisher, 1925) holds:

$$\nabla \ell_i(\theta) = -\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \tag{11}$$

2. For all $\vartheta \in \Theta$, the function $\theta \to Q_i(\theta, \vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$ and is L_i−smooth, i.e., for all $(\theta, \theta') \in \Theta$ and $L_i > 0$:

$$\|\nabla Q_i(\theta, \vartheta) - \nabla Q_i(\theta', \vartheta)\| \leq L_i \|\theta - \theta'\| \tag{12}$$

3. For all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, Louis Formula (Louis, 1982) yields that:

$$\begin{aligned}\nabla^2 l_i(\theta) = &- \int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla^2 \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \\ &- \int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \\ &+ \left(\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \right)^\top \int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i)\end{aligned} \tag{13}$$

Thus, sufficient conditions to verify M 2d. are M 2c. and the following condition: There exist a measurable function $N_i : \mathsf{Z}_i \to \mathbb{R}$ such that for all $\theta \in \Theta$, $|\nabla \log f_i(z_i, \theta)| \leq N_i(z_i)$ for μ_i almost every z_i with $\int_{\mathsf{Z}_i} p_i(z_i, \theta) N_i^2(z_i) \mu_i(\mathrm{d}z_i) < \infty$.

M 3. *For all $i \in [\![1, N]\!]$, the objective function ℓ_i is bounded from below, i.e. there exist $M_i \in \mathbb{R}$ such that for all $\theta \in \Theta$:*

$$\ell_i(\theta) \geq M_i \tag{14}$$

For $\theta \in \Theta$, we say that a function $B_{i,\theta}$ is a surrogate of ℓ_i at θ if the following three properties are satisfied:

S.1 the function $\vartheta \to B_{i,\theta}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$

S.2 for all $\vartheta \in \Theta$, $B_{i,\theta}(\vartheta) \geq \ell_i(\vartheta)$

S.3 $B_{i,\theta}(\theta) = \ell_i(\theta)$ and $\nabla B_{i,\theta}(\vartheta)\Big|_{\vartheta=\theta} = \nabla \ell_i(\vartheta)\Big|_{\vartheta=\theta}$.

For all $i \in [\![1, N]\!]$ and $(\theta, \theta') \in \Theta^2$, define the Kullback-Leibler Divergence from $\mathbb{P}_{i,\theta'}$ to $\mathbb{P}_{i,\theta}$ as:

$$\mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\theta'}) \triangleq \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{p_i(z_i, \theta')} \mu_i(\mathrm{d}z_i) \tag{15}$$

and the negated entropy function $H_i(\theta)$ as:

$$H_i(\theta) \triangleq \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \tag{16}$$

To analyze the MBEM algorithm, we introduce for $i \in [\![1, N]\!]$ and $\theta \in \Theta$ the function $\vartheta \to B_{i,\theta}(\vartheta)$ defined by:

$$B_{i,\theta}(\vartheta) \triangleq Q_i(\vartheta, \theta) + H_i(\theta) \tag{17}$$

We will show below that for $i \in [\![1, N]\!]$ and $\theta \in \Theta$, $B_{i,\theta}$ is a surrogate of l_i at θ. Let us note that this function can be rewritten as follows:

$$\begin{aligned}B_{i,\theta}(\vartheta) &= \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{f_i(z_i, \vartheta)} \mu_i(\mathrm{d}z_i) \\ &= \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{p_i(z_i, \vartheta)} \mu_i(\mathrm{d}z_i) + \ell_i(\vartheta) \\ &= \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta}) + \ell_i(\vartheta)\end{aligned} \tag{18}$$

We verify **S.1** using assumption M 2. Since $\vartheta \to \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta})$ is always positive and is equal to zero if $\theta = \vartheta$, we verify **S.2** and the first part of **S.3**. The second part of **S.3** follows from the Fisher identity (11). The difference between the surrogate function and the objective function denoted, for all $\vartheta \in \Theta$, $h_i(\vartheta) \triangleq B_{i,\theta}(\vartheta) - l_i(\vartheta)$ plays a key role in the convergence analysis. Here, for all $i \in [\![1, N]\!]$ and $\vartheta \in \Theta$ the error reads $h_i(\vartheta) = \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta})$. Under M 2c. and M 2d., we obtain that for all $i \in [\![1, N]\!]$, the function $\vartheta \to h_i(\vartheta)$ is L_i-smooth. Since for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, the surrogate function $\vartheta \to B_{i,\theta}(\vartheta)$ is equal to $\vartheta \to Q_i(\vartheta, \theta)$ up to a constant, the MBEM algorithm is equivalent to the following theoretical algorithm:

```
Algorithm 2 Theoretical MBEM algorithm
Initialization: given an initial parameter estimate $\theta^0$, for all $i \in [\![1, N]\!]$
compute a surrogate function $\vartheta \to A_i^0(\vartheta) = B_{i,\theta^0}(\vartheta)$ defined by (18).
Iteration k: given the current estimate $\theta^{k-1}$:
    1. Pick a set $I_k$ uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
    2. For all $i \in I_k$, compute a surrogate function $\vartheta \to B_{i,\theta^{k-1}}(\vartheta)$ defined
       by (18).
    3. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N A_i^k(\vartheta)$ where $A_i^k(\vartheta)$ are defined recursively
       as follows:
           $A_i^k(\vartheta) = \begin{cases} B_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ A_i^{k-1}(\vartheta) & \text{otherwise} \end{cases}$    (19)
```

We remark that, for all $i \in [\![1, N]\!]$ and $\vartheta \in \Theta$:

$$A_i^k(\vartheta) = B_{i,\theta^{\tau_{i,k}}}(\vartheta) \tag{20}$$

using the notation introduced in (9). Denote by $\langle .,. \rangle$ the scalar product. We now state the convergence theorem of the MBEM algorithm:

Theorem 1. *Assume* **M1-M3***. Let* $\left(\theta^k\right)_{k \geq 1}$ *be a sequence generated from* $\theta^0 \in \Theta$ *by the iterative application described by algorithm 1. Then:*

(i) $\left(\ell(\theta^k)\right)_{k \geq 1}$ *converges almost surely.*

(ii) $\left(\theta^k\right)_{k \geq 1}$ *satisfies the Asymptotic Stationary Point Condition, i.e.*

$$\liminf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq 0 \tag{21}$$

Proof. The proof is postponed to section 5.1 □

We observe that in the unconstrained case, we have:

$$\inf_{\theta \in \mathbb{R}^d} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} = -\|\nabla \ell(\theta^k)\| \tag{22}$$

which yields to $\lim_{k \to \infty} \|\nabla \ell(\theta^k)\| = 0$.

2.2 MBEM for a curved exponential family

In the particular case where for all $i \in [\![1, N]\!]$ and $z_i \in \mathsf{Z}_i$, the function $\theta \to f_i(z_i, \theta)$ belongs to the curved exponential family, we assume that:

E 1. *For all $i \in [\![1, N]\!]$ and $\theta \in \Theta$:*

$$\log f_i(z_i, \theta) = H_i(z_i) - \psi_i(\theta) + \langle \tilde{S}_i(z_i), \phi_i(\theta) \rangle. \tag{23}$$

where $\psi_i : \Theta \mapsto \mathbb{R}$ and $\phi_i : \Theta \mapsto \mathbb{R}$ are twice continuously differentiable functions of θ, $H_i : \mathsf{Z}_i \mapsto \mathbb{R}$ is a twice continuously differentiable function of z_i and $\tilde{S}_i : \mathsf{Z}_i \mapsto \mathsf{S}_i$ is a statistic taking its values in a convex subset S_i of $\mathbb{R}$ and such that $\int_{\mathsf{Z}_i} |\tilde{S}_i(z_i)| p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) < \infty$.

Define for all $\theta \in \Theta$ and $i \in [\![1, N]\!]$ the function $\bar{s}_i : \Theta \to \mathsf{S}_i$ as:

$$\bar{s}_i(\theta) \triangleq \int_{\mathsf{Z}_i} \tilde{S}_i(z_i) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i). \tag{24}$$

Define, for all $\theta \in \Theta$ and $s = (s_i, 1 \leq i \leq N) \in \mathsf{S}$ where $\mathsf{S} = \bigtimes_{n=1}^{N} \mathsf{S}_i$, the function $L(s; \theta)$ by:

$$L(s; \theta) \triangleq \sum_{i=1}^{N} \psi_i(\theta) - \sum_{i=1}^{N} \langle s_i, \phi_i(\theta) \rangle. \tag{25}$$

E 2. *There exist a function $\hat{\theta} : \mathsf{S} \mapsto \Theta$ such that for all $\bar{s} \in \mathsf{S}$, :*

$$L(s; \hat{\theta}(s)) \leq L(s; \theta). \tag{26}$$

In many models of practical interest for all $s \in \mathsf{S}$, $\theta \mapsto L(s, \theta)$ has a unique minimum. In the context of the curved exponential family, the MBEM algorithm can be formulated as follows:

```
Algorithm 3 mini-batch EM for a curved exponential family
Initialisation: given an initial parameter estimate $\theta^0$, for all $i \in [\![1, N]\!]$ compute $s_i^0 = \bar{s}(\theta^0)$.
Iteration k: given the current estimate $\theta^{k-1}$:
    1. Pick a set $I_k$ uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
    2. For $i \in [\![1, N]\!]$, compute $s_i^k$ such as:
        $s_i^k = \begin{cases} \bar{s}_i(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases}$    (27)
    3. Set $\theta^k = \hat{\theta}(s^k)$ where $s^k = (s_i^k, 1 \leq i \leq N)$.
```

Example 1. We observe N independent and identically distributed (i.i.d.) random variables $(y_i, 1 \leqslant i \leqslant N)$. Each one of these observations is distributed according to a mixture model. Denote by $(c^j, 1 \leqslant j \leqslant J)$ the distribution of the component of the mixture and $(\pi_j, 1 \leqslant j \leqslant J)$ the associated weights. Consider the complete data likelihood for each individual $f_i(z_i, \theta)$:

$$f_i(z_i, \theta) = \prod_{j=1}^{J} (\pi_j c^j(y_i, \delta))^{\mathbb{1}_{z_i = j}} \tag{28}$$

We restrict this study to a mixture of Gaussian distributions. In such case $\theta = ((\pi_j, \mu_j, \sigma_j), 1 \leqslant j \leqslant J)$ and the individual complete log likelihood is expressed as:

$$\log f_i(z_i, \theta) = \sum_{j=1}^{J} \mathbb{1}_{z_i = j} \log(\pi_j) + \sum_{j=1}^{J} \mathbb{1}_{z_i = j} \left[-\frac{(y_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2} \log \sigma_j^2 \right] \tag{29}$$

The complete data sufficient statistics are given for all $i \in [\![1, N]\!]$ and $j \in [\![1, J]\!]$, by $\tilde{S}_i^{1,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i = j}$, $\tilde{S}_i^{2,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i = j} y_i$ and $\tilde{S}_i^{3,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i = j} y_i^2$. At each iteration k, algorithm 2.2 consists in picking a set I_k and for $i \in I_k$, computing the following quantities:

$$\begin{aligned} (\bar{s}_i^k)^{1,j} &= \int_{\mathsf{Z}_i} \mathbb{1}_{z_i = j} p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) \\ (\bar{s}_i^k)^{2,j} &= \int_{\mathsf{Z}_i} \mathbb{1}_{z_i = j} y_i p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i \\ (\bar{s}_i^k)^{3,j} &= \int_{\mathsf{Z}_i} \mathbb{1}_{z_i = j} y_i^2 p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i^2 \end{aligned} \tag{30}$$

where the quantity $p_{ij}(\theta^{k-1}) \triangleq \mathbb{P}_{i, \theta^{k-1}}(z_i = j)$ is obtained using the Bayes rule:

$$p_{ij}(\theta^{k-1}) = \frac{\mathbb{P}_i(z_i = j) p_i(y_i | z_i = j; \theta^{k-1})}{p_i(y_i; \theta^{k-1})} = \frac{\pi_j^{k-1} c^j(y_i; \mu_j^{k-1}, \sigma_j^{k-1})}{\sum_{l=1}^{J} \pi_l^{k-1} c^l(y_i; \mu_l^{k-1}, \sigma_l^{k-1})} \tag{31}$$

Finally the maximisation step yields:

$$\begin{aligned}
\pi_j^k &= \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}}{N} \\
\mu_j^k &= \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{2,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} \\
\sigma_j^k &= \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{3,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} - (\mu_j^k)^2
\end{aligned} \tag{32}$$

3 Convergence of the mini-batch MCEM algorithm

We now consider the stochastic version of the MBEM algorithm called the mini-batch MCEM algorithm. At iteration k, the MBMCEM approximates the quantity defined by (6) by Monte Carlo integration, i.e. for all $i \in I_k$, $\vartheta \in \Theta$ and $k \geq 1$:

$$\hat{Q}_i^k(\vartheta, \theta^{k-1}) \triangleq \frac{1}{M_k} \sum_{m=0}^{M_k-1} \log f_i(z_i^{k,m}, \vartheta) \tag{33}$$

where $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ is a Monte Carlo batch. In simple scenarios, the samples $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ are conditionally independent and identically distributed with distribution $p_i(z_i, \theta^{k-1})$. Nevertheless, in most cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov Chains (MCMC) algorithm. MCMC algorithms are a class of methods allowing to sample from complex distribution over (possibly) large dimensional space.

Recall that a Markov kernel P on a measurable space $(\mathsf{E}, \mathcal{E})$ is an application on $\mathsf{E} \times \mathcal{E}$, taking values in $[0,1]$ such that for any $z \in \mathsf{E}$, $P(z, \cdot)$ is a probability measure on $\mathcal{E}$ and for any $A \in \mathcal{E}$, $P(\cdot, A)$ is measurable. We denote by P^k the $k-$th iterate of P defined recursively as $P^0(z, A) \triangleq \mathbb{1}_A(z)$ and for $k \geq 1$, $P^k(z, A) \triangleq \int_A P^{k-1}(z, \mathrm{d}z')P(z', A)$. The probability π is said to be stationary for P if $\int_{\mathsf{E}} \pi(\mathrm{d}z)P(z, A) = \pi(A)$ for any $A \in \mathcal{E}$. We refer the reader to (Meyn and Tweedie, 2009) for the definitions of basic properties of Markov chains.

For $i \in [\![1, N]\!]$ and $\theta \in \Theta$, let $P_{i,\theta}$ be a Markov kernel with stationary distribution $\pi_{i,\theta}(A_i) = \int_{A_i} p_i(z_i, \theta)\mu_i(\mathrm{d}z_i)$ where $A_i \in \mathcal{Z}_i$. For example, $P_{i,\theta}$ might be either a Gibbs or a Metropolis-Hastings samplers with target distribution $\pi_{i,\theta}$. For $\theta \in \Theta$, let $\lambda_{i,\theta}$ be a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$. We will use $\lambda_{i,\theta}$ as an initial distribution and allow this initial distribution to depend on the parameter θ. For example, $\lambda_{i,\theta}$ might be the Dirac mass at some given point but more clever choice can be made. We denote by $\mathbb{E}_{i,\theta}$ the expectation of the canonical Markov chain $\{z_i^m\}_{m=0}^{\infty}$ with initial distribution $\lambda_{i,\theta}$ and transition kernel $P_{i,\theta}$.

In this setting, the Monte Carlo mini-batch $\{z_i^{k,m}\}_{m=0}^{M_{k-1}}$ is a realisation of a Markov Chain with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$. The MBMCEM algorithm can be summarised as follows:

Algorithm 4 mini-batch MCEM algorithm

```
Initialization: given an initial parameter estimate $\theta^0$, for all $i \in [\![1, N]\!]$ and $m \in [\![0, M_0 - 1]\!]$, sample a Markov Chain $\{z_i^{0,m}\}_{m=0}^{M_0-1}$ with initial distribution $\lambda_{i,\theta^0}$ and transition kernel $P_{i,\theta^0}$ and compute a function $\vartheta \to \hat{R}_i^0(\vartheta) = \hat{Q}_i^0(\vartheta, \theta^0)$ defined by (33).
Iteration k: given the current estimate $\theta^{k-1}$:
    1. Pick a set $I_k$ uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
    2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sample a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$.
    3. For all $i \in I_k$, compute the function $\vartheta \to \hat{Q}_i^k(\vartheta, \theta^{k-1})$ defined by (33).
    4. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N \hat{R}_i^k(\vartheta)$ where $\hat{R}_i^k(\vartheta)$ are defined recursively as follows:
        $\hat{R}_i^k(\vartheta) = \begin{cases} \hat{Q}_i^k(\vartheta, \theta^{k-1}) & \text{if } i \in I_k \\ \hat{R}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases}$    (34)
```

Whether we use Markov Chain Monte Carlo or direct simulation, we need to control the supremum norm of the fluctuations of the Monte Carlo approximation. Let $i \in [\![1, N]\!]$, $\{q_i(z_i, \vartheta), z_i \in \mathsf{Z}_i, \vartheta \in \Theta\}$ be a family of measurable functions, λ_i a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$. We define:

$$\begin{aligned} C_i(q_i) \triangleq \sup_{\theta \in \Theta} \sup_{M > 0} M^{-p/2} \mathbb{E}_{i,\theta} \Bigg[\sup_{\vartheta \in \Theta} \Bigg| \sum_{m=0}^{M-1} & \{ q_i(z_i^m, \vartheta) \\ & - \int_{\mathsf{Z}_i} q_i(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \Big\} \Bigg| \Bigg] \end{aligned} \tag{35}$$

M 4. *For all $i \in [\![1, N]\!]$:*

$$C_i(\log f_i) < \infty \quad \textit{and} \quad C_i(\nabla \log f_i) < \infty \tag{36}$$

The assumption M 4 is based on maximal inequality for beta-mixing sequences obtained in (Doukhan et al., 1995). This condition can be translated in terms of drift and minorization conditions (see (Meyn and Tweedie, 2009)). Finally, we consider the following assumption on the number of simulations:

M 5. *$\{M_k\}_{k \geq 0}$ is a non deacreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$.*

We now state the convergence theorem for the MBMCEM algorithm:

Theorem 2. *Assume* **M1-M5***. Let* $\left(\theta^k\right)_{k\geq 1}$ *be a sequence generated from* $\theta^0 \in \Theta$ *by the iterative application described by algorithm 4. Then:*

(i) $\left(\ell(\theta^k)\right)_{k\geq 1}$ *converges almost surely.*

(ii) $\left(\theta^k\right)_{k\geq 1}$ *satisfies the Asymptotic Stationary Point Condition.*

Proof. The proof is postponed to section 5.2 □

3.1 MBMCEM for a curved exponential family

Using the notations introduced in section 2.2, we can write the mini-batch MCEM algorithm can be described as follows:

Algorithm 5 mini-batch MCEM for a curved exponential family

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![1, N]\!]$ and $m \in [\![0, M_0 - 1]\!]$, sample a Markov Chain $\{z_i^{0,m}\}_{m=0}^{M_0-1}$ with initial distribution λ_{i,θ^0} and transition kernel P_{i,θ^0} and compute $s_i^0 = \frac{1}{M_0}\sum_{m=1}^{M_0} \tilde{S}_i(z_i^{0,m})$.
Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$

2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sample a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$.

3. Compute s_i^k such as:

$$s_i^k = \begin{cases} \frac{1}{M_k}\sum_{m=1}^{M_k-1} \tilde{S}_i(z_i^{k,m}) & \text{if } i \in I_k \\ s_i^{k-1} & \text{otherwise} \end{cases} \tag{37}$$

4. Set $\theta^k = \hat{\theta}(\bar{s}^k)$ where $s^k = (s_i^k, 1 \leq i \leq N)$

4 Numerical examples

4.1 A Linear mixed effects model

4.1.1 The model

We consider a linear mixed effects model (Verbeke and Molenberghs, 2000; B.T. West and Galecki, 2006). We denote by $y = (y_i, 1 \leq i \leq N)$ the observations, for each $i \in [\![1, N]\!]$, y_i is a $n_i \times 1$ vector where for all $i \in [\![1, N]\!]$:

$$y_i = A_i\theta + B_i z_i + \epsilon_i. \tag{38}$$

where $A_i \in \mathbb{R}^{n_i \times d_A}$ and $B_i \in \mathbb{R}^{n_i \times d_B}$ are design matrices, $\theta \in \mathbb{R}^{d_A}$ is a vector of parameters, $z_i \in \mathbb{R}^{d_B}$ are the latent data (i.e. the random effects in the context of mixed effects models) which are assumed to be distributed according to a normal distribution $\mathcal{N}(0, \Omega)$. We also assume that the residual errors $\epsilon_i \in \mathbb{R}^{n_i}$ are distributed according to $\mathcal{N}(0, \Sigma)$ and that the sequences of variables $(z_i, 1 \leq i \leq N)$ and $(\epsilon_i, 1 \leq i \leq N)$ are i.i.d. and mutually independent. The covariance matrices Ω and Σ are assumed to be known. For all $i \in [\![1, N]\!]$, the conditional distribution of the observations given the latent variables $y_i|z_i$ and of the latent variables given the observations $z_i|y_i$ are respectively given by:

$$\begin{aligned} y_i|z_i &\sim \mathcal{N}(A_i\theta + B_i z_i, \Sigma), \\ z_i|y_i &\sim \mathcal{N}(\mu_i, \Gamma_i). \end{aligned} \tag{39}$$

where:

$$\begin{aligned} \Gamma_i &= (B_i^\top \Sigma^{-1} B_i + \Omega^{-1})^{-1}, \\ \mu_i &= \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta). \end{aligned} \tag{40}$$

This model belongs to the curved exponential family introduced in section 2.2 where for all $i \in [\![1, N]\!]$:

$$\begin{aligned} &\tilde{S}_i(z_i) \triangleq z_i \quad \text{and} \quad \bar{s}_i(\theta) = \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta) \\ &\psi_i(\theta) \triangleq (y_i - A_i\theta)^\top \Sigma^{-1}(y_i - A_i\theta) \\ &\phi_i(\theta) \triangleq B^\top \Sigma^{-1}(y_i - A_i\theta) \end{aligned} \tag{41}$$

Maximising $L(s, \theta)$, defined in (25), with respect to θ yields the following maximisation function for all $s = (s_i \in \mathbb{R}^{d_B}, 1 \leq i \leq N)$:

$$\hat{\theta}(s) \triangleq \left(\sum_{i=1}^{N} A_i^\top \Sigma^{-1} A_i\right)^{-1} \sum_{i=1}^{N} A_i^\top \Sigma^{-1}(y_i - B_i s_i).$$

Thus, the $k - th$ update of the MBEM algorithm consists in sampling a subset of indices I_k and computing $\theta^k = \hat{\theta}(s^k)$ where:

$$s_i^k = \begin{cases} \bar{s}_i(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases}$$

4.1.2 Simulation and runs

We generate a synthetic dataset, with $d_A = 2$, $\theta = (\theta_1 : 4, \theta_2 : 9)$, $N = 1000$ and for all $i \in \llbracket 1, N \rrbracket$, $n_i = 10$ observations per individual. We also generate random design matrices $(A_i, 1 \leq i \leq N)$ and $(B_i, 1 \leq i \leq N)$. Two runs of the MBEM are executed starting from different initial values $((\theta_1^0 : 1, \theta_2^0 : 5)$ and $(\theta_1^0 : 3, \theta_2^0 : 7))$ to study the convergence behaviour of these algorithms depending on how far from the true solution they start. Figure 1 shows the convergence of the vector of parameter estimates $(\theta_1^k, \theta_2^k)_{k=0}^K$ over passes of the EM algorithm, the MBEM algorithm where half of the data is considered at each iteration and the Incremental EM algorithm (i.e. the mini-batch is reduced to a single data point at each iteration). The speed of convergence is a monotone function of the batch size in this case, the smaller the batch the faster the convergence.

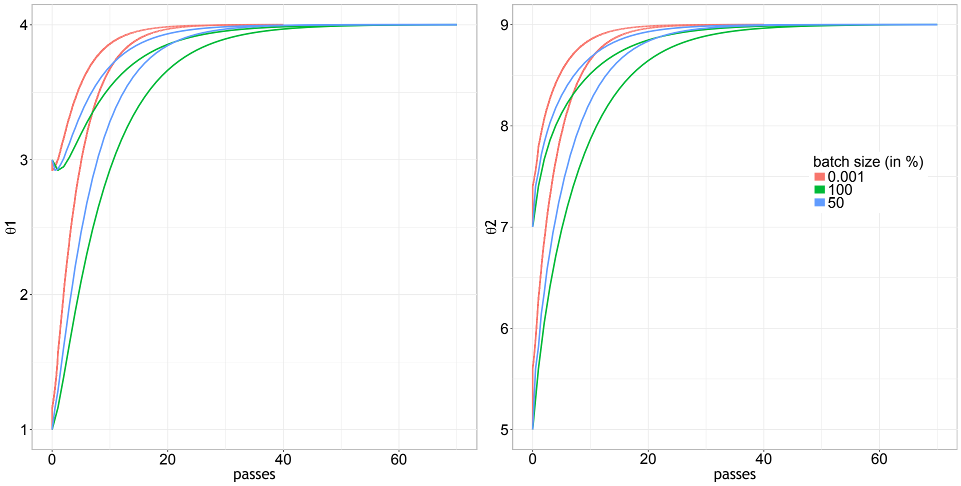

Figure 1: Convergence of the vector of parameter estimates θ^k function of passes over the data.

4.2 Logistic regression for a binary variable

4.2.1 The model

Let $y = (y_i, 1 \leq i \leq N)$ be the vector of binary responses where for each individual i, $y_i = (y_{ij}, 1 \leq j \leq n_i)$ is a sequence of conditionally independent random variables taking values in $\{0, 1\}$ and correspond to the j-th responses for the i-th subject. Let $z = (z_i, 1 \leq i \leq N)$ be a vector of latent data where $z_i = (z_{i,1}, z_{i,2}, z_{i,3}) \in \mathbb{R}^3$. Let $((c_{ij,1}, c_{ij,2}), 1 \leq j \leq n_i, 1 \leq i \leq n)$ be explanatory variables. For instance, responses are 'healthy' and 'sick' of a patient i and the predictors represent the time and the dosage of a certain injected drug. Formally, the conditional distribution of the observations y_i given the latent variables z_i is given by:

$$\operatorname{logit}(\mathbb{P}(y_{ij} = 0 | z_i)) = z_{i,1} + z_{i,2} c_{ij,1} + z_{i,3} c_{ij,2} = c_{ij}^\top z_i \tag{42}$$

where $c_{ij} = (1, c_{ij,1}, c_{ij,2})$. This is a logistic regression but the parameters of this regression which are the latent variables depend upon each individual. The distribution of the vector of latent variables is assumed to be Gaussian for every individual of the population, $z_i \sim \mathcal{N}(\beta, \Omega)$ with $\beta = (\beta_1, \beta_2, \beta_3)$ and $\Omega = \operatorname{diag}(\omega_1^2, \omega_2^2, \omega_3^2)$. The complete log likelihood is expressed as:

$$\log f(z, \theta) \propto \sum_{i=1}^{N} \left\{ y_i c_i^\top z_i + \log\left(\frac{1}{1 + \exp(c_i^\top z_i)}\right) - \frac{1}{2}\log(|\Omega|) - \frac{1}{2}\operatorname{Tr}(\Omega^{-1}(z_i - \beta)(z_i - \beta)^\top) \right\}. \tag{43}$$

Since the expectation of the complete log likelihood with respect to the conditional distribution of the latent variables given the observations is intractable, we use the MCEM and the MBMCEM algorithms, which require to simulate random draws from this conditional distribution. We use the saemix R package (E. Comets and Lavielle, 2017) to run a Metropolis-Hastings within Gibbs sampler (Brooks et al., 2011) where for all $i \in [\![1, N]\!]$ and dimension $d \in [\![1, 3]\!]$ of the parameter, the Markov Chain is constructed as follows:

```
Algorithm 6 Random Walk Metropolis
Initialization: given the current parameter estimates $(\beta_d^{k-1}, \omega_d^{k-1})$ set $\mathcal{T}_d^{(0)} = \omega_d^{k-1}$ and sample the initial state $z_{i,d}^{(0)} \sim \mathcal{N}(\beta_d^{k-1}, \mathcal{T}_d^{(0)})$
Iteration t: given the current chain state $z_d^{(t-1)}$:
    1. Sample a candidate state:
           $z_{i,d}^{(c)} \sim \mathcal{N}(z_{i,d}^{(t-1)}, \mathcal{T}_d^{(t-1)})$        (44)
    2. Accept with probability $\min\left(1, \alpha^{(t)}(z_{i,d}^{(c)}, z_{i,d}^{(t-1)})\right)$
    3. Update the variance of the proposal as follows:
           $\mathcal{T}_d^{(t)} = \mathcal{T}_d^{(t-1)} + (1 + \delta(\alpha^{(t)} - \alpha^*))$        (45)
```

where $\delta = 0.4$, α^t is the MH acceptance ratio and $\alpha^* = 0.4$ is the optimal acceptance ratio (Robert and Casella, 2005). This model belongs to the curved exponential family introduced in section 2.2 where for all $i \in [\![1, N]\!]$, $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$. At iteration k, the MBMCEM algorithm consists in:

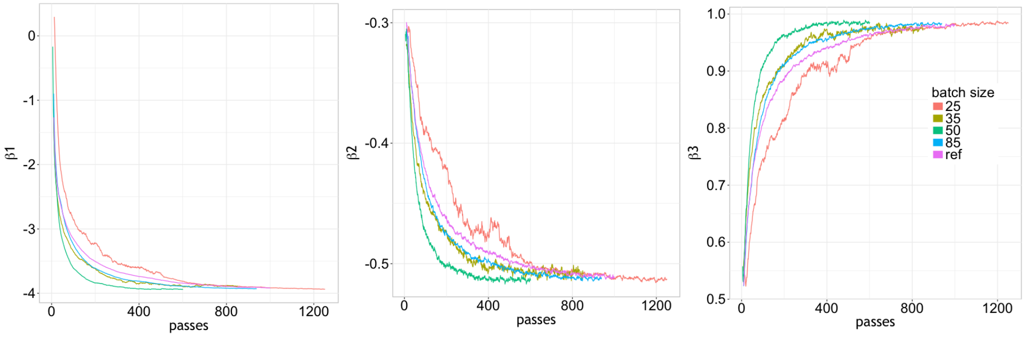

Figure 2: Convergence of the vector of fixed parameters β for different batch sizes function of passes over the data.

1. Picking a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sampling a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ using algorithm 6.
3. Computing $s_i^k = (s_i^{1,k}, s_i^{2,k})$ such as:

$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k}\sum_{m=0}^{M_k-1} z_i^{k,m}, \frac{1}{M_k}\sum_{m=0}^{M_k-1} (z_i^{k,m})^\top z_i^{k,m}\right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases} \tag{46}$$

4. Updating the parameters as follows:

$$\begin{aligned} \beta^k &= \frac{1}{N}\sum_{i=1}^{N} s_i^{1,k} \\ \Omega^k &= \frac{1}{N}\sum_{i=1}^{N} s_i^{2,k} - (\beta^k)^\top \beta^k \end{aligned} \tag{47}$$

4.2.2 Simulation and runs

In the sequel, $N = 1200$ and for all $i \in [\![1, N]\!]$, $n_i = 15$. For all $i \in [\![1, N]\!]$, the vector of predictors $(c_{ij,2}, 1 \leq j \leq n_i, 1 \leq i \leq n)$ is ranging from -20 to 50, 5 by 5. The other predictor takes different value for 3 groups of individuals of size $N/3$. If $c_{ij,2} \geq 0$, $c_{ij,1}$ is equal to 10, 20 or 30 and is null otherwise. We generate a synthetic dataset using the following generating values for the fixed and random effects $(\beta_1 = -4, \beta_2 = -0.5, \beta_3 = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2)$. We run the MBMCEM algorithm in this section. The size of the Monte Carlo batch increases polynomially, $M_k \triangleq M_0 + k^2$ with $M_0 = 50$. Figure 2 shows the convergence of the fixed effects $(\beta_1, \beta_2, \beta_3)$ estimates obtained with both the MCEM and the MBMCEM algorithms for different batch sizes. The effect of the batch size on the convergence rate differs from the previous example where the smaller batches implied faster convergence of the mini batch EM algorithm. Here, an optimal batch size of 50% accelerates the algorithm. Figure 2 highlights a non monotonic evolution of the convergence rate with respect to the size of the batch.

5 Proofs

5.1 Proof of Theorem 1

5.1.1 Proof of (i)

First, let us define for $\theta \in \Theta$

$$\bar{A}^k(\theta) \triangleq \sum_{i=1}^{N} A_i^k(\theta) \tag{48}$$

where for all $i \in [\![1, N]\!]$, A_i^k is defined in (19). For any $k \geq 1$ and for all $\theta \in \Theta$ the following decomposition plays a key role:

$$\bar{A}^k(\theta) = \bar{A}^{k-1}(\theta) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta) - \sum_{i \in I_k} A_i^{k-1}(\theta). \tag{49}$$

Since by construction $\bar{A}^k(\theta^k) \leq \bar{A}^k(\theta^{k-1})$, we get:

$$\bar{A}^k(\theta^k) \leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}). \tag{50}$$

Since for $i \in I_k$, $B_{i,\theta^{k-1}}$ is a surrogate of ℓ_i at θ^{k-1} we get that $B_{i,\theta^{k-1}}(\theta^{k-1}) = \ell_i(\theta^{k-1})$. On the other hand, for $i \in [\![1, N]\!]$, $A_i^{k-1} \equiv B_{i,\theta^{\tau_{i,k-1}}}$ and $B_{i,\theta^{\tau_{i,k-1}}}$ is a surrogate of ℓ_i at $\theta^{\tau_{i,k-1}}$, thus we obtain that $\ell_i(\theta^{k-1}) - A_i^{k-1}(\theta^{k-1}) \leq 0$. Plugging these two relations in (50) we obtain:

$$\begin{aligned} \bar{A}^k(\theta^k) &\leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} \ell_i(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}) \\ &\leq \bar{A}^{k-1}(\theta^{k-1}) \end{aligned} \tag{51}$$

As a result, the sequence $\left(\bar{A}^k(\theta^k)\right)_{k\geq 0}$ is monotonically decreasing. Since, under assumption M 3, this quantity is bounded from below with probability one, we obtain its almost sure convergence. Taking the expectations with respect to the sampling distributions of the previous inequalities implies the convergence of the (deterministic) sequence $\left(\mathbb{E}[\bar{A}^k(\theta^k)]\right)_{k\geq 0}$. Let us denote for all $\theta \in \Theta$ and a subset $J \subset [\![1, N]\!]$:

$$\begin{aligned} \ell_J(\theta) &\triangleq \sum_{i \in J} \ell_i(\theta) \\ A_J^{k-1}(\theta) &\triangleq \sum_{i \in J} A_i^{k-1}(\theta) \end{aligned} \tag{52}$$

Inequality (50) gives :

$$0 \leq \sum_{k=1}^{n} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \leq \sum_{k=1}^{n} \bar{A}^{k-1}(\theta^{k-1}) - \bar{A}^k(\theta^k) = \bar{A}^0(\theta^0) - \bar{A}^n(\theta^n) \tag{53}$$

Consequently, the sum of positive terms $\left(\sum_{k=1}^{n} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right)_{n \geq 1}$ converges almost surely and $\left(A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right)_{k \geq 1}$ converges almost surely to zero. The Beppo-Levi theorem and the Tower property of the conditional expectation imply:

$$\begin{aligned} \mathsf{M} \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] &= \sum_{k=0}^{\infty} \mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] \\ &= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right]\right] \end{aligned} \tag{54}$$

with $\mathbb{E}\left[\ell_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{p}{N}\ell(\theta^{k-1})$ and $\mathbb{E}\left[[A_{I_k}^{k-1}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{p}{N}\sum_{i=1}^{N} A_i^{k-1}(\theta^{k-1}) = \frac{p}{N}\bar{A}^{k-1}(\theta^{k-1})$ where $\mathcal{F}_{k-1} = \sigma(I_j, j \leq k-1)$ is the filtration generated by the sampling of the indices. We thus obtain:

$$\mathsf{M} = \frac{p}{N}\sum_{k=0}^{\infty} \mathbb{E}\left[\bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] = \frac{p}{N}\mathbb{E}\left[\sum_{k=0}^{\infty} \bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] < \infty \tag{55}$$

This last equation shows that:

$$\lim_{k \to \infty} \bar{A}^k(\theta^k) - \ell(\theta^k) = 0 \quad \text{a.s.} \tag{56}$$

which implies the almost sure convergence of $\left(\ell(\theta^k)\right)_{k \geq 0}$.

5.1.2 Proof of (ii)

Let us define, for all $k \geq 0$, $\bar{h}_k$ as:

$$\bar{h}^k : \vartheta \to \sum_{i=1}^{N} A_i^k(\vartheta) - \ell_i(\vartheta) \tag{57}$$

$\bar{h}^k$ is L-smooth with $L = \sum_{i=1}^{N} L_i$ since each of its component is L_i-smooth by definition of the surrogate functions. Using the particular parameter $\vartheta^k = \theta^k - \frac{1}{L}\nabla\bar{h}_k(\theta^k)$ we have the following classical inequality for smooth functions (cf. Lemma 1.2.3 in (Nesterov, 2007)):

$$\begin{aligned} 0 \leq \bar{h}^k(\vartheta^k) \leq \bar{h}^k(\theta^k) - \frac{1}{2L}\|\nabla\bar{h}^k(\theta^k)\|_2^2 \\ \implies \|\nabla\bar{h}^k(\theta^k)\|_2^2 \leq 2L\bar{h}^k(\theta^k) \end{aligned} \tag{58}$$

Using (56), we conclude that $\lim_{k \to \infty} \|\nabla\bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle \nabla\ell(\theta^k), \theta - \theta^k \rangle$ for any $\theta \in \Theta$ yields:

$$\langle \nabla\ell(\theta^k), \theta - \theta^k \rangle = \langle \nabla\bar{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla\bar{h}^k(\theta^k), \theta - \theta^k \rangle \tag{59}$$

Note that θ^k is the result of the minimisation of the sum of surrogates $\bar{A}^k(\theta)$ on the constrained set Θ, therefore $\langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k \rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\begin{aligned} \langle \nabla \ell(\theta^k), \theta - \theta^k \rangle &\geq -\langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \\ &\geq -\|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \end{aligned} \tag{60}$$

By minimizing over Θ and taking the infimum limit on k, we get:

$$\liminf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq -\lim_{k \to \infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0 \tag{61}$$

which is the Asymptotic Stationary Point Condition (ASPC).

5.2 Proof of Theorem 2

We preface the proof by the following lemma which is of independent interest:

Lemma 1. *Let $(V_k)_{k\geq 0}$ be a non negative sequence of random variables such that $\mathbb{E}[V_0] < \infty$. Let $(X_k)_{k\geq 0}$ a non negative sequence of random variables and $(E_k)_{k\geq 0}$ be a sequence of random variables such that $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|] < \infty$. If for any $k \geq 1$:*

$$V_k \leq V_{k-1} - X_k + E_k \tag{62}$$

then:

(i) for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.

(ii) the sequence $(\mathbb{E}[V_k])_{k\geq 0}$ converges and $\lim_{k\to\infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.

(iii) the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.

Remark 2. *Note that the result still holds if $(V_k)_{k\geq 0}$ is a sequence of random variables which is bounded from below by a deterministic quantity $M \in \mathbb{R}$.*

Proof. We first show that for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$. Note indeed that:

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \tag{63}$$

showing that $\mathbb{E}[V_k] \leq \mathbb{E}[V_0] + \mathbb{E}\left[\sum_{j=1}^{k} E_j\right] < \infty$.

Since $0 \leq X_k \leq V_{k-1} - V_k + E_k$ we also obtain for all $k \geq 0$, $\mathbb{E}[X_k] < \infty$. Moreover, since $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$, the series $\sum_{j=1}^{\infty} E_j$ converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{64}$$

Note that $\mathbb{E}[|W_k|] \leq \mathbb{E}[V_k] + \mathbb{E}\left[\sum_{j=k+1}^{\infty} |E_j|\right] < \infty$. For all $k \geq 1$, we get:

$$\begin{aligned} &W_k \leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ &\mathbb{E}[W_k] \leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \tag{65}$$

Hence the sequences $(W_k)_{k\geq 0}$ and $(\mathbb{E}[W_k])_{k\geq 0}$ are non increasing. Since for all $k \geq 0$, $W_k \geq -\sum_{j=1}^{\infty} |E_j| > -\infty$ and $\mathbb{E}[W_k] \geq -\sum_{j=1}^{\infty} \mathbb{E}[|E_j|] > -\infty$, the (random) sequence $(W_k)_{k\geq 0}$ converges a.s. to a limit W_∞ and the (deterministic) sequence $(\mathbb{E}[W_k])_{k\geq 0}$ converges to a limit w_∞. Since $|W_k| \leq V_0 + \sum_{j=1}^{\infty} |E_j|$, the Fatou lemma implies that:

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \tag{66}$$

showing that the random variable W_∞ is integrable.

In the sequel, set $U_k \triangleq W_0 - W_k$. By construction we have for all $k \geq 0$, $U_k \geq 0$, $U_k \leq U_{k+1}$ and $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$ and by the monotone convergence theorem, we get:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \tag{67}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \tag{68}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (65) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\begin{aligned} &\sum_{j=1}^{\infty} X_j \leq W_0 - W_\infty < \infty \\ &\sum_{j=1}^{\infty} \mathbb{E}[X_j] \leq \mathbb{E}[W_0] - w_\infty < \infty \end{aligned} \tag{69}$$

which concludes the proof of the lemma.

□

5.2.1 Proof of (i)

To study the convergence of the MBMCEM algorithm, we consider for all $k \geq 1$, the function $\vartheta \to \hat{B}_{i,\theta^{k-1}}(\vartheta)$ defined for all $i \in I_k$ and $\vartheta \in \Theta$ by:

$$\begin{aligned} \hat{B}_{i,\theta^{k-1}}(\vartheta) &\triangleq \hat{Q}_i^k(\vartheta, \theta^{k-1}) + H_i(\theta^{k-1}) \\ &= -\frac{1}{M_k} \sum_{m=0}^{M_k - 1} \log p_i(z_i^{k,m}, \vartheta) + l_i(\vartheta) + H_i(\theta^{k-1}) \end{aligned} \tag{70}$$

where $H_i(\theta^{k-1})$ is defined by (16). This function is a Monte Carlo approximation of the surrogate function $B_{i,\theta^{k-1}}$ defined for all $\vartheta \in \Theta$ and $i \in I_k$ as:

$$\begin{aligned} B_{i,\theta^{k-1}}(\vartheta) &\triangleq - \int_{\mathsf{Z}_i} \log p_i(z_i, \vartheta) p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) + l_i(\vartheta) + H_i(\theta^{k-1}) \\ &= \mathrm{KL}\big(\mathbb{P}_{i,\theta^{k-1}} \,\big\|\, \mathbb{P}_{i,\vartheta}\big) + l_i(\vartheta) \end{aligned} \tag{71}$$

Under assumption M 2, $\vartheta \to \hat{B}_{i,\theta^{k-1}}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$. Let us define, for $\theta \in \Theta$, $\hat{A}^k(\theta) \triangleq \sum_{i=1}^N \hat{A}_i^k(\theta)$ where $\hat{A}_i^k(\vartheta)$ are defined recursively as follows:

$$\hat{A}_i^k(\vartheta) = \begin{cases} \hat{B}_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ \hat{A}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{72}$$

(70) implies that for $k \geq 1$, $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N \hat{A}_i^k(\vartheta)$. Set for all $\theta \in \Theta$, $i \in [\![1, N]\!]$ and $k \geq 1$:

$$A_i^k(\theta) \triangleq B_{i,\theta^{\tau_{i,k}}}(\theta) \quad \text{and} \quad \bar{A}^k(\theta) = \sum_{i=1}^N A_i^k(\theta) \tag{73}$$

where $\tau_{i,k}$ is defined by (9). For any $k \geq 1$ and $\theta \in \Theta$ the following decomposition plays a key role:

$$\hat{A}^k(\theta) = \hat{A}^{k-1}(\theta) + \sum_{i \in I_k} \{\hat{B}_{i,\theta^{k-1}}(\theta) - \hat{A}_i^{k-1}(\theta)\} \tag{74}$$

Set the following notations:

$$\begin{aligned}
V_k &\triangleq \bar{A}^k(\theta^k), \\
X_k &\triangleq -\sum_{i \in I_k} \{B_{i,\theta^{k-1}}(\theta^{k-1}) - A_i^{k-1}(\theta^{k-1})\}, \\
E_k &\triangleq \sum_{i \in I_k} \{\hat{B}_{i,\theta^{k-1}}(\theta^{k-1}) - B_{i,\theta^{k-1}}(\theta^{k-1})\} \\
&\quad + \sum_{i \in I_k} \{A_i^{k-1}(\theta^{k-1}) - \hat{A}_i^{k-1}(\theta^{k-1})\} \\
&\quad + \bar{A}^k(\theta^k) - \hat{A}^k(\theta^k) + \hat{A}^{k-1}(\theta^{k-1}) - \bar{A}^{k-1}(\theta^{k-1}).
\end{aligned}$$

Combining (74) with $\bar{A}^k(\theta^k) = \bar{A}^k(\theta^k) - \hat{A}^k(\theta^k) + \hat{A}^k(\theta^k)$ and $\hat{A}^k(\theta^k) \leq \hat{A}^k(\theta^{k-1})$, we obtain:

$$V_k \leq V_{k-1} - X_k + E_k. \tag{75}$$

where A_i^{k-1} and $\bar{A}^k$ are defined in (73). We now check the assumptions of Lemma 1. Note first that the sequence $(V_k)_{k\geq 0}$ is bounded from below under assumption M 3. We now check that $X_k \geq 0$ thanks to the following relation:

$$\begin{aligned}
X_k &= -0 - \sum_{i \in I_k} \ell_i(\theta^{k-1}) + \sum_{i \in I_k} \mathrm{KL}\big(\mathbb{P}_{i,\theta^{\tau_{i,k-1}}} \,\big\|\, \mathbb{P}_{i,\theta^{k-1}}\big) + \sum_{i \in I_k} \ell_i(\theta^{k-1}) \\
&= \sum_{i \in I_k} \mathrm{KL}\big(\mathbb{P}_{i,\theta^{\tau_{i,k-1}}} \,\big\|\, \mathbb{P}_{i,\theta^{k-1}}\big) \geq 0.
\end{aligned} \tag{76}$$

We finally have to prove the convergence of the series $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|]$. For this purpose, we will show that for all $i \in [\![1, N]\!]$:

$$\sum_{k=0}^{\infty} \mathbb{E}\Big[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\Big] < \infty \tag{77}$$

We have, using the Tower property of the conditional expectation and the Jensen inequality:

$$\mathbb{E}\Big[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\Big] \leq \mathbb{E}\Big[\mathbb{E}_{i,\theta^{\tau_{i,k}}}\Big[\sup_{\vartheta \in \Theta} |\hat{A}_i^k(\vartheta) - A_i^k(\vartheta)|\Big]\Big] \tag{78}$$

Under assumption M 4 applied with the function $\vartheta \to \hat{A}_i^k(\vartheta)$, for all $i \in [\![1, N]\!]$ we have:

$$\mathbb{E}_{i,\theta^{\tau_{i,k}}}\left[\sup_{\vartheta\in\Theta} |\hat{A}_i^k(\vartheta) - A_i^k(\vartheta)|\right] \leq C_i M_{\tau_{i,k}}^{-1/2} \tag{79}$$

where C_i is a finite constant defined by (35) and $\tau_{i,k}$ is defined by (9). Thus, we have that:

$$\mathbb{E}\left[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\right] \leq C_i \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] \tag{80}$$

Since, any index i is included in a mini-batch with a probability equal to $\frac{p}{N}$ conditionally independently from the past, we obtain that:

$$\mathbb{E}[M_{\tau_{i,k}}^{-1/2}] = \sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \tag{81}$$

Taking the infinite sum of this term yields:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] &= \sum_{k=1}^{\infty}\sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \\
&= \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{k-(l+1)} \frac{p}{N} \mathbb{1}_{\{l \leq k-1\}} M_l^{-1/2} \\
&= \frac{p}{N} \sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{-(l+1)} M_l^{-1/2} \sum_{k=l+1}^{\infty} \left(1 - \frac{p}{N}\right)^{k} \\
&= \sum_{l=0}^{\infty} M_l^{-1/2}
\end{aligned} \tag{82}$$

which proves identity (77), using assumption M 5. By summing over the indices $i \in [\![1, N]\!]$, (77) implies:

$$\sum_{k=0}^{\infty} \mathbb{E}\left[|\hat{A}^k(\theta^k) - \bar{A}^k(\theta^k)|\right] < \infty \tag{83}$$

Hence, we obtain that $\sum_{k=0}^{\infty} |\hat{A}^k(\theta^k) - \bar{A}^k(\theta^k)| < \infty$ almost surely which implies that:

$$\lim_{k\to\infty} \hat{A}^k(\theta^k) - \bar{A}^k(\theta^k) = 0 \quad \text{a.s.} \tag{84}$$

Similarly, using assumption M 4 applied for all $i \in [\![1, N]\!]$, with the function $\vartheta \to \nabla \hat{A}_i^k(\vartheta)$ we obtain:

$$\lim_{k\to\infty} \nabla\hat{A}^k(\theta^k) - \nabla\bar{A}^k(\theta^k) = 0 \quad \text{a.s.} \tag{85}$$

It follows from (77) and (83) that $\sum_{k=0}^{\infty} \mathbb{E}\left[|E_k|\right] < \infty$ and that the series $\sum_{k=0}^{\infty} \epsilon_k$ converges to an almost surely finite limit. Hence by Lemma 1 and (84) we get:

- the sequence $\left(\bar{A}^k(\theta^k)\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \chi_k$ converge a.s.
- the sequence $\left(\mathbb{E}\left[\bar{A}^k(\theta^k)\right]\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \mathbb{E}\left[X_k\right]$ converge with $\lim\limits_{k\to\infty} \mathbb{E}\left[\bar{A}^k(\theta^k)\right] = \mathbb{E}[\lim\limits_{k\to\infty} \bar{A}^k(\theta^k)]$.
- the sequence $\left(\hat{A}^k(\theta^k)\right)_{k\geq 0}$ converges a.s. and the sequence $\left(\mathbb{E}\left[\hat{A}^k(\theta^k)\right]\right)_{k\geq 0}$ converges.

Now, we have to prove the almost-sure convergence of the sequence $\left(\ell(\theta^k)\right)_{k\geq 0}$ and the convergence of $\left(\mathbb{E}\left[\ell(\theta^k)\right]\right)_{k\geq 0}$. Using the same argument as in (54) and (55), we have:

$$\mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] = \frac{p}{N}\,\mathbb{E}\left[\sum_{k=1}^{\infty}\{\bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\}\right] < \infty \tag{86}$$

showing that:

$$\begin{aligned} &\lim_{k\to\infty} \mathbb{E}\left[\bar{A}^k(\theta^k) - \ell(\theta^k)\right] = 0 \\ &\lim_{k\to\infty} \bar{A}^k(\theta^k) - \ell(\theta^k) = 0 \quad \text{a.s.} \end{aligned} \tag{87}$$

showing that the sequence $\left(\mathbb{E}\left[\ell(\theta^k)\right]\right)_{k\geq 0}$ converges and that $\left(\ell(\theta^k)\right)_{k\geq 0}$ converges a.s.

5.2.2 Proof of (ii)

Consider for any $k \geq 0$, the L smooth function $\bar{h}^k$ defined by (57). Using (58) and (87) we get $\lim\limits_{k\to\infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle \nabla \ell(\theta^k), \theta - \theta^k\rangle$ for any $\theta \in \Theta$ yields:

$$\begin{aligned} \langle \nabla \ell(\theta^k), \theta - \theta^k\rangle &= \langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k\rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k\rangle \\ &= \langle \nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k), \theta - \theta^k\rangle + \langle \nabla \hat{A}^k(\theta^k), \theta - \theta^k\rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k\rangle \end{aligned} \tag{88}$$

Note that θ^k is the result of the minimisation of $\hat{A}^k(\theta)$ on the constrained set Θ, therefore for all $\theta \in \Theta$, $\langle \nabla \hat{A}^k(\theta^k), \theta - \theta^k\rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\begin{aligned} \langle \nabla \ell(\theta^k), \theta - \theta^k\rangle &\geq \langle \nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k), \theta - \theta^k\rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k\rangle \\ &\geq -\|\nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 - \|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \end{aligned} \tag{89}$$

By minimizing over Θ and taking the infimum limit, we get, using (85):

$$\liminf_{k\to\infty} \inf_{\theta\in\Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k\rangle}{\|\theta - \theta^k\|_2} \geq -\lim_{k\to\infty}\left(\|\nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k)\|_2 + \|\nabla \bar{h}^k(\theta^k)\|_2\right) = 0 \tag{90}$$

which is the Asymptotic Stationary Point Condition (ASPC).

References

F. Bach A. Bietti and A. Cont. An online EM algorithm in hidden (semi-)Markov models for audio segmentation and clustering. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

W. Byrne A. Gunawardana. Convergence theorems for Generalized Alternating Minimization procedures. 2005.

S. Ahn and J. A. Fessler. Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms. *IEEE Trans. Med. Imag.*, 22, 2003.

J. Biscarat. Almost sure convergence of a class of stochastic algorithms. *Stochastic Process. Appl.*, 1994.

Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8. URL https://doi.org/10.1201/b10905.

K.B. Welch B.T. West and A.T. Galecki. Linear Mixed Models: A Practical Guide Using Statistical Softwar. *Chapman Hall/CRC*, 2006.

S. Trevezas C. Baey and P.H. Cournede. A non linear mixed effects model of plant growth and estimation via stochastic variants of the EM algorithm. *Comm. Statist. Theory Methods*, 45, 2016.

O. Cappe. Online EM algorithm for hidden Markov models. 2009.

O. Cappe and E. Moulines. Online EM Algorithm for Latent Data Models. *J. Roy. Statist. Soc*, 2009.

A. Chakraborty and K. Das. Inferences for joint modelling of repeated ordinal scores and time to event data. *Comput. Math. Methods Med.*, 11, 2010.

K. Chan and J. Ledolter. Monte Carlo EM estimation for time series models involving count. *J. Amer. Statist. Asooc.*, 1995.

Laird Dempster and Rubin. Maximum likelihood from incomplete likelihood data via EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser.*, 1977.

P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31(2):393–427, 1995. ISSN 0246-0203. URL http://www.numdam.org/item?id=AIHPB_1995__31_2_393_0.

A. Lavenu E. Comets and M. Lavielle. Parameter estimation in nonlinear mixed effect models using saemix. *Journal of Statistical Software*, 2017.

H. Erdogan and J. A. Fessler. Ordered subsets algorithms for transmission tomography. *Phys. Med. Biol.*, 44, 1999.

R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22, 1925.

G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for Curved Exponential Families. *The Annals of Statistics*, 2003.

G.R. Gene H. Ing-Tsung, R. Anand. Provably convergent OSEM-like reconstruction algorithm for emission tomography. *Medical Imaging 2002: Image Processing*, 2002.

H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imag.*, 4, 1994.

Y. Wang G. Herman B. Zhang J. Yang J. Xu, G. Ye. Incremental EM for Probabilistic Latent Semantic Analysis on Human Action Recognition. *Advanced Video and Signal Based Surveillance*, 2009.

J. S. Kole and F. J. Beekman. Evaluation of the ordered subset convex algorithm for cone-beam CT. *Phys. Med. Biol.*, 50, 2005.

R. Levine and G. Casella. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10, 2001.

T.A. Louis. Finding the Observed Information Matrix when using the EM algorithm. *JRSS*, 44, 1982.

J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM*, 2015.

G. McLachlan and T. Krishnan. The EM Algorithm and Extensions. 2007.

Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-73182-9. URL https://doi.org/10.1017/CBO9780511626630. With a prologue by Peter W. Glynn.

R. Neath. On Convergence Properties of the Monte Carlo EM Algorithm. 2012.

Y. Nesterov. Gradient methods for minimizing composite objective function. 2007.

A.R. De Pierro and ME. Yamagishi. Fast EM-like methods for maximum" a posteriori" estimates in emission tomography. *IEEE Trans. Med. Imag.*, 4, 2001.

G.Hinton R. Neal. A view of the EM algorithm that justifies incremental, sparse, and other variants. 2007.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387212396.

A. H. Lee S. Ng, G. J. McLachlan. An incremental EM-based learning approach for on-line prediction of hospital resource utilization. *Artificial Intelligence in Medicine*, 2005.

G. Verbeke and G. Molenberghs. Linear Mixed Models for Lon- gitudinal Data. *Springer*, 2000.

G. Wei and M. Tanner. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.*, 1990.

A. J. Brown X. R. Wang and B. Upcroft. Applying Incremental EM to Bayesian Classifiers in the Learning of Hyperspectral Remote Sensing Data. *IEEE*, 2005.

Glossary

directional derivative Let consider a function $f : \Theta \to \mathbb{R}$. For all $(\theta, \theta') \in \Theta^2$, the following limit is called the directional derivative of f at θ in the direction $\theta' - \theta$: $\nabla f(\theta, \theta' - \theta) \triangleq \lim_{t \to 0} (f(\theta + t(\theta' - \theta)) - f(\theta))/t$. 1

iterative application Let $\mathcal{X}$ be a set and $x_0 \in \mathcal{X}$ a given point. Then an iterative algorithm A with initial point x_0 is a point-to-set mapping $A \colon \mathcal{X} \to \mathcal{X}$ which generates a sequence $\{x_n\}_{n=1}^{\infty}$ according to

$$x_{n+1} \in A(x_n) \tag{91}$$

. 1, 7, 13

smooth A function $f : \Theta \to \mathbb{R}$ is called L-smooth when it is differentiable and when its gradient ∇f is L-Lipschitz continuous.. 1, 5, 7

stationary point Let consider a function $f : \Theta \to \mathbb{R}$ such as f admits a directional derivative $\nabla f(\theta, \theta' - \theta)$ for all $(\theta, \theta') \in \Theta^2$. We say that θ is a stationary point if for all $\theta' \in \Theta$, $\nabla f(\theta, \theta' - \theta) \geq 0$. 1