Properties of the Stochastic Approximation EM Algorithm with Mini-batch Sampling

Estelle Kuhn

INRAE, MaIAGE

April 9, 2021

joint work with Catherine Matias and Tabea Rebafka

Outline

Incomplete data framework

$\Longrightarrow$ Observe data which are related to unobserved data

- signal deconvolution
- source separation
- pharmacokinetic
- graph analysis
- images matching
- ...

$\Longrightarrow$ Some statistical models with latent variables

- hidden Markov model
- mixed effects model
- frailty model
- stochastic block model
- ...

Optimization algorithms for estimation

- very large datasets available
- long computing times of iterative algorithms (EM,...) when using of all data points in every iteration
- use only a part of the observations during one iteration in order to accelerate convergence
  - **online algorithms :** process a single observation per iteration handled in the order of arrival (**????**)
  - **mini-batch algorithms :** use (randomly chosen) subsets of observations (**????**)

General latent variable model

Observed data $y \rightarrow$ observed variable
Missing data $z \rightarrow$ latent variable

Assume the complete likelihood f of (y, z) belongs to a parametric family $\{f(y, z; \theta), \theta \in \Theta\}$.

Objective : Compute the value θ^{ML} that maximises the likelihood $g(y; \theta)$ of the observed data

Heuristics : if z were observed, then consider $\log f(y, z; \theta)$
$\Longrightarrow$ consider $E[\log f(y, z; \theta)|y; \theta]$.

The EM algorithm [Dempster et al. (1977), Wu (1983), Vaida (2005)]

$\Longrightarrow$ Estimation in missing data model

Iteration k of the algorithm :

- Expectation step :

$$Q(\theta|\theta_{k-1}) = E[\log f(y, z; \theta)|y; \theta_{k-1}]$$

- Maximization step :

$$\theta_k = Argmax \;\; Q(\theta|\theta_{k-1})$$

$+$ increase of $Q \Longrightarrow$ increase of the observed likelihood g

$+$ converges toward a stationary point $\widehat{\theta}_g$ of g

$-$ theory in exponential model

$-$ nature of the limit point

$-$ convergence depends on the initial guess

$-$ expression of $Q(\theta|\theta')$ often analytically intractable

Some existing methods

- Methods based on approximations of the likelihood
  No convergence property or with non realistic assumptions, default of convergence.

- Methods based on the exact likelihood
  - MCEM algorithm (Walker, 1996 ; Fort and Moulines, 2004)
  - SAEM algorithm (Delyon, Lavielle and Moulines, 1999)
  - ...

  Convergence property for some but high computation times and/or non realistic assumptions.

Heuristics of the stochastic approximation

Quantity of interest in the EM algorithm :

$$Q(\theta|\theta') = E[\log f(y, z; \theta)|y; \theta']$$

Sequential approximation of this quantity : at iteration k

- simulate z_k
- compute
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log f(y, z_k; \theta) - Q_{k-1}(\theta)\right]$

Then, we have :

$$\frac{Q_k(\theta) - Q_{k-1}(\theta)}{\gamma_k} = E[\log f(y, z; \theta)|y; \theta] - Q_{k-1}(\theta) + \log f(y, z_k; \theta) - E[\log f(y, z; \theta)|y; \theta]$$

$$\frac{Q_k(\theta) - Q_{k-1}(\theta)}{\gamma_k} \approx E[\log f(y, z; \theta)|y; \theta] - Q_{k-1}(\theta) + e_k$$

If $z_k \sim p(\cdot|y, \theta)$ then $e_k \approx 0$

Stochastic Approximation of the EM algorithm (Delyon et al (1999))

Iteration k of the algorithm :

- Simulation step : $z^k \sim p_{\theta_{k-1}}$
  where p_θ is the posterior distribution
- Stochastic approximation :
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k[\log f(y, z^k, \theta) - Q_{k-1}(\theta)]$ where (γ_k) is a decreasing sequence of positive step-sizes.
- Maximisation step : $\theta_k = \arg\max Q_k(\theta)$

$+$ converges almost surely toward a stationary point $\widehat{\theta}_g$ of g
– theory in exponential model
– nature of the limit point
– convergence depends on the initial guess

Coupling MCMC with Stochastic Approximation of the EM algorithm (K. et al (2004), Allassonnière et al. (2015))

Iteration k of the algorithm :

- Simulation step : $z^k \sim \Pi(z^{k-1}, \cdot; \theta_{k-1})$ where $\Pi(\cdot, \cdot; \theta)$ is a transition probability of an ergodic Markov Chain having the posterior distribution p_θ as stationary distribution,
- Stochastic approximation
- Maximisation step

SAEM algorithm using Metropolis-Hastings-within-Gibbs

Recall that $z \in \mathbb{R}^n$

Iteration k of the algorithm :

- Simulation step : for i in $1 : n$
  $z^k \sim \Pi_{\mathbf{i}}(z^{k-1}, \cdot; \theta_{k-1})$
  where Π_i is the Metropolis kernel acting only on the i-th coordinate
- Stochastic approximation
- Maximisation step

Mini-batch MCMC-SAEM algorithm (Kuhn, Matias, Rebafka (2019))

Recall that $z \in \mathbb{R}^n$

Let $0 < \alpha \leq 1$ be the mini-batch proportion

Iteration k of the algorithm :

- Simulation step :
  Select a subset $\mathcal{I}_k$ of indices :
  Simulate the mini-batch size $r \sim \text{Bin}(n, \alpha)$
  Choose r indices from $\{1, \ldots, n\}$ denoted by $\mathcal{I}_k$
  Simulate for i in $\mathcal{I}_k$
  $z^k \sim \Pi_i(z^{k-1}, \cdot; \theta_{k-1})$
- Stochastic approximation
- Maximisation step

Mini-batch MCMC-SAEM algorithm (Kuhn, Matias, Rebafka (2019))

Recall that $z \in \mathbb{R}^{\mathbf{n}}$

Let $0 < \alpha \leq 1$ be the mini-batch proportion

Iteration k of the algorithm :

- Simulation step :
  Select a subset $\mathcal{I}_k$ of indices :
  Simulate the mini-batch size $r \sim \text{Bin}(n, \alpha)$
  Choose r indices from $\{1, \ldots, n\}$ denoted by $\mathcal{I}_k$
  for i in $\mathcal{I}_k$
  $z^k \sim \Pi_{\mathbf{i}}(z^{k-1}, \cdot; \theta_{k-1})$
  where Π_i is the Metropolis kernel acting only on the i-th coordinate
- Stochastic approximation :
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k[\log f(y, z^k, \theta) - Q_{k-1}(\theta)]$ where (γ_k) is a decreasing sequence of positive step-sizes.
- Maximisation step : $\theta_k = \arg\max Q_k(\theta)$

Equivalent description of the kth simulation step

First description :

- Select a subset $\mathcal{I}_k$ of indices :
  Simulate the mini-batch size $r \sim \text{Bin}(n, \alpha)$
  Choose r indices from $\{1, \ldots, n\}$ denoted by $\mathcal{I}_k$
- for i in $\mathcal{I}_k$
  $z^k \sim \Pi_{\mathbf{i}}(z^{k-1}, \cdot; \theta_{k-1})$
  where Π_i is the Metropolis kernel acting only on the i-th coordinate

Equivalent description : for **i** in 1 : **n**

- Sample an indicator $U_{k,i} \sim \text{Bernoulli}(\alpha)$
- Sample $\tilde{\mathbf{z}} \sim \Pi_i(\mathbf{z}_k, \cdot | \theta_{k-1})$
- Set $\mathbf{z}_k = U_{k,i}\tilde{\mathbf{z}} + (1 - U_{k,i})\mathbf{z}_k$

Details of the simulation step

- Simulation step of the **batch** MCMC-SAEM :
  Simulate $\mathbf{z}_k \sim \Pi(\mathbf{z}_{k-1}, \cdot|\theta_{k-1})$, where

$$\Pi = \Pi_n \circ \cdots \circ \Pi_1$$

  and Π_i are Metropolis kernels acting only on the i-th coordinate.
- Simulation step of the **mini-batch** MCMC-SAEM :
  - Denote

$$\Pi_{\alpha,i}(\mathbf{z}, \mathbf{z}'|\theta) = \alpha\Pi_i(\mathbf{z}, (z_1, \ldots, z_i', \ldots, z_n)|\theta) + (1-\alpha)\delta_{\mathbf{z}}(\mathbf{z}').$$

  - Simulate $\mathbf{z}_k \sim \Pi_\alpha(\mathbf{z}_{k-1}, \cdot|\theta_{k-1})$ where

$$\Pi_\alpha = \Pi_{\alpha,n} \circ \cdots \circ \Pi_{\alpha,1}$$

$\Longrightarrow$ The mini-batch MCMC-SAEM algorithm formally belongs to the family of MCMC-SAEM algorithms

Convergence result (Kuhn, Matias, Rebafka (2019)

Theorem

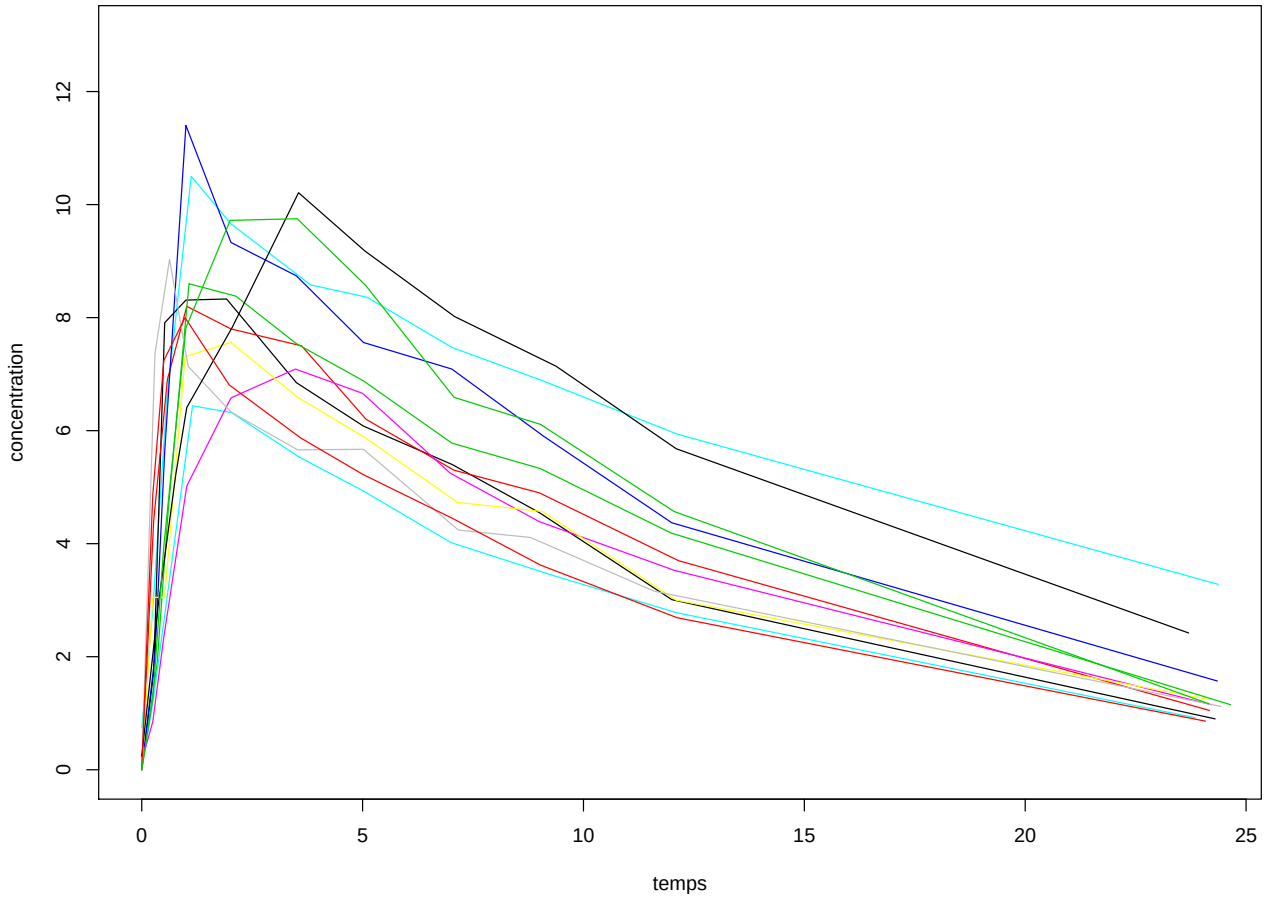
Let $0 < \alpha \leq 1$ *be the mini-batch proportion and* $(\theta_k)_{k\geq 1}$ *a sequence generated by the mini-batch MCMC-SAEM algorithm with corresponding Markov kernel* $\Pi_\alpha(\cdot, \cdot|\theta)$*. Then,*

under the same assumptions as for the batch algorithm,

$$\lim_{k\to\infty} \theta_k \in \{\theta : \nabla \ell(\theta) = 0\},$$

that is, $(\theta_k)_{k\geq 1}$ *converges almost surely towards the set of critical points of the observed likelihood* ℓ *as the number of iterations increases.*

Theophylline concentration along time (Davidian and Giltinian (1995))



12 subjects, same oral dose (mg/kg) times in hours theophylline concentration in mg/L

Nonlinear mixed effects models

classical one-compartment model presented in **?**
For $i = 1, \ldots, n$ and $j = 1, \ldots, J$:

$$Y_{ij} = \frac{d\mathrm{ka}_i}{V_i\mathrm{ka}_i - \mathrm{Cl}_i} \left[\exp(-\mathrm{Cl}_i t_{ij}/V_i) - \exp(-\mathrm{ka}_i t_{ij})\right] + \varepsilon_{ij}$$

with Y_{ij} measure of drug concentration at time t_{ij}
d drug dose
$Z_i = (V_i, \mathrm{ka}_i, \mathrm{Cl}_i)$ with V_i volume of the central compartment, ka_i constant of the drug absorption rate, Cl_i drug's clearance

$$\begin{aligned}
\log V_i &= \log(\mu_V) + \eta_{i,1}, \quad \eta_{i,1} \sim \mathcal{N}(0, \omega_V^2) \\
\log \mathrm{ka}_i &= \log(\mu_{\mathrm{ka}}) + \eta_{i,2}, \quad \eta_{i,2} \sim \mathcal{N}(0, \omega_{\mathrm{ka}}^2) \\
\log \mathrm{Cl}_i &= \log(\mu_{\mathrm{Cl}}) + \eta_{i,3}, \quad \eta_{i,3} \sim \mathcal{N}(0, \omega_{\mathrm{Cl}}^2)
\end{aligned}$$

$\Longrightarrow \theta = (\mu_V, \mu_{\mathrm{ka}}, \mu_{\mathrm{Cl}}, \omega_V^2, \omega_{\mathrm{ka}}^2, \omega_{\mathrm{Cl}}^2, \sigma^2)$.

Theophylline model results

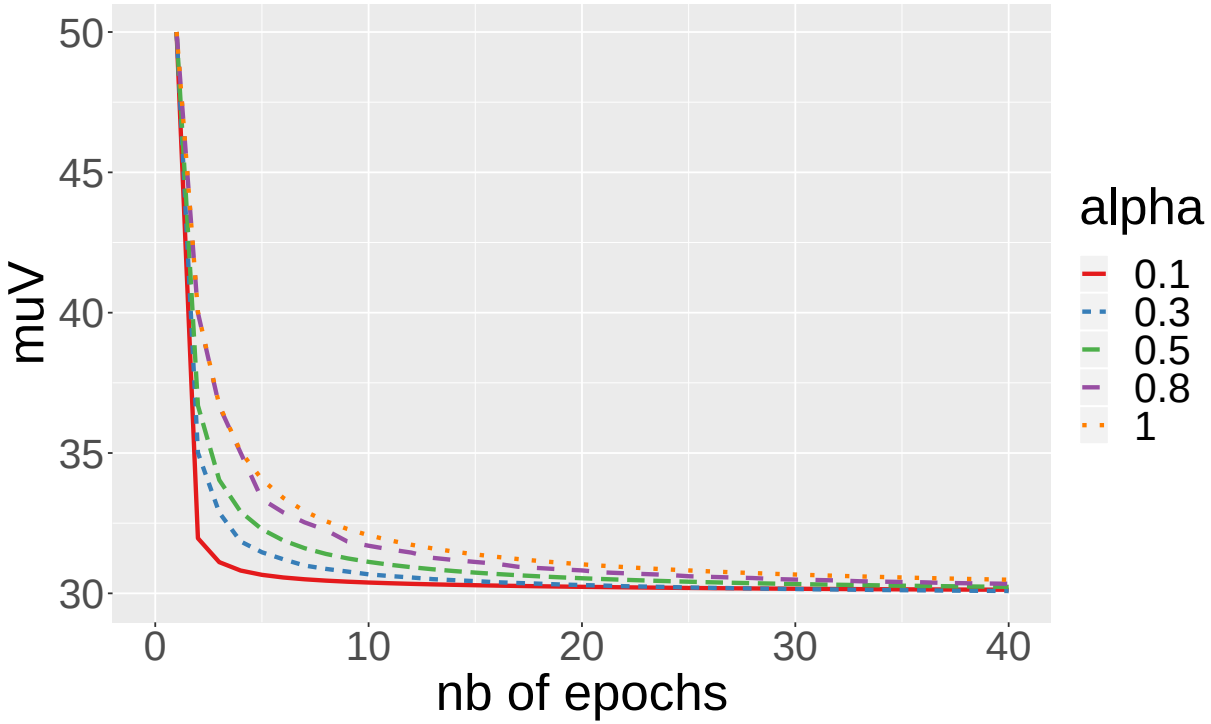


FIGURE – Estimates of the parameter μ_V using mini-batch MCMC-SAEM with $\alpha \in \{0.1, 0.3, 0.5, 0.8, 1\}$ as a function of the number of epochs.

Simulation setting : $n = 1000$, $\mu_V = 30$

Theophylline model results

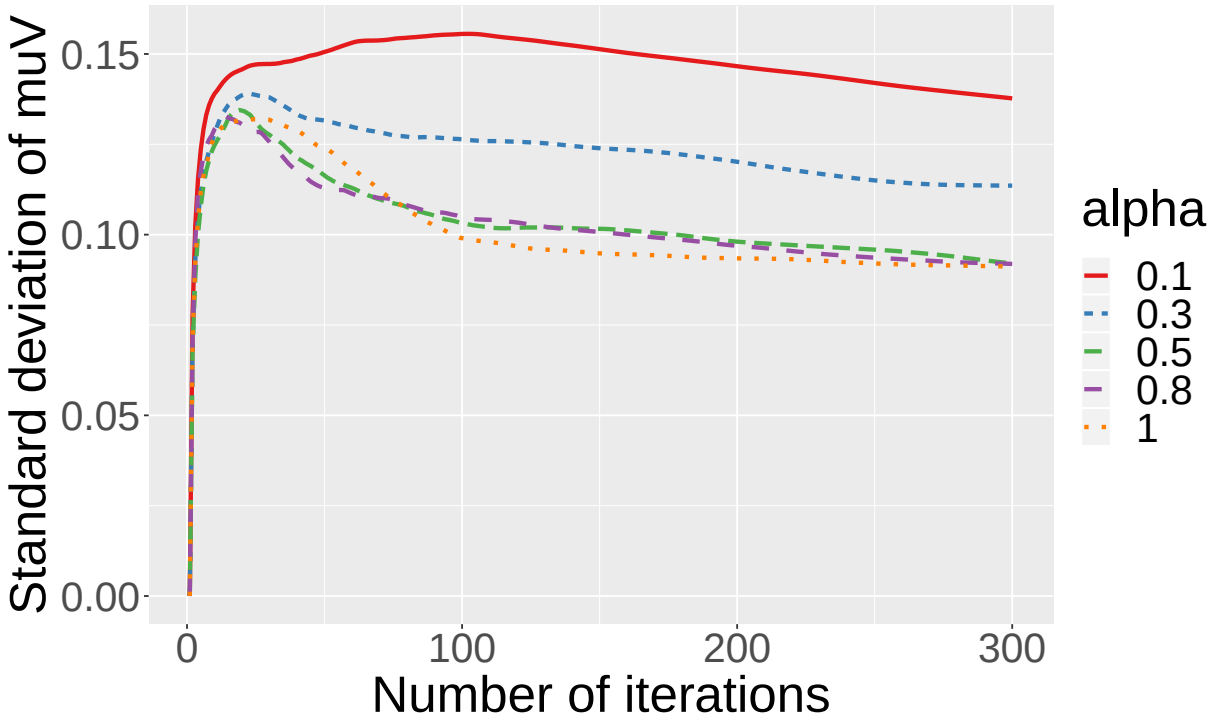


FIGURE – Sample standard deviation of the estimate of parameter μ_V using mini-batch MCMC-SAEM with $\alpha \in \{0.1, 0.3, 0.5, 0.8, 1\}$ as a function of the number of iterations.

Observations of USPS dataset

Deformable model for image analysis (Allassonnière et al (2007))

$$y_i(s) = I_0(x_s - \Phi_i(x_s)) + \sigma\varepsilon_i(s)$$

- y_i image
- I_0 reference image
- Φ_i deformation
- σ noise level and ε_i noise term

Let $(p_k)_{1\leq k\leq k_p}$ be some landmarks on the domain D.
Then for $\xi \in \mathbb{R}^{k_p}$ we define the template by

$$I_0(x) = (K_p\xi)(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\xi(k).$$

Let $(g_k)_{1\leq k\leq k_g}$ be some geometrical landmarks on D.
Then for $z_i \in \mathbb{R}^{dk_g}$ we define the field of deformation by

$$\Phi_i(x) = (K_g z_i)(x) = \sum^{k_g} K_g(x, g_k)(z_i^{(1)}(k), z_i^{(2)}(k)).$$

Deformable model for image analysis

$$y_i(s) = I_0(x_s - \Phi_i(x_s)) + \sigma\varepsilon_i(s)$$

- y_i image
- I_0 reference image
- Φ_i deformation
- σ noise level and ε_i noise term

$$I_0(x) = (K_p\xi)(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\xi(k).$$

$$\Phi_i(x) = (K_g z_i)(x) = \sum_{k=1}^{k_g} K_g(x, g_k)(z_i^{(1)}(k), z_i^{(2)}(k)).$$

Latent variable $z_i \sim \mathcal{N}_{2k_g}(0, \Gamma)$

Model parameters $\theta = (\xi, \Gamma, \sigma^2)$

Image analysis results

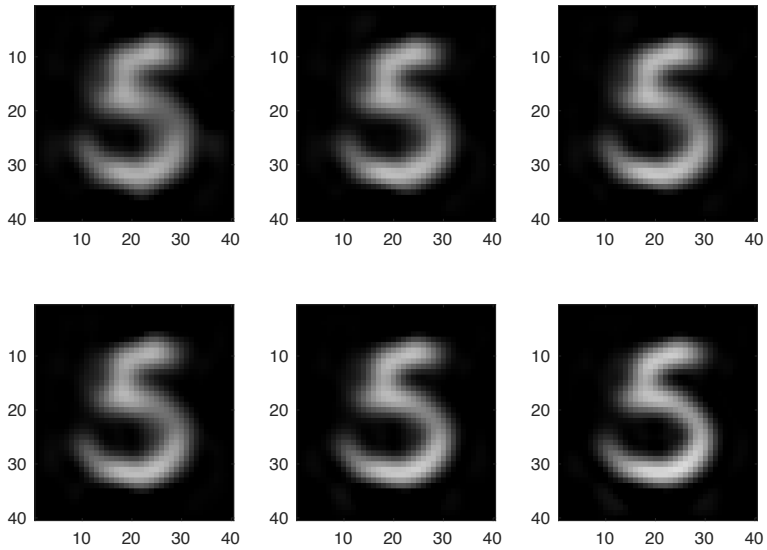


FIGURE – Estimation of the template : first row : using batch MCMC-SAEM ; second row : using mini-batch MCMC-SAEM with $\alpha = 0.1$; columns correspond to 1, 2 and 3 epochs, respectively.

Simulation setting : $n = 20$

Image analysis results

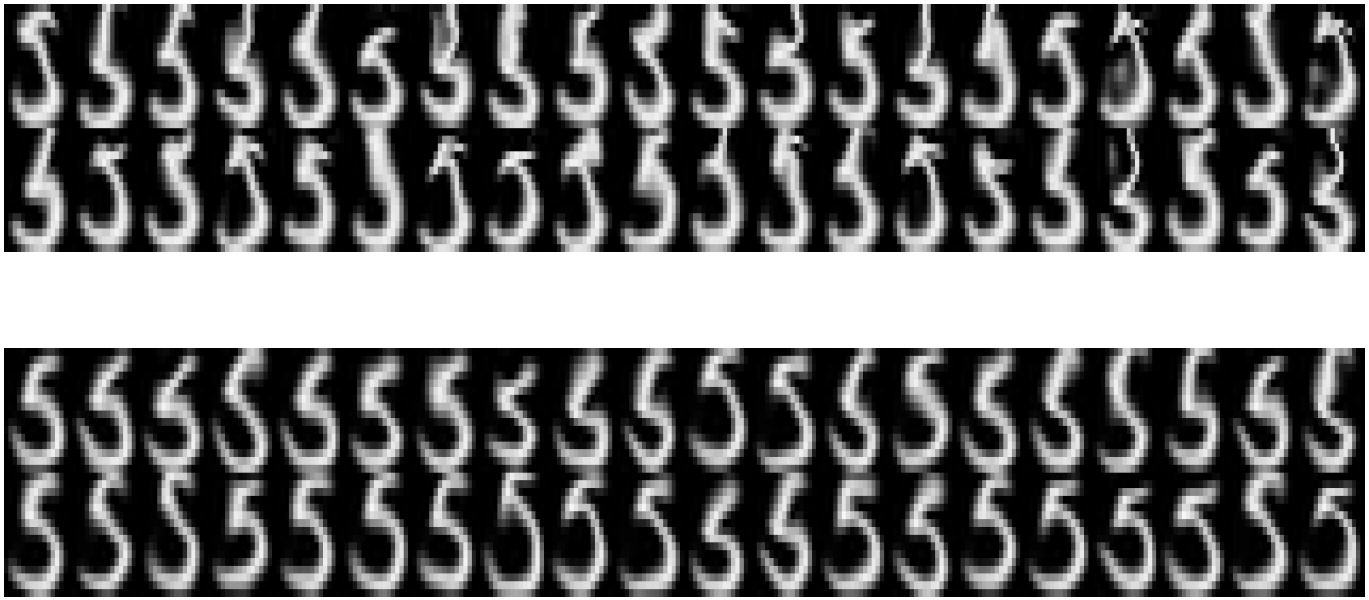
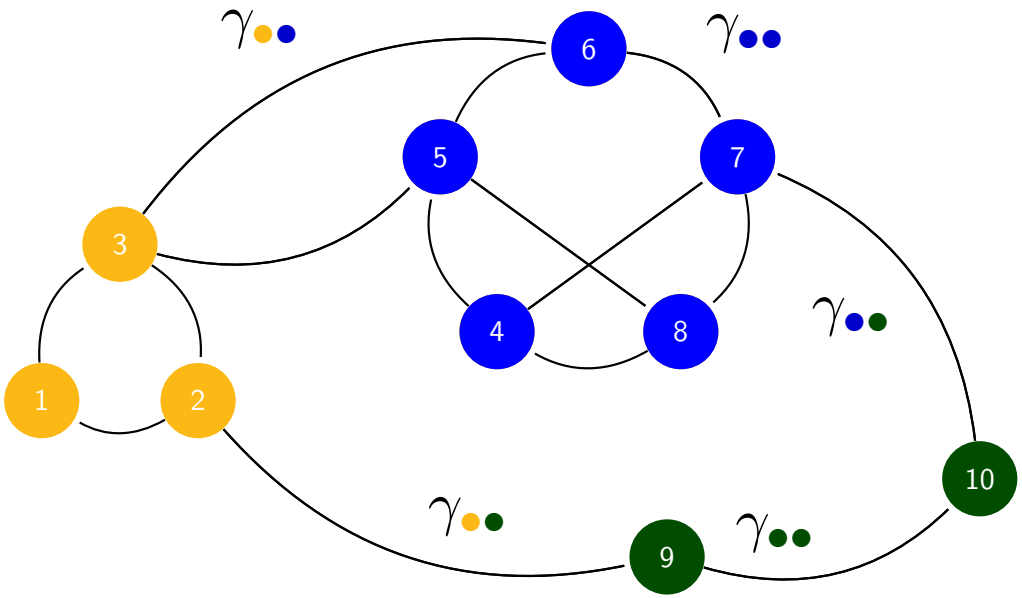


FIGURE – Synthetic images sampled from the model for digit 5 using the parameter estimates obtained with the batch version on 20 images (top) and with the mini-batch version with $\alpha = 0.2$ on 100 images (bottom).

Stochastic block model



$n = 10,\ Z_{5\bullet} = 1,$
$Y_{12} = 1, Y_{15} = 0$

parametric model with $\theta = (\pi, \gamma)$

- K groups (=colors $\bullet\bullet\bullet$). $\{Y_{ij}\}_{1 \le i < j \le n}$ edges
- Not observed $\{Z_i\}_{1 \le i \le n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iK}) \sim \mathcal{M}(1, \pi)$, with $\pi = (\pi_1, \ldots, \pi_K)$ groups proportions.
- Observations : presence/absence of an edge $\{Y_{ij}\}$
- Conditional on $\{Z_i\}$'s, the r.v. Y_{ij} are independent $\mathcal{B}(\gamma_{Z_i Z_j})$.

Model (?)

- Block membership Z_i of node i, i.i.d. with

$$P(Z_i = \bullet) = \pi_\bullet, \qquad 1 \leq \bullet \leq Q, 1 \leq i \leq n.$$

- Elements $Y_{i,j}$ of the adjacency matrix of a directed graph are such that $Y_{i,j}$ are independent conditional on $\mathbf{Z}$ and

$$Y_{i,j}|(Z_i = \bullet, Z_j = \bullet) \sim \text{Bernoulli}(\gamma_{\bullet,\bullet}), \quad 1 \leq i, j \leq n$$

- Observations : adjacency matrix $\mathbf{y} = (y_{i,j})_{1 \leq i,j \leq n}$
- Latent variables : block labels $z_i, 1 \leq i \leq n$.
- Model parameter $\theta = ((\pi_q)_{1 \leq q \leq Q}, (\gamma_{q,\ell})_{1 \leq q,\ell \leq Q})$.

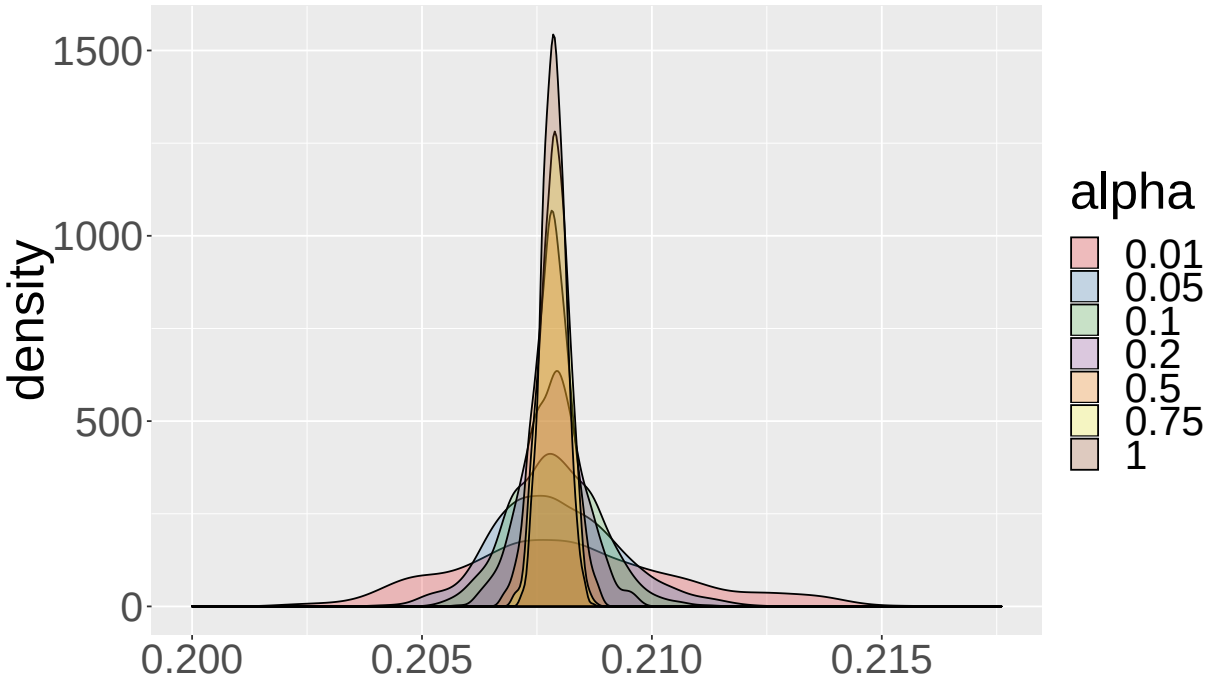
Simulation setting

- $Q = 2$ latent blocks
- Block proportions $\pi_1 = 1 - \pi_2 = 0.6$
- Connectivity matrix

$$(\nu_{q,\ell})_{q,\ell} = \begin{pmatrix} 0.25 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}.$$

- Number of nodes $n = 100$
  $\Longrightarrow$ number of observations $n^2 = 10.000$

Simulation results



Limit distribution of the estimate of $\nu_{2,2} = 0.2$ after 10000 iterations.

Heuristic of asymptotic normality result

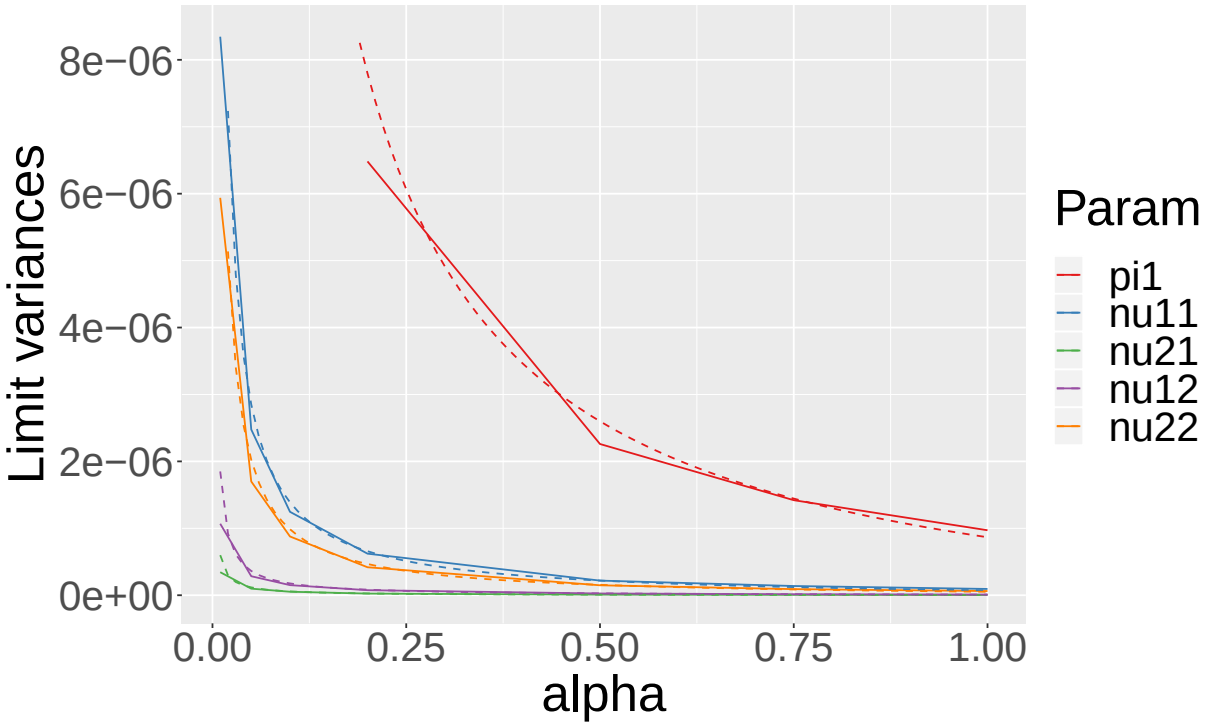
Conjecture :

Under reasonable assumptions, $(\theta_k)_{k\geq 1}$ is asymptotically normal at rate $1/\sqrt{k}$ and the limiting covariance matrix, say V_α, depends on the mini-batch proportion α in the following form

$$V_\alpha = \frac{2-\alpha}{\alpha} V_1,$$

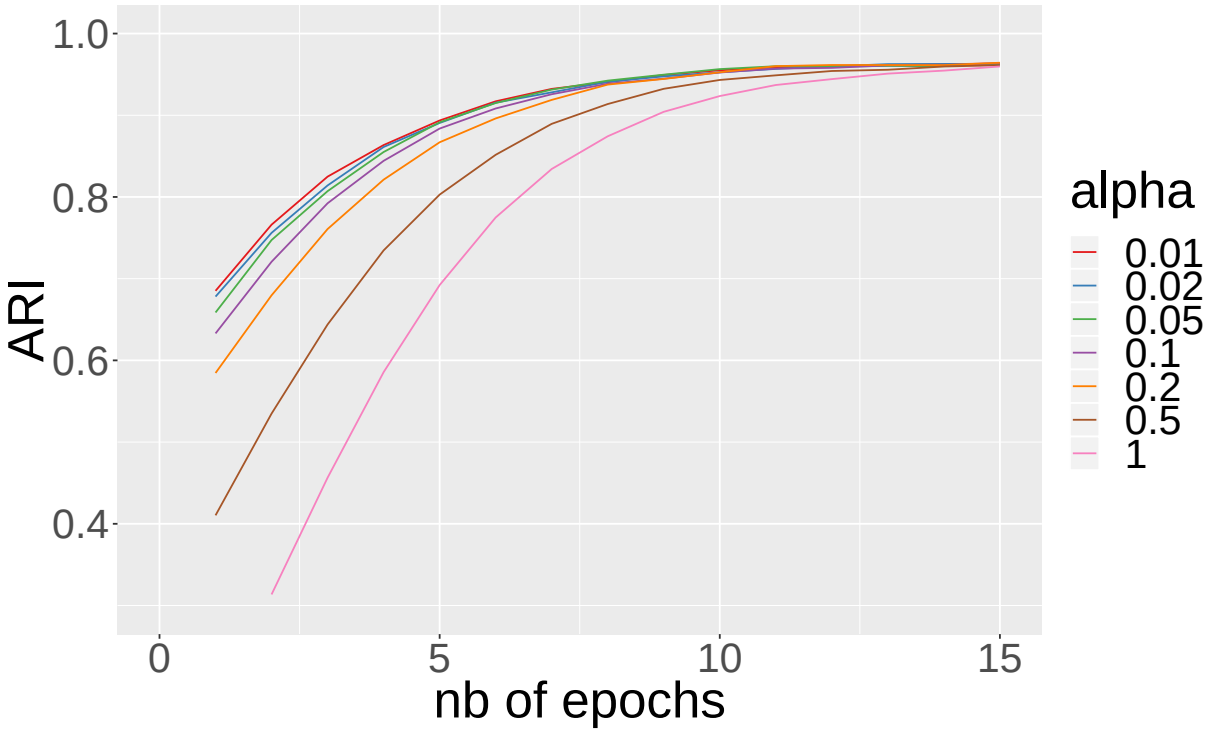
where V_1 denotes the limiting covariance of the batch algorithm.

Simulation results



Sample variances of the parameter estimates after 10 000 iterations as a function of the mini-batch proportion α (solid lines) and adjusted theoretical limit variances (dashed lines).

Simulation set up



Mean ARI obtained by mini-batch MCMC-SAEM algorithms as a function of the number of epochs for $\alpha \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$.

Conclusion and perspectives

Conclusion :

- minibatch SAEM algorithm
- theoretical convergence result
- heuristic for asymptotic normality
- perfom well in practice

Perspectives

- good use of mini-batch sampling in practice
- compare algorithms relying on the same computing time rather than on the same number of epochs
- understand the impact of the mini-batch proportion α and the sample size on the convergence of the algorithm

Bibliographie