On the Convergence Properties of the Stochastic Version of the Mini-Batch EM algorithm

Belhal Karimi [1] **Marc Lavielle** [1] **Eric Moulines** [1]

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 mini-batch version of the Monte Carlo EM (MBMCEM) 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, 1998), 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 terms of efficiency of computation, (Cappe & Moulines, 2007) introduced an online version where the whole dataset is not analysed at each iteration but a growing batch of it only. In 1994, (Hudson & Larkin, 1994) introduced the ordered subsets EM (OS-EM) algorithm in the context of emission tomography for accelerated image reconstruction. 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 & Yamagishi, 2001). Since then, block-iterative EM methods have been very popular in the medical imaging community for tomographic image reconstruction (Erdogan & Fessler, 1999; H. Ing-Tsung, 2002; Ahn & Fessler, 2003; Kole & Beekman, 2005) due to their remarkably fast convergence rates. In the same line of work, (Byrne, 1998) presents a rescaled block-iterative method for the Poisson model with faster convergence to the maximum likelihood estimate. Moreover, incremental algorithm have been very useful for fitting mixture model as described in (McLachlan, 2003) where the authors give an overview on the choice of number of blocks at each iteration of the algorithm. 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 with a sampling scheme repeated from a pass over the data to another. In this article, we apply the MISO (Minimisation by Incremental Surrogate Optimisation) framework introduced in (Mairal, 2015) to provide convergence properties of the mini-batch EM algorithm. 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 & 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

[*]Equal contribution [1]Department of Applied Mathematics, Ecole Polytechnique, Palaiseau, France. Correspondence to: Belhal Karimi <belhal.karimi@polytechnique.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

effects model of plant growth (C. Baey & Cournede, 2016) or to do inference for joint modelling of time to event data coming from clinical trials in (Chakraborty & Das, 2010). This algorithm has been vastly studied in (Levine & Casella, 2001) or (McLachlan & Krishnan, 2007) and its convergence properties have been derived, initially, in (Biscarat, 1994; Chan & Ledolter, 1995) and more recently in (Neath, 2012) and (Fort & Moulines, 2003). In this contribution, we adapt the MISO framework, using stochastic incremental surrogates, to the MBMCEM algorithm in order to prove its almost-sure 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 $\mathsf{Z} = \bigtimes_{i=1}^N \mathsf{Z}_i$ 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 & 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 θ^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 θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$
2. For all $i \in I_k$, 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 $R_i^k(\theta)$ are defined recursively as follows:

$$R_i^k(\vartheta) = \begin{cases} Q_i(\vartheta, \theta^{k-1}) & \text{if } i \in I_k. \\ R_i^{k-1}(\vartheta) & \text{otherwise.} \end{cases} \tag{7}$$

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, 1998), 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 $\vartheta \in \Theta$:

$$\begin{aligned} &\int_{\mathsf{Z}_i} |\nabla f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty, \\ &\int_{\mathsf{Z}_i} p_i(z_i, \vartheta) |\log f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty. \end{aligned} \tag{10}$$

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:

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

Thus, sufficient conditions to verify M2d. are M2c. 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:

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

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 analyse 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:

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

We verify **S.1** using assumption M2. Since $\vartheta \to \mathrm{KL}(\mathbb{P}_{i,\theta} \,\|\, \mathbb{P}_{i,\vartheta})$ 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 $h_i(\vartheta) = \mathrm{KL}(\mathbb{P}_{i,\theta} \,\|\, \mathbb{P}_{i,\vartheta})$. Under M2c. and M2d., 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 θ^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 θ^{k-1}:

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

$$A_i^k(\vartheta) = \begin{cases} B_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k. \\ A_i^{k-1}(\vartheta) & \text{otherwise.} \end{cases} \tag{19}$$

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.*

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

Proof. The proof is postponed to the supplementary material □

We observe that in the unconstrained case, we have:

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

wihch 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 $s = (s_i, 1 \le i \le N) \in \mathsf{S}$ where $\mathsf{S} = \bigtimes_{i=1}^N \mathsf{S}_i$, the function $L(s; \theta)$ by:

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

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

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

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 θ^0, for all $i \in [\![1, N]\!]$ compute $s_i^0 = \bar{s}(\theta^0)$.

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 $i \in [\![1, N]\!]$, compute s_i^k such as:

$$s_i^k = \begin{cases} \bar{s}(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases} \tag{27}$$

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

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 $k \ge 1$:

$$\hat{Q}_i^k(\vartheta, \theta^{k-1}) \triangleq \frac{1}{M_k} \sum_{m=0}^{M_k - 1} \log f_i(z_i^{k,m}, \vartheta). \tag{28}$$

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 $k \ge 1$, $P^k(z, A) \triangleq \int_A P^{k-1}(z, \mathrm{d}z') P(z', A)$. 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 & 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 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 a function $\vartheta \to \hat{R}_i^0(\vartheta) = \hat{Q}_i^0(\vartheta, \theta^0)$ defined by (28).

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 $m \in [\![1, M_k]\!]$, 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 (28).
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:

$$\hat{R}_i^k(\vartheta) = \begin{cases} \hat{Q}_i^k(\vartheta, \theta^{k-1}) & \text{if } i \in I_k. \\ \hat{R}_i^{k-1}(\vartheta) & \text{otherwise.} \end{cases} \tag{29}$$

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:

$$C_i(q_i) \triangleq \sup_{\theta \in \Theta} \sup_{M > 0} M^{-p/2} \mathbb{E}_{i,\theta} \left[\sup_{\vartheta \in \Theta} \left| \sum_{m=0}^{M-1} \left\{ q_i(z_i^m, \vartheta) - \int_{\mathsf{Z}_i} q_i(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \right\} \right| \right] < \infty. \tag{30}$$

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

$$C_i(\log f_i) < \infty \quad \text{and} \quad C_i(\nabla \log f_i) < \infty. \tag{31}$$

The assumption M4 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 & Tweedie, 2009)). Finally, we consider the following assumption on the number of simulations:

M 5. $\{M_k\}_{k \geq 0}$ is a non decreasing 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 iterapp 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 the supplementary material □

4. Numerical examples

4.1. A Linear mixed effects model

4.1.1. The model

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

$$y_i = A_i \theta + B_i z_i + \epsilon_i. \tag{32}$$

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 known. For all $i \in [\![1, N]\!]$, the random variables y_i, $y_i|z_i$ and $z_i|y_i$ are distributed according to the following Gaussian distributions:

$$\begin{aligned} y_i|z_i &\sim \mathcal{N}(A_i\beta + B_i z_i, \Sigma), \\ z_i|y_i &\sim \mathcal{N}(\mu_i, \Gamma_i). \end{aligned} \tag{33}$$

where:

$$\begin{aligned} \Gamma_i &= (B_i^\top \Sigma^{-1} B_i + \Omega^{-1})^{-1}, \\ \mu_i &= \Gamma_i B_i^\top \Sigma^{-1} (y_i - A_i \theta). \end{aligned} \tag{34}$$

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

$$\begin{aligned} &\tilde{S}_i(z_i) \triangleq z_i \quad \text{and} \quad \bar{s}_i(\theta) = \Gamma_i B_i^\top \Sigma^{-1} (y_i - A_i \theta) \\ &\psi_i(\theta) \triangleq (y_i - A_i \theta)^\top \Sigma^{-1} (y_i - A_i \theta) \\ &\phi_i(\theta) \triangleq B^\top \Sigma^{-1} (y_i - A_i \theta) \end{aligned} \tag{35}$$

Maximising $L(s, \theta)$, defined in (25), with respect to θ yields to 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 picking a set I_k and compute $\theta^k = \hat{\theta}(s^k)$ 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 Convergence

We generate a synthetic dataset using the following generating values: $\theta = (\theta_1 : 4, \theta_2 : 9)$ with $N = 100$ individuals and for all $i \in [\![1, N]\!]$, $n_i = 10$ observations per individual. We also generate random integer-valued 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 $(\theta_1^k, \theta_2^k)_{k=0}^K$ obtained with $K = 6000$ iterations of the EM, the MBEM where half of the data is considered at each iteration and the Incremental EM (i.e. when a single individual is considered at each iteration).

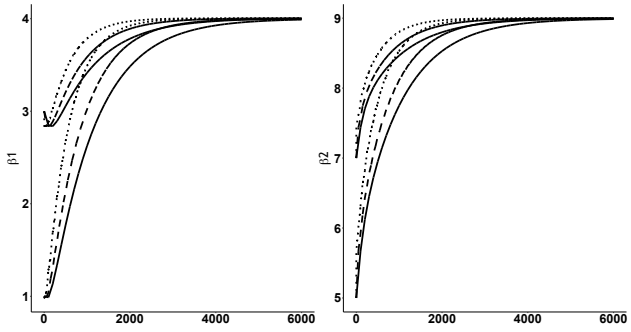

Figure 1. Convergence of the vector of parameter estimates β^k for the EM (solid), the MBEM 50 % (dashed) and Incremental EM (dotted).

4.1.3. Monte Carlo study

After having run the MBEM algorithm on a single synthetic dataset starting from different initial values and in order to justify the effectiveness of the MBEM algorithm, we present in this section a Monte Carlo study that consists in generating several simulated datasets using the same generating values. Those simulated datasets were generated using the following values: $\beta = (\beta_1 : 4, \beta_2 : 9)$. We perform a single run of the full EM, the MBEM using half of the data at each iteration and the incremental EM on the $D = 100$ simulated datasets starting with initial values $(\beta_1^0 : 1, \beta_2^0 : 4)$ and obtain a sequence $(\beta_{(d)}^K, 1 \leq d \leq D)$ of estimates. In order to compare the speed of convergence of the parameter estimate to the target value, we compute the error between the parameter $\beta_{(d)}^k$ at iteration k and its target value $\beta_{(d)}^K$ (the resulting parameter after K iterations) for each dataset d. This error E^k is defined as follows:

$$E^k \triangleq \frac{1}{D} \sum_{d=1}^{D} \left[\beta_{(d)}^k - \beta_{(d)}^K \right]^2. \tag{36}$$

Figure 2 shows the evolution of this error at each iteration depending on the batch size considered. The speed of convergence is a monotone function of the batch size in this case, the bigger the batch the slower the convergence.

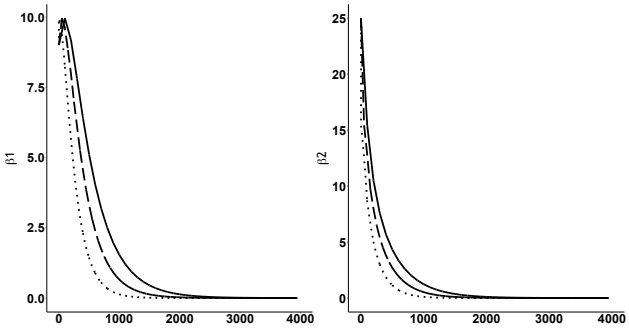

Figure 2. Average error E^k for the EM (solid), MBEM (50%) (dashed) and Incremental EM (dotted).

4.2. A logit normal generalized linear mixed model

4.2.1. The model

Let $y = (y_i, 1 \leq i \leq N)$ be the vector of observations where for each individual i, $y_i = (y_{ij}, 1 \leq j \leq n_i)$ is a sequence of conditionally independent random variables. Let $z = (z_i, 1 \leq i \leq N)$ be a vector of latent data where $z_i = (z_{i,1}, z_{i,2}, z_{i,3}) \in \mathbb{R}^3$. In this example, observations are ordinal data that take their value in the set $\{0, 1, 2, 3\}$ where each cumulative odds ratio, defining the model, are given by:

$$\begin{aligned} &\text{logit}(\mathbb{P}(y_{ij} \leq 0 | z_i)) = z_{i,1}, \\ &\text{logit}(\mathbb{P}(y_{ij} \leq 1 | z_i)) = z_{i,1} + \exp(z_{i,2}), \\ &\text{logit}(\mathbb{P}(y_{ij} \leq 2 | z_i)) = z_{i,1} + \exp(z_{i,2}) + \exp(z_{i,3}). \end{aligned} \tag{37}$$

Where the vector of latent variables is assumed to be distributed according to a Gaussian distribution:

$$z_i \sim \mathcal{N}(\beta, \Omega). \tag{38}$$

where $\beta = (\beta_1, \beta_2, \beta_3)$ and $\Omega = \text{diag}(\omega_1^2, \omega_2^2, \omega_3^2)$. The complete log likelihood is expressed as:

$$\log f(z, \theta) \propto \sum_{i=1}^{N} \left\{ \log(\mathbb{P}(y_i|z_i)) - \frac{1}{2}\log(|\Omega|) \right\} - \sum_{i=1}^{N} \left\{ \frac{1}{2}(z_i - \beta)^\top \Omega^{-1}(z_i - \beta) \right\}. \tag{39}$$

The quantity $p_i(z_i, \theta)$ is proportional to:

$$p_i(z_i, \theta) \propto -\frac{1}{|\Omega|^{1/2}} \mathbb{P}(y_i|z_i) \exp(-\frac{1}{2}(z_i - \beta)^\top \Omega^{-1}(z_i - \beta)). \tag{40}$$

Clearly the expectation of the complete log likelihood with respect to the conditional distribution (40) is intractable. Thus we consider the MCEM algorithm and the MBMCEM algorithm, which requires to simulate random draws from the distribution given by (40). We use a Random Walk Metropolis-Hastings algorithm (Brooks et al., 2011) with Gaussian proposal to perform those draws.

4.2.2. Simulation and Convergence

In the sequel, $N = 400$ and for all $i \in [\![1, N]\!]$, $n_i = 10$. We generate a synthetic dataset using the following generating values for the fixed and random effects $(\beta_1 = 0, \beta_2 = 1, \beta_3 = 1, \omega_1 = 0.2, \omega_2 = 0.2, \omega_3 = 0.1)$. We run the MBMCEM algorithm in this section. Three runs are executed starting from different initial values to study the convergence behaviour of both algorithms depending on how far they start from the true solution. Figure 3 shows the convergence of the fixed effect β_2 and the random effect ω_2 estimates obtained with both the MCEM and the MBMCEM algorithms for different batch sizes (half and a quarter of the data).

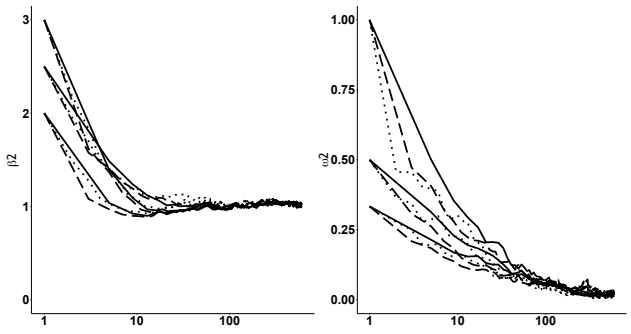

Figure 3. Convergence of the fixed parameter β_2 and its variance ω_2 for the MCEM (solid), MBMCEM 50% (dashed) and MBMCEM 25% (dotted). Logarithmic scale for the x-axis.

4.2.3. Monte Carlo study

After having run the MBMCEM algorithm on a single synthetic dataset starting from different initial values and in order to justify the effectiveness of our technique, we present here a similar monte carlo study as in the previous section. The $D = 100$ simulated datasets were generated using the following values: $(\beta_1 = 0, \beta_2 = 1, \beta_3 = 1, \omega_1 = 0.2, \omega_2 = 0.2, \omega_3 = 0.1)$. We perform a single run of the full MCEM, the MBMCEM using a quarter and a half of the data at each iteration on the D simulated datasets starting with initial values $(\beta_1^0 = 2, \beta_2^0 = 2, \beta_3^0 = 2, \omega_1^0 = 1, \omega_2^0 = 1, \omega_3^0 = 1)$ and obtain a sequence $(\theta_{(d)}^K, 1 \leq d \leq D)$ of estimates with $\theta^K = (\beta_1^K, \beta_2^K, \beta_3^K, \omega_1^K, \omega_2^K, \omega_3^K)$. In order to compare the speed of convergence of the parameter estimate to the target value, we compute the error defined by (36). Figure 4 shows the evolution of this error at each iteration for each. We observe relatively similar convergence rates for the population parameter and a faster convergence for smaller batches as far as covariance estimation.

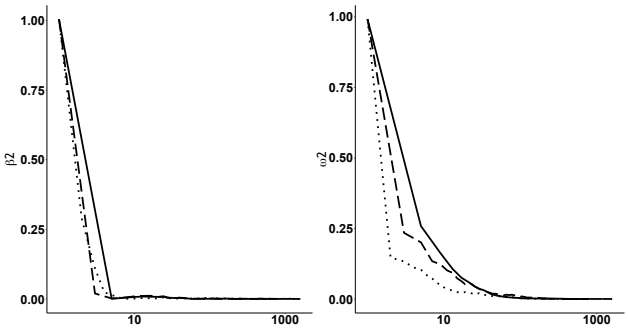

Figure 4. Average error for the fixed parameter β_2 and its variance ω_2 for the MCEM (solid), MBMCEM 50% (dashed) and MBMCEM 25% (dotted). Logarithmic scale for the x-axis.

5. Conclusion

In this paper, we have derived the convergence properties of the mini-batch EM algorithm using the MISO framework based on the minimisation of surrogate functions at each iteration. In order to derive the convergence properties of the stochastic version of this mini-batch algorithm, we have a presented an adaptation of this framework using the Monte Carlo integration of those surrogates. Numerical examples presented in this work showcased different convergence behaviours with respect to the number of observations considered at each iteration of both algorithms.

References

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

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

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

Brooks, Steve, Gelman, Andrew, Jones, Galin L., and Meng, Xiao-Li (eds.). *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.

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

Byrne, Charles L. Accelerating The EMML Algorithm and Related Iterative Algorithms by Rescaled Block-Iterative Methods. *IEEE Trans. Imag. Proc.*, 7, 1998.

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

Cappe, O. and Moulines, E. Online EM Algorithm for Latent Data Models. 2007.

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

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

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

Doukhan, P., Massart, P., and Rio, E. Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31, 1995.

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

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

Fort, G. and Moulines, E. Convergence of the Monte Carlo Expectation Maximization for Curved Exponential Families. *The Annals of Statistics*, 31:1220–1259, 2003.

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

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

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

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

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

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

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

McLachlan, S. K. NgG. J. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Statistics and Computing*, 2003.

Meyn, Sean and Tweedie, Richard L. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

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

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

R. Neal, G.Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pp. 355–368, 1998.

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

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