Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation

Stéphanie Allassonniere, Estelle Kuhn

Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation*

Stéphanie Allassonnière[??] and Estelle Kuhn[??]

Ecole Polytechnique
Route de Saclay
91128 Palaiseau, FRANCE
e-mail: stephanie.allassonniere@polytechnique.edu

INRA
Domaine de Vilvert
78352 Jouy-en-Josas, FRANCE
e-mail: estelle.kuhn@jouy.inra.fr

Abstract: Estimation in the deformable template model is a big challenge in image analysis. The issue is to estimate an atlas of a population. This atlas contains a template and the corresponding geometrical variability of the observed shapes. The goal is to propose an accurate algorithm with low computational cost and with theoretical guaranties of relevance. This becomes very demanding when dealing with high dimensional data which is particularly the case of medical images. We propose to use an optimized Monte Carlo Markov Chain (MCMC) method into a stochastic Expectation Maximization (EM) algorithm in order to estimate the model parameters by maximizing the likelihood. We present a new Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA) which is used as transition in the MCMC method. We first prove that this new sampler leads to a geometrically uniformly ergodic Markov chain. We prove also that under mild conditions, the estimated parameters converge almost surely and are asymptotically Gaussian distributed. The methodology developed is then tested on handwritten digits and some 2D and 3D medical images for the deformable model estimation. More widely, the proposed algorithm can be used for a large range of models in many field of applications such as pharmacology or genetic.

AMS 2000 subject classifications: Primary 62F10, 60J22; secondary 62F15, 62M40, 62P10.
Keywords and phrases: Maximum likelihood estimation; missing variable; high dimension; stochastic EM algorithm; MCMC; Anisotropic MALA; Deformable template; geometric variability..

1. Introduction

We consider here the deformable template model introduced for Computational Anatomy by Grenander and Miller (1998). This model, which has demonstrated great impact in image analysis, was developed and analyzed later on by many groups (among other Marsland and Twining (2004); Miller et al (2009, 2002); Vercauteren et al (2009)). It offers several major advantages. First, it enables to describe the population of interest by a digital anatomical template. Then, it captures the geometric variability of the population shapes through the modeling of deformations of the template which match it to the observations. Moreover, the metric on the space of deformations is specified in the model as a quantification of the deformation cost. Not only describing the population, this generative model also allows to generate synthetic data using both the template and the geometrical metric of the deformation space which together define the atlas. Nevertheless, the key statistical issue is how to estimate efficiently and accurately these parameters of the model from an observed population of images.

*The authors would like to thank ANR HM-TC and ANR IRMGroup

Several numerical methods have been developed mainly for the estimation of the template image (for example Cootes et al (1995); Joshi et al (2004)). Even if these methods lead to interesting results on some training samples, they suffer from a lack of theoretical properties and are not robust to noisy data. Another important contribution toward the statistical formulation of the issue of template estimation was proposed by Glasbey and Mardia (2001). However interesting this approach is not entirely satisfactory since the deformations are applied to discrete observations requiring some interpolation. Moreover it does not formulate the analysis in terms of a generative model which appears very attractive as mentioned above. To overcome these lacks, Allassonnière et al (2007) have formulated a coherent statistical generative model. For estimating all the model parameters, the template image together with the geometrical metric, the authors proposed a deterministic algorithm based on an approximation of the well-known Expectation Maximization (EM) algorithm (Dempster et al (1977)), where the posterior distribution is replaced by a Dirac measure on its mode (called FAM-EM). However, such an approximation leads to the non-convergence of the estimates highlighted when considering noisy observations.

One solution to face this problem is to consider a convergent stochastic approximation of the EM (SAEM) algorithm which was proposed by Delyon et al (1999). An extension using Monte Carlo Markov Chain (MCMC) methods was developed and studied by Kuhn and Lavielle (2004) and Allassonnière et al (2010b) allowing for wider applications. To apply this extension to the deformable template model, Allassonnière et al (2010b) chose a Metropolis Hastings within Gibbs sampler (also called hybrid Gibbs) as MCMC method since the variables to sample were of large dimension (the usual Metropolis Hastings algorithm providing low acceptation rates). This estimation algorithm has been proved convergent and performs very well on very different kind of data as presented by Allassonnière et al (2010a). Nevertheless, the hybrid Gibbs sampler becomes computationally very expensive when sampling very high dimensional variables. Although it reduces the dimension of the sampling to one which enables to stride easier the target density support, it loops over the sampling variable coordinates, which becomes computationally unusable as soon as the dimension is very large or as the acceptation ratio involves heavy computations. To overcome the problem of computational cost of this estimation algorithm, some authors propose to simplify the statistical model constraining the correlations of the deformations (c.f. Maire et al (2011); Richard et al (2009)). Our purpose in this paper is to propose an efficient and convergent estimation algorithm for the deformable template model in high dimension without any constrains. With regards to the above considerations, the computational cost of the estimation algorithm can be reduced by optimizing the sampling scheme in the MCMC method.

The sampling of high dimensional variables is a well-known difficult challenge. In particular, many authors have proposed to use the Metropolis Adjusted Langevin Algorithm (MALA) (see Roberts and Tweedie (1996) and Stramer and Tweedie (1999a)). This algorithm is a particular random walk Metropolis-Hastings sampler. Starting from the current iterate of the Markov chain, one simulates a candidate with respect to a Gaussian proposal with an expectation equal to the sum of this current iterate and a drift related to the target distribution. The covariance matrix is diagonal and isotropic. This candidate is accepted or rejected with a probability given by the Metropolis Hastings acceptance ratio.

Some modifications have been proposed in particular to optimize the covariance matrix of the proposal in order to better stride the support of the target distribution (see Atchadé (2006); Girolami and Calderhead (2011); Marshall and Roberts (2012); Stramer and Tweedie (1999b)). Atchadé (2006) and Marshall and Roberts (2012) proposed to construct adaptive MALA chains for which they prove the geometric ergodicity of the chain uniformly on any compact subset of its parameters. Unfortunately, this technique does not take the whole advantage of changing the proposal using the target distribution. In particular, the covariance matrix of the proposal is given by a stochastic approximation of the empirical covariance matrix. This choice seems completely relevant as soon as the convergence toward the stationary distribution is reached. However, it does not provide a good guess of the variability during the first iterations of the chain since it is still very dependent on the initialization. This leads to chains that may be numerically trapped. Moreover, this particular algorithm may require a lot of tuning parameters. Although

the theoretical convergence is proved, this algorithm may be very difficult to optimize in practice *into* an estimation process.

Recently, Girolami and Calderhead (2011) proposed the Riemann manifold Langevin algorithm in order to sample from target density in high dimensional setting with strong correlations. This algorithm is also a MALA based one for which the choice of the proposal covariance is guided by the metric of the underlying Riemann manifold. It requires to evaluate the metric, its inverse as well as its derivatives. The proposed well-suited metric is the Fisher-Rao information matrix or its empirical value. However, in the context we are dealing with, the real metric, namely the metric of the space of non-rigid deformations, is not explicit (a particular case is calculated by Micheli et al (2012)) preventing from any use of it. Moreover, if we consider the constant curvature simplification suggested by Girolami and Calderhead (2011), one still needs to invert the metric which may be neither explicit nor computationally tractable. Note that these constrains are common with other application fields such as genetic or pharmacology, where the models are often complex.

For all these reasons, we propose to adapt the MALA algorithm in the spirit of both works of Atchadé (2006) and Girolami and Calderhead (2011) to get an efficient sampler into the stochastic EM algorithm. Therefore, we propose to sample from a proposal distribution which has the same expectation as the MALA but using a full anisotropic covariance matrix based on the anisotropy and correlations of the target distribution (called in the sequel AMALA). The expectation is obtained as the sum of the current iterate plus a drift which is proportional to the gradient of the logarithm of the target distribution. We construct the covariance matrix as a regularization of the Gram matrix of this drift. We prove the geometric ergodicity uniformly on any compact set of the AMALA assuming some regularity conditions on the target distribution. We also prove the almost sure convergence of the parameter estimated sequence generated by the coupling of AMALA and SAEM algorithms (AMALA-SAEM) toward the maximum likelihood estimate under some regularity assumptions on the model. Moreover, we prove a Central Limit Theorem for this sequence under weak conditions on the model. We test our estimation algorithm on the deformable template model for estimating hand-written digit atlases from the USPS database and medical images of corpus callosum (2D) and of dendrite spine excrescences (3D). The proposed estimation method is compared with the results obtained from the FAM-EM algorithm and from the MCMC-SAEM algorithm using different samplers namely the hybrid Gibbs sampler, the MALA and the adaptive MALA proposed by Atchadé (2006) previously introduced. The comparison is also made via classification rates on the USPS database. These experiments demonstrate the good behavior of our method in both the accuracy of the estimation and the low computational cost in high dimension.

The paper is organized as follows. In Section 2, we recall the Bayesian Mixed Effect (BME) template model. In Section 3, we consider the estimation issue in general framework of missing data models. We present our stochastic version of the EM algorithm using the AMALA sampler. The convergence properties are established in Section 4. Section 5 is devoted to the experiments on the BME template estimation. Finally, we give some conclusion in Section 6. The proofs are postponed in Section 7.

2. Description of the Bayesian Mixed Effect (BME) Template model

The deformable template model aims at summarizing a population of images by two quantities. The first one is a mean image which has to represent a relevant shape as one could find in the population. The second quantity represents the variance in the space of shapes. This corresponds to the geometrical variability around the mean shape. In practice, numerical approximations of both quantities are obtained through statistical estimation process. Let us now describe the deformable template model.

We consider the hierarchical Bayesian framework for dense deformable template developed by Allassonnière et al (2007) where each image in a population is assumed to be generated as a noisy and randomly deformed version of the mean image also called template.

The database is composed of n grey level images $(y_i)_{1\leq i\leq n}$ observed on a grid Λ of pixels (or voxels) $\{v_u \in D, u \in \Lambda\}$ included in a continuous domain $D \subset \mathbb{R}^d$, (typically $D = [-1,1]^d$ where d equals 2 or 3). The expected template $I_0 : \mathbb{R}^d \to \mathbb{R}$ takes its values in the continuous domain. Each observation y is assumed to be a discretization on Λ of a random deformation of this template plus an independent noise. Therefore, there exists an unobserved deformation field (also called mapping) $m : \mathbb{R}^d \to \mathbb{R}^d$ such that for $u \in \Lambda$

$$y(u) = I_0(v_u - m(v_u)) + \sigma\epsilon(u) \ ,$$

where $\sigma\epsilon$ denotes the independent additive noise and v_u is the location of pixel (or voxel) u.

Considering the template and the deformations as continuous functions would lead to a dense problem. The dimension is reduced assuming that both elements belong to a subset of fixed Reproducing Kernel Hilbert Spaces (RKHS) V_p and V_g defined by their respective kernels K_p and K_g. More precisely, let $(r_{p,j})_{1\leq j\leq k_p}$ -respectively $(r_{g,j})_{1\leq j\leq k_g}$- be some fixed control points in the domain D: there exist $\alpha \in \mathbb{R}^{k_p}$ -resp. $z \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$- such that for all v in D:

$$I_\alpha(v) = (\mathbf{K_p}\alpha)(v) = \sum_{j=1}^{k_p} K_p(v, r_{p,j})\alpha^j \quad \text{and}$$

$$m_z(v) = (\mathbf{K_g}z)(v) = \sum_{j=1}^{k_g} K_g(v, r_{g,j})z^j \ .$$

For clarity, write $\mathbf{y} = (y_i)_{1\leq i\leq n}$ for the $n-$tuple of observations and $\mathbf{z} = (z_i)_{1\leq i\leq n}$ for the $n-$tuple of unobserved variables defining the deformations. The statistical model on the observations is chosen as follows:

$$\begin{cases} \mathbf{z} \sim \otimes_{i=1}^n \mathcal{N}_{dk_g}(0, \Gamma_g) \mid \Gamma_g \ , \\ \\ \mathbf{y} \sim \otimes_{i=1}^n \mathcal{N}_{|\Lambda|}(m_{z_i} I_\alpha, \sigma^2 Id) \mid \mathbf{z}, \alpha, \sigma^2 \ , \end{cases} \tag{2.1}$$

where $\otimes$ denotes the product of independent variables and $mI_\alpha(u) = I_\alpha(v_u - m(v_u))$, for u in Λ. The parameters of interest are α (the template), σ^2 (the noise variance) and Γ_g (the deformation covariance matrix). We assume that $\theta = (\alpha, \sigma^2, \Gamma_g)$ belongs to the parameter space Θ:

$$\Theta \triangleq \{ \ \theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p}, \ |\alpha| < R, \ \sigma > 0, \ \Gamma_g \in \mathrm{Sym}^+_{dk_g,*}(\mathbb{R}) \ \} \ , \tag{2.2}$$

where $\mathrm{Sym}^+_{dk_g,*}(\mathbb{R})$ is the cone of real positive $dk_g \times dk_g$ definite symmetric matrices, R an arbitrary positive constant and d is the space dimension (typically 2 or 3 for images).

Since we aim at dealing with small size samples and high dimensional parameters, we work in the Bayesian framework and we introduce priors on the parameters. In addition of guiding the estimation it regularizes the estimation as shown by Allassonnière et al (2007). The priors are all independent: $\theta = (\alpha, \sigma^2, \Gamma_g) \sim \nu_p \otimes \nu_g$ where

$$\begin{cases} \nu_p(d\alpha, d\sigma^2) \propto \exp\left(-\frac{1}{2}(\alpha - \mu_p)^T(\Sigma_p)^{-1}(\alpha - \mu_p)\right) \times \left(\exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right)\frac{1}{\sqrt{\sigma^2}}\right)^{a_p} d\sigma^2 d\alpha, \ a_p \geq 3 \ , \\ \nu_g(d\Gamma_g) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g \rangle_F / 2)\frac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g, a_g \geq 4k_g + 1 \ . \end{cases} \tag{2.3}$$

For two matrices A, B we define the Frobenius inner product by $\langle A, B \rangle_F \triangleq tr(A^T B)$.

Parameter estimation for this model is then performed by Maximum A Posteriori (MAP) :

$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}}\, q(\theta|\mathbf{y}) \ ,$$

where $q(\theta|\mathbf{y})$ is the posterior density of θ conditional on $\mathbf{y}$. The existence and consistence of the MAP estimator for the BME template model has been proved by Allassonnière et al (2007).

Note that this model belongs to a more general class called the mixed effect models. The fixed effects are the parameters θ and the random effects are the deformation coefficients $\mathbf{z}$. The estimation issue in this class is treated in the same way as the likelihood maximization problem in the more general framework of incomplete-data models. Therefore, the next section will be presented in this general setting in which the proposed algorithm applies.

3. Maximum likelihood estimation

3.1. Maximum likelihood estimation for incomplete data setting

We consider in this section the standard incomplete-data (or partially-observed-data) setting and recall the usual notation. We denote by $y \in \mathbb{R}^q$ the observed data and by $z \in \mathbb{R}^l$ the missing data, so that we obtain the complete data $(y, z) \in \mathbb{R}^{q+l}$ for some $q \in \mathbb{N}^*$ and $l \in \mathbb{N}^*$. We consider these data as random vectors. Let μ' be a σ-finite measure on $\mathbb{R}^{q+l}$ and μ the restriction of μ' to $\mathbb{R}^l$ generated by the projection $(y, z) \mapsto z$. We assume that the probability density function of the random vector (y, z) belongs to $\mathcal{P} = \{f(y, z; \theta), \theta \in \Theta\}$, a family of parametric probability density functions on $\mathbb{R}^{q+l}$ w.r.t. μ', where $\Theta \subset \mathbb{R}^p$. Therefore, the observed likelihood (i.e. the incomplete-data likelihood) is defined for some $\theta \in \Theta$ by:

$$g(y; \theta) \triangleq \int f(y, z; \theta)\mu(dz). \tag{3.1}$$

Our purpose is to find the maximum likelihood estimate that is the value $\hat{\theta}_g$ in Θ that maximizes the observed likelihood g given a sample of observations. However, this maximization can often not be done analytically because of the integration involved in (3.1). A powerful tool which enables to compute this maximization in such a setting is the Expectation Maximization (EM) algorithm (see Dempster et al (1977)). It is an iterative procedure which consists of two steps. First, the so-called E-step computes the conditional expectation of the complete log-likelihood using the current parameter value. Second, the M-step achieves the update of the parameter by maximizing this expectation over Θ. However, the computation of this expectation is often intractable analytically. Therefore, alternative procedures have been proposed. We are particularly interested in the Stochastic Approximation EM (SAEM) algorithm (see Delyon et al (1999)) because of its theoretical convergence property and its small computation time. In this stochastic algorithm, the usual E-step is replaced by two steps, the first one corresponding to the simulation of realizations of the missing data, the second one to the computation of a stochastic approximation of the complete log-likelihood using these simulated values. It can be shown that under weak regularity conditions the sequence generated by this algorithm converges almost surely toward a local maximum of the observed likelihood (see Delyon et al (1999)).

Nevertheless the simulation step requires some attention. In the SAEM algorithm the simulated values of the missing data have to be drawn from the posterior distribution defined by:

$$p(z|y; \theta) \triangleq \begin{cases} f(y, z; \theta)/g(y; \theta) & \text{if } g(y; \theta) \neq 0 \\ 0 & \text{if } g(y; \theta) = 0\,. \end{cases}$$

When not possible, the extension using MCMC method (Allassonnière et al (2010b); Kuhn and Lavielle (2004)) allows to apply the SAEM algorithm using simulations obtained from some transition probability of an ergodic Markov Chain having the targeted posterior distribution as stationary distribution. Methods like Metropolis Hastings algorithm or Gibbs sampler are useful to perform this assignment. However, this becomes very challenging in high dimensional setting. Indeed, when the MCMC procedure has to explore a space of high dimension, its convergence may occur in practice only after a possibly infinite time. Thus, it is necessary to optimize this MCMC procedure. This is what we will propose in the following paragraph.

3.2. Description of the sampling method: Anisotropic Metropolis Adjusted Langevin Algorithm

We propose an anisotropic version of the well-known Metropolis Adjusted Langevin Algorithm (MALA). So let us first recall the steps of the Metropolis Adjusted Langevin Algorithm (MALA). Let $\mathcal{X}$ be an open subset of $\mathbb{R}^l$, the $l-$dimensional Euclidean space equipped with its Borel $\sigma-$algebra $\mathcal{B}$. Let us denote π the probability density function (pdf) with respect to the Lebesgue measure on $\mathcal{X}$ of the target distribution. We assume that π is positive continuously differentiable. At each iteration k of this algorithm, a candidate X_c is simulated with respect to the Gaussian distribution with expectation $X_k + \frac{\sigma^2}{2} D(X_k)$ and covariance $\sigma^2 Id_l$ where X_k is the current value,

$$D(x) = \frac{b}{\max(b, |\nabla \log \pi(x)|)} \nabla \log \pi(x)\,, \tag{3.2}$$

Id_l is the identity matrix in $\mathbb{R}^l$ and $b > 0$ a fixed truncated threshold. In the following, we denote $q_{MALA}(x, \cdot)$ the pdf of this Gaussian candidate distribution starting from x. Given this candidate, the next value of the Markov chain is updated using an acceptance ratio $\alpha_{MALA}(X_k, X_c)$ as follows: $X_{k+1} = X_c$ with probability $\alpha_{MALA}(X_k, X_c) = \min\left(1, \frac{\pi(X_c) q_{MALA}(X_c, X_k)}{q_{MALA}(X_k, X_c)\pi(X_k)}\right)$ and $X_{k+1} = X_k$ with probability $1 - \alpha_{MALA}(X_k, X_c)$. This provides a transition kernel Π_{MALA} of this form: for any Borel set $A \in \mathcal{B}$

$$\Pi_{MALA}(x, A) = \int_A \alpha_{MALA}(x, z) q_{MALA}(x, z) dz + \mathbb{1}_A(x) \int_{\mathcal{X}} (1 - \alpha_{MALA}(x, z)) q_{MALA}(x, z) dz\,. \tag{3.3}$$

Note that the truncation of the drift was already suggested by Gilks et al (1996) to provide more stability.

The Gaussian proposal of the MALA algorithm is optimized with respect to its expectation guided by the Langevin diffusion. One step further is to optimize also its covariance matrix. A first work in this direction was proposed by Atchadé (2006). The covariance matrix of the proposal is given by a projection of a stochastic approximation of the empirical covariance matrix. It produces an adaptive Markov chain. This process involves some additional tuning parameters which have to be calibrated. Since our goal is to use this sampler in an estimation algorithm, the sampler has at each iteration a different target distribution (depending on the current estimate of the parameter). Therefore, the optimal tuning parameter may be different along the iterations of the estimation process. Although we agree with the idea of using adaptive chain, we prefer taking the advantage of the dynamic of the estimation algorithm. On the other side, an intrinsic solution has been proposed by Girolami and Calderhead (2011) where the covariance matrix is given by the metric of the Riemann manifold of the variable to sample. Unfortunately, this metric may not be accessible and its empirical approximation not easy to compute. This is particularly the case in the BME template model.

For these reasons, we propose a sampler in the spirit of Atchadé (2006), Marshall and Roberts (2012) or Girolami and Calderhead (2011) however not providing an adaptive chain as motivated above. The adaption comes from the dependency of the target distribution with respect to the parameters of the model which are updated along the estimation algorithm. The proposal is still a Gaussian distribution but both the drift and the covariance matrix depend on the gradient of the target distribution. At the k^{th} iteration, we are provided with X_k. The candidate is sampled from the Gaussian distribution with expectation $X_k + \delta D(X_k)$ and covariance matrix $\delta\Sigma(X_k)$ denoted in the sequel
$\mathcal{N}\left(X_k + \delta D(X_k), \delta\Sigma(X_k)\right)$ where $\Sigma(x)$ is given as :

$$\Sigma(x) = \varepsilon Id_l + D(x) D(x)^T\,, \tag{3.4}$$

with D defined in Eq (3.2) and $\varepsilon > 0$ is a small regularization parameter. Note that the threshold parameter b leads to a symmetric positive definite covariance matrix with bounded non zero eigen values. We introduced the gradient of $\log \pi$ into the covariance matrix to provide an anisotropic covariance matrix depending on the amplitude of the drift at the current value. When the drift is large, the candidate is likely to be far from the current value. This large step may not be of the right amplitude and a large variance will enable more flexibility. On the other hand, when the drift is small in a particular direction, it means that the current value is close to an admissible state for the Markov chain. Therefore, the candidate should neither be allowed to move too far nor with a large variance. This is taken into account here by introducing the drift into the covariance matrix.

We denote by q_c the pdf of this proposal distribution. The transition kernel becomes:

$$\Pi(x, A) = \int_A \alpha(x, z) q_c(x, z) dz + \mathbb{1}_A(x) \int_{\mathcal{X}} (1 - \alpha(x, z)) q_c(x, z) dz \,, \tag{3.5}$$

where $\alpha(X_k, X_c) = \min\left(1, \frac{\pi(X_c) q_c(X_c, X_k)}{q_c(X_k, X_c)\pi(X_k)}\right)$.

3.3. Description of the stochastic estimation algorithm

Back to the stochastic estimation algorithm, the target distribution of the sampler is the posterior distribution $p(\cdot|y, \theta)$.

The four steps of the proposed AMALA-SAEM algorithm are detailed in this subsection : simulation, stochastic approximation, truncation on random boundaries and maximization steps. At each iteration k of the algorithm, simulated values of the missing data are drawn from the transition probability of the AMALA algorithm described in Section 3.2 with the current value of the parameters. Then, a stochastic approximation of the complete log-likelihood is computed using these simulated values for the missing data and is truncated using random boundaries. Finally, the parameters are updated by maximizing this quantity over Θ.

We consider here only parametric models $\mathcal{P}$ which belong to the set of curved exponential family, this means that the complete likelihood $f(y, z; \theta)$ can be written as:

$$f(y, z; \theta) = \exp\left[-\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right] ,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product, the sufficient statistics S is a function on $\mathbb{R}^l$, taking its values in a subset $\mathcal{S}$ of $\mathbb{R}^m$ and ψ, ϕ are two functions on Θ (note that S, ϕ and ψ may depend also on y, but we omit this dependency for simplicity). This condition is usual in the framework of EM algorithm applications and it is fulfilled by large range of models even complex ones as the BME template model. Therefore the stochastic approximation can be done either on the sufficient statistics S of the model or on the complete log-likelihood.

Concerning the truncation procedure, we introduce a sequence of increasing compact subsets of $\mathcal{S}$ denoted by $(\mathcal{K}_q)_{q\geq 0}$ such that $\cup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$. Let also $\boldsymbol{\varepsilon} = (\varepsilon_q)_{q\geq 0}$ be a monotone non-increasing sequence of positive numbers and K a compact subset of $\mathbb{R}^l$. At iteration k we simulate a value $\bar{z}$ for the missing data from the Anisotropic Metropolis Adjusted Langevin Algorithm using the current value of the parameter θ_{k-1}. We compute the associated stochastic approximation of the sufficient statistics of the model $\bar{s}$. If it does not wander outside the current compact set $\mathcal{K}_k$ and if it is not too far from its previous value s_{k-1}, we keep the possible proposed values for (z_k, s_k). As soon as one of these conditions is not fulfilled, we reinitialize the sequences of z and s using a projection (for more details see Andrieu et al (2005)) and we increase the size of the compact set used for the truncation.

Concerning the maximization step, we denote by L the function defined on $\mathcal{S} \times \Theta$ taking values in $\mathbb{R}$ equals for all (s, θ) to $L(s, \theta) = -\psi(\theta) + \langle s, \phi(\theta)\rangle$. We assume that there exists a function $\hat{\theta}$ defined on $\mathcal{S}$ taking values in Θ such that

$$\forall \theta \in \Theta \; \forall s \in \mathcal{S} \; L(s, \hat{\theta}(s)) \geq L(s, \theta).$$

Finally we update the parameter using the value of the function $\hat{\theta}$ evaluated in s_k.

The complete algorithm is summarized in Algorithm 1. It only involves three parameters: b the threshold for the gradient which appears in the mean as well as in the covariance matrix, δ the scale on this gradient and ε a small regularization parameter to ensure a positive definite covariance matrix. The scale δ can be easily optimized looking at the data we are dealing with to adapt to the range of the drift. The value of the threshold b is in practice never reached. The practical choices for the sequence of positive step sizes used in the stochastic approximation step and the tuning parameters will be detailed in the section devoted to the experiments.

Algorithm 1 AMALA within SAEM

for all $k = 1 : k_{end}$ **do**

`Sample` z_c with respect to $\mathcal{N}(z_{k-1} + \delta D(z_{k-1}, \theta_{k-1}), \delta\Sigma(z_{k-1}, \theta_{k-1}))$ whose pdf is denoted $q_c(z_{k-1}, .; \theta_{k-1})$ where

$$\begin{cases} D(z_{k-1}, \theta_{k-1}) & = \frac{b}{\max(b, |\nabla \log p(z_{k-1}|y;\theta_{k-1})|)} \times \nabla \log p(z_{k-1}|y;\theta_{k-1}) \\ \Sigma(z_{k-1}, \theta_{k-1}) & = \varepsilon Id_l + D(z_{k-1}, \theta_{k-1}) D(z_{k-1}, \theta_{k-1})^T. \end{cases}$$

`Compute` the acceptance ratio

$$\alpha(z_{k-1}, z_c, \theta_{k-1}) = \min\left(1, \frac{p(z_c|y;\theta_{k-1}) q_c(z_c, z_{k-1}; \theta_{k-1})}{q_c(z_{k-1}, z_c; \theta_{k-1}) p(z_{k-1}|y;\theta_{k-1})}\right) \tag{3.6}$$

`Sample` $\bar{z} = z_c$ with probability $\alpha(z_{k-1}, z_c, \theta_{k-1})$ and $\bar{z} = z_{k-1}$ with probability $1 - \alpha(z_{k-1}, z_c, \theta_{k-1})$

`Do the stochastic approximation`

$$\bar{s} = s_{k-1} + \gamma_k \left(S(\bar{z}) - s_{k-1}\right),$$

where $(\gamma_k)_k$ is a sequence of positive step sizes.

if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ and $\|\bar{s} - s_{k-1}\| \leq \varepsilon_{\zeta_{k-1}}$ **then**

Set $(z_k, s_k) = (\bar{z}, \bar{s})$ and $\kappa_k = \kappa_{k-1}$, $\nu_k = \nu_{k-1} + 1$, $\zeta_k = \zeta_{k-1} + 1$

else

set $(z_k, s_k) = (\tilde{z}, \tilde{s}) \in \mathrm{K} \times \mathcal{K}_0$ and $\kappa_k = \kappa_{k-1} + 1$, $\nu_k = 0$, $\zeta_k = \zeta_{k-1} + \Psi(\nu_{k-1})$

where $\Psi : \mathbb{N} \to \mathbb{Z}$ is a function such that $\Psi(k) > -k$ for any k

and $(\tilde{z}, \tilde{s})$ is chosen arbitrarily.

end if

`Update` the parameter

$$\theta_k = \hat{\theta}(s_k).$$

end for

4. Theoretical Properties

4.1. Geometric ergodicity of the AMALA

The AMALA algorithm described in Subsection 3.2 creates a geometric ergodic Markov chain uniformly on any compact subset. Atchadé (2006) proved the geometric ergodicity for a bounded drift and a projection of the empirical covariance matrix. In our case, the same results hold using the previously detailed algorithm where the main difference stands in the anisotropic covariance of the proposal involving the gradient of $\log \pi$. An assumption is required on the stationary distribution namely the so-called super-exponential property given by:

(B1) The density π is positive with continuous first derivative such that:

$$\lim_{|x| \to \infty} n(x).\nabla \log \pi(x) = -\infty \tag{4.1}$$

and

$$\limsup_{|x| \to \infty} n(x).m(x) < 0 \tag{4.2}$$

where ∇ is the gradient operator in $\mathbb{R}^l$, $n(x) = \frac{x}{|x|}$ is the unit vector pointing in the direction of x and $m(x) = \frac{\nabla\pi(x)}{|\nabla\pi(x)|}$ is the unit vector in the direction of the gradient of the stationary distribution at point x.

Proposition 1. *Assume (**B1**). Let $0 < \beta < 1$ and $V(x) = c\pi(x)^{-\beta}$ where c is such that $\forall\ x \in \mathcal{X},\ V(x) \geq 1$. There exist a set $\mathcal{C} \subseteq \mathcal{X}$, a probability measure ν such that $\nu(\mathcal{C}) > 0$ and there exist constants $\lambda \in]0,1[,\ b \in [0,\infty[$ and $\varepsilon \in]0,1]$ such that*

$$\Pi(x,A) \geq \varepsilon\nu(A)\mathbb{1}_{\mathcal{C}}(x) \quad \forall x \in \mathcal{X} \quad \forall A \in \mathcal{B}\,, \tag{4.3}$$

$$\Pi V(x) \leq \lambda V(x) + b\mathbb{1}_{\mathcal{C}}(x)\,. \tag{4.4}$$

The first equation defines $\mathtt{C}$ as a small set for the transition kernel Π. Note that both ε and ν can depend on $\mathtt{C}$. The second inequality is a drift condition which states that the transition kernel tends to bring back elements into the small set. As a consequence of these well known drift conditions, the transition kernel Π is V-uniformly ergodic. That is to say: there exist $0 < \rho < 1$ and $0 < c < \infty$ such that for all $n \in \mathbb{N}^*$ and f such that $\|f\|_V = \sup_{x\in\mathcal{X}} \frac{\|f(x)\|}{V(x)} < \infty$:

$$\|\Pi^n f(.) - \pi f\|_V \leq c\rho^n \|f\|_V\,. \tag{4.5}$$

The proof of Proposition 1 is given in Appendix.

Remark 1. *The same property holds for any power p of the function V satisfying $0 < p\beta < 1$. Indeed, the proof follows the same way as it can be seen in Section Appendix 7. This is a property that will appear useful in the sequel to prove some properties of the estimation algorithm.*

4.2. Convergence property of the general stochastic approximation using AMALA

We first prove a general theorem on the convergence of the stochastic approximation using the Anisotropic Metropolis Adjusted Langevin Algorithm. Then, we apply this result to prove the convergence of the estimated sequence generated by the algorithm AMALA-SAEM.

Let $\mathcal{S}$ be a subset of $\mathbb{R}^m$ for some positive integer m. Let $\mathcal{X}$ be a measurable subspace of $\mathbb{R}^l$ for some positive integer l. For all $s \in \mathcal{S}$, let $H_s :\ \mathcal{X} \to \mathcal{S}$ be a measurable function. Let $\boldsymbol{\gamma} = (\gamma_k)_k$ be a sequence of positive step sizes. Let $(\pi_s)_{s\in\mathcal{S}}$ be a family of positive continuously differentiable probability density functions with respect to the Lebesgue measure on $\mathcal{X}$. For any $s \in \mathcal{S}$, denote by Π_s the transition kernel corresponding to the AMALA procedure described in Section 3.2 with stationary distribution π_s.

Define now the stochastic approximation sequence $(s_k)_k$ as follows :

$$\begin{cases} s_k = s_{k-1} + \gamma_{k-1} H_{s_{k-1}}(z_k) & \text{with} \quad z_k \sim \Pi_{s_{k-1}}(z_{k-1}, \cdot)\,, \text{ if } s_{k-1} \in \mathcal{S} \\ s_k = s_c & \text{with} \quad z_k = z_c\,, \text{ if } s_{k-1} \notin \mathcal{S}\,, \end{cases} \tag{4.6}$$

where $s_c \notin \mathcal{S}$, $z_c \notin \mathcal{X}$. Denote by Q_γ the resulting transition which generates $((z_k, s_k))_k$. We consider the natural filtration of the *non-homogeneous* chain $((z_k, s_k))_k$ and denote respectively by $\mathbb{P}^{\boldsymbol{\gamma}}_{z,s}$ and $\mathbb{E}^{\boldsymbol{\gamma}}_{z,s}$ the probability measure and the corresponding expectation generated by this Markov chain starting at $(z_0, s_0) \in \mathcal{X} \times \mathcal{S}$ and using the sequence $\boldsymbol{\gamma}$.

For any $s \in \mathcal{S}$ such that the function H_s is integrable with respect to π_s, we denote by h the mean field associated with our stochastic approximation so that :

$$h(s) = \int H_s(z)\pi_s(z)dz\,.$$

Using the notation introduced in Subsection 3.3, let $\Phi : \mathcal{X} \times \mathcal{S} \to \mathrm{K} \times \mathcal{K}_0$ be a measurable function and $\Psi :\ \mathbb{N} \to \mathbb{Z}$ be a function such that $\Psi(k) > -k$ for any k. We define the *homogeneous* Markov chain

$$(Z_k = (z_k, s_k, \kappa_k, \zeta_k, \nu_k))_k \tag{4.7}$$

on $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{S} \times \mathbb{N}^3$ with the following transition at iteration k :

- If $\nu_{k-1} = 0$ then draw $(z_k, s_k) \sim Q_{\gamma_{\zeta_{k-1}}}(\Phi(z_{k-1}, s_{k-1}), \cdot)$; otherwise draw $(z_k, s_k) \sim Q_{\gamma_{\zeta_{k-1}}}((z_{k-1}, s_{k-1}), \cdot)$;
- If $\|s_k - s_{k-1}\| \leq \varepsilon_{\zeta_{k-1}}$ and $s_k \in \mathcal{K}_{\kappa_{k-1}}$ then set $\kappa_k = \kappa_{k-1}$, $\zeta_k = \zeta_{k-1} + 1$ and $\nu_k = \nu_{k-1} + 1$; otherwise set $\kappa_k = \kappa_{k-1} + 1$, $\zeta_k = \zeta_{k-1} + \Psi(\nu_{k-1})$ and $\nu_k = 0$.

We consider the following assumptions, generalized from Andrieu et al (2005) and Allassonnière et al (2010b).

(A1') $\mathcal{S}$ is an open subset of $\mathbb{R}^m$, $h : \mathcal{S} \to \mathbb{R}^m$ is continuous and there exists a continuously differentiable function $w : \mathcal{S} \to [0, \infty[$ with the following properties:

(i) There exists an $M_0 > 0$ such that

$$\mathcal{L} \triangleq \{s \in \mathcal{S}, \langle \nabla w(s), h(s) \rangle = 0\} \subset \{s \in \mathcal{S},\ w(s) < M_0\}.$$

(ii) There exists a closed convex set $\mathcal{S}_a \subset \mathcal{S}$ for which $s \to s + \rho H_s(z) \in \mathcal{S}_a$ for any $\rho \in [0, 1]$ and $(z, s) \in \mathcal{X} \times \mathcal{S}_a$ ($\mathcal{S}_a$ is absorbing) and such that for any $M_1 \in]M_0, \infty]$, the set $\mathcal{W}_{M_1} \cap \mathcal{S}_a$ is a compact set of $\mathcal{S}$ where $\mathcal{W}_{M_1} \triangleq \{s \in \mathcal{S},\ w(s) \leq M_1\}$.

(iii) For any $s \in \mathcal{S} \backslash \mathcal{L}$ $\langle \nabla w(s), h(s) \rangle < 0$.

(iv) The closure of $w(\mathcal{L})$ has an empty interior.

(A2') For any $s \in \mathcal{S}$, $H_s : \mathcal{X} \to \mathcal{S}$ is measurable and $\int \|H_s(z)\| \pi_s(z) dz < \infty$.

(A3") There exist a function $V : \mathcal{X} \to [1, \infty]$ such that $\{z \in \mathcal{X}, V(z) < \infty\} \neq \emptyset$, constants $a \in]0, 1]$, $p \geq 2$ and $q \geq p$ such that for any compact subset $\mathcal{K} \subset \mathcal{S}$,

(i)

$$\sup_{s \in \mathcal{K}} \|H_s\|_V < \infty, \tag{4.8}$$

$$\sup_{s \in \mathcal{K}} (\|g_s\|_V + \|\Pi_s g_s\|_V) < \infty, \tag{4.9}$$

$$\sup_{s, s' \in \mathcal{K}} \|s - s'\|^{-a} \{\|g_s - g_{s'}\|_{V^q} + \tag{4.10}$$

$$\|\Pi_s g_s - \Pi_{s'} g_{s'}\|_{V^q}\} < \infty, \tag{4.11}$$

where for anyÂ $s \in \mathcal{S}$ a solution of the Poisson equation $g - \Pi_s g = H_s - \pi_s(H_s)$ is denoted by g_s.

(ii) For any sequence $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k \geq 0}$ such that there exists $\bar{\epsilon} > 0$ satisfying $\varepsilon_k \leq \bar{\epsilon}$ for all $k \in \mathbb{N}^*$, for any sequence $\boldsymbol{\gamma} = (\gamma_k)_{k \geq 0}$, there exists a constant C such that for any $z \in \mathcal{X}$,

$$\sup_{s \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}^{\boldsymbol{\gamma}}_{z,s} \left[V^p(z_k) \mathbb{1}_{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\varepsilon}) \geq k} \right] \leq C_{\mathcal{K}} V^q(z), \tag{4.12}$$

where $\nu(\boldsymbol{\varepsilon}) = \inf\{k \geq 1, \|s_k - s_{k-1}\| \geq \varepsilon_k\}$ and $\sigma(\mathcal{K}) = \inf\{k \geq 1, s_k \notin \mathcal{K}\}$ and the expectation is related to the non-homogeneous Markov chain $((z_k, s_k))_{k \geq 0}$ using the step-size sequence $\boldsymbol{\gamma} = (\gamma_k)_{k \geq 0}$.

(A4) The sequences $\boldsymbol{\gamma} = (\gamma_k)_{k \geq 0}$ and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k \geq 0}$ are non-increasing, positive and satisfy: $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\lim_{k \to \infty} \varepsilon_k = 0$, $\sum_{k=1}^{\infty} \{\gamma_k^2 + \gamma_k \varepsilon_k^a + (\gamma_k \varepsilon_k^{-1})^p\} < \infty$, where a and p are defined in (**A3"**).

We assume also some regularity properties of the stationary distribution with respect to s.

(B2) For all $z \in \mathcal{X}$, the functions $s \mapsto \pi_s$ and $s \mapsto \nabla_z \log \pi_s$ are continuous on $\mathcal{S}$.

Theorem 1 (Convergence Result for Stochastic Approximation). *Assume **(A1')**,**(A2')**, **(A3")** and **(A4)**. Assume also that for all $s \in \mathcal{S}$, the density function π_s satisfies assumptions **(B1-B2)**. Let $\mathrm{K} \subset \mathcal{X}$ be such that $\sup_{z \in \mathrm{K}} V(z) < \infty$ and $\mathcal{K}_0 \subset \mathcal{W}_{M_0} \cap \mathcal{S}_a$ (where M_0 is defined in **(A1')**), and let $(Z_k)_{k \geq 0}$ be the sequence defined in Equation (4.7). Then, for all $z_0 \in \mathrm{K}$ and $s_0 \in \mathcal{K}_0$, we have $\lim_{k \to \infty} d(s_k, \mathcal{L}) = 0$ $\bar{\mathbb{P}}_{z_0,s_0,0,0,0}$-a.s, where $\bar{\mathbb{P}}_{z_0,s_0,0,0,0}$ is the probability measure associated with the chain $(Z_k = (z_k, s_k, \kappa_k, \zeta_k, \nu_k))_{k \geq 0}$ starting at $(z_0, s_0, 0, 0, 0)$.*

Proof. The proof of this result is based on the theorem on general convergence result for stochastic approximation stated by Allassonnière et al (2010b) which generalized the one stated by Andrieu et al (2005). Assumptions (**A1'**) and (**A4**) are the same. We prove that the choice of the Anisotropic Metropolis Adjusted Langevin Algorithm as the transition probability for the dynamic of the current Markov Chain combined with assumptions (**A2'**) and (**A3"**) implies that assumptions (**A2**) and (**A3'**) of the theorem stated by Allassonnière et al (2010b) are fulfilled. The details are in Appendix 7.

Remark 2. *The differences between assumptions **(A2')**, **(A3")** and **(A2)** and **(A3')** from Allassonnière et al (2010b) come from the existence of a unique stationary distribution and the existence of a solution to the Poisson equation which are actually induced by the theoretical properties of the AMALA sampling procedure.*

□

4.3. Convergence property of the estimated sequence generated by the AMALA-SAEM Algorithm

We do the following assumptions on the model which are quite usual in the context of missing data model using EM-like algorithms (see Delyon et al (1999), Kuhn and Lavielle (2004)).

For sake of simplicity we denote in the sequel $p_\theta(\cdot)$ instead of $p(\cdot|y;\theta)$ the posterior distribution.

- (**M1**) The parameter space Θ is an open subset of $\mathbb{R}^p$. The complete data likelihood function is given by:
$$f(y, z; \theta) = \exp\left\{-\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right\},$$
where S is a Borel function on $\mathbb{R}^l$ taking its values in an open subset $\mathcal{S}$ of $\mathbb{R}^m$. Moreover, the convex hull of $S(\mathbb{R}^l)$ is included in $\mathcal{S}$, and, for all θ in Θ,
$$\int ||S(z)|| p_\theta(z) \mu(dz) < \infty.$$
- (**M2**) The functions ψ and ϕ are twice continuously differentiable on Θ.
- (**M3**) The function $\bar{s} : \Theta \to \mathcal{S}$ defined as
$$\bar{s}(\theta) \triangleq \int S(z) p_\theta(z) \mu(dz)$$
is continuously differentiable on Θ.
- (**M4**) The function $l : \Theta \to \mathbb{R}$ defined as the observed-data log-likelihood
$$l(\theta) \triangleq \log g(y;\theta) = \log \int f(y, z; \theta) \mu(dz)$$
is continuously differentiable on Θ and
$$\partial_\theta \int f(y, z; \theta) \mu(dz) = \int \partial_\theta f(y, z; \theta) \mu(dz).$$

- **(M5)** There exists a function $\hat{\theta}: \mathcal{S} \to \Theta$, such that:

$$\forall s \in \mathcal{S}, \quad \forall \theta \in \Theta, \quad L(s; \hat{\theta}(s)) \geq L(s; \theta).$$

  Moreover, the function $\hat{\theta}$ is continuously differentiable on $\mathcal{S}$.
- **(M6)** The functions $l: \Theta \to \mathbb{R}$ and $\hat{\theta}: \mathcal{S} \to \Theta$ are m times differentiable.
- **(M7)**
  (i) There exists an $M_0 > 0$ such that

$$\left\{ s \in \mathcal{S}, \partial_s l(\hat{\theta}(s)) = 0 \right\} \subset \{ s \in \mathcal{S}, \ -l(\hat{\theta}(s)) < M_0 \}.$$

  (ii) For all $M_1 > M_0$, the set $\bar{Conv}(S(\mathbb{R}^l)) \cap \{ s \in \mathcal{S}, \ -l(\hat{\theta}(s)) \leq M_1 \}$ is a compact set of $\mathcal{S}$.
- **(M8)** There exists a polynomial function P of degree 2 such that for all $z \in \mathcal{X}$

$$||S(z)|| \leq P(z).$$

- **(B3)** For any compact subset $\mathcal{K}$ of $\mathcal{S}$, there exists a polynomial function Q of the hidden variable such that $\sup_{s \in \mathcal{K}} |\nabla_z \log p_{\hat{\theta}(s)}(z)| \leq Q(z)$.

Theorem 2 (Convergence Result for the Estimated Sequence generated by Algorithm 1)**.** *Assume **(M1-M8)** and **(A4)**. Assume that the family of posterior density functions $\{p_{\hat{\theta}(s)}, \ s \in \mathcal{S}\}$ satisfies assumptions **(B1-B3)**.*

*Let $\mathrm{K} \subset \mathcal{X}$ be such that $\sup_{z \in \mathrm{K}} p^{-\beta}_{\hat{\theta}(s_1)}(z) < \infty$ and $\mathcal{K}_0 \subset \mathcal{W}_{M_0} \cap \bar{Conv}(S(\mathbb{R}^l))$ (where M_0 is defined in **(M7)**). Then, for all $z_0 \in \mathrm{K}$ and $s_0 \in \mathcal{K}_0$, we have $\lim_{k \to \infty} d(\theta_k, \mathcal{L}') = 0$ a.s. where $(\theta_k)_k$ is the sequence generated by Algorithm 1 and $\mathcal{L}' \triangleq \{\theta \in \Theta, \partial_\theta l(\theta) = 0\}$.*

The proof is postponed to Appendix 7.3.

4.4. Central Limit Theorem for the estimated sequence generated by the AMALA-SAEM

Theorem 2 ensures that the number of re-initializations of the sequence of stochastic approximation of Algorithm 1 is finite almost surely. We can therefore consider only the non truncated sequence when we are interested in its asymptotic behavior.

Let us write the stochastic approximation procedure :

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k \eta_k$$

where $\eta_k = S(z_k) - \mathbb{E}_{p_{\hat{\theta}(s_{k-1})}}(S(z))$.

Let us introduce some assumptions.

(N1) The function h is C^1 in some neighborhood of s^* which satisfies $h(s^*) = 0$ with first derivatives Lipschitz and J the Jacobean matrix of the mean field h in s^* has all its eigenvalues with negative real part.

(N2) Let $\xi_k = g_{s_{k-1}}(z_k) - \Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1})$ where for any $s \in \mathcal{S}$, g_s is a solution of the Poisson equation $g - \Pi_s g = H_s - p_s(H_s)$. Assume that for some matrix U and some random variable $X > 0$,

$$\lim_{k \to \infty} \mathbb{E}\left[|X \sum_{i=1}^{k} (\xi_i \xi_i' - U)| \right] = 0.$$

(N3) The step size sequence (γ_k) is decreasing and satisfies either case 1 or case 2:

Case 1 $\lim_{k \to \infty} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) = 0,$

Case 2 for all $k \in \mathbb{N}^*$ $\gamma_k = 1/k$.

In Case 2, it requires also all eigenvalues of J have a real part lower than $-1/2$.

Theorem 3. *Under the assumptions of Theorem 2 and under **(N1)-(N3)**, the sequence $(s_k - s^*)/\sqrt{\gamma_k}$ converges in distribution to a Gaussian random vector with zero mean and covariance matrix Γ and*

$$\sqrt{k}(\theta_k - \theta^*) \to_{\mathcal{L}} \mathcal{N}(0, \partial_s \hat{\theta}(s^*) \Gamma \partial_s \hat{\theta}(s^*)')$$

where $\theta^* = \hat{\theta}(s^*)$.

The matrix Γ is characterized as the solution of some Lyapunov equation depending on the matrices U and J (see Theorem 4 in Section Appendix). The proof of Theorem 3 is given in Appendix 7.4.

5. Applications on Bayesian Mixed Effect Template model

We apply our estimation process on different databases. The first one is the USPS hand-written digit base as used by Allassonnière et al (2007) and Allassonnière et al (2010b). The others two are medical images of 2D corpus callosum and 3D murine dendrite spine excrescences used by Allassonnière et al (2010a).

We begin with presenting the experiments on the USPS database. In order to make comparison, we estimate the parameters in the same conditions as in the previous mentioned works that is to say the same 20 images per digit. Each image has grey level between 0 (background) and 2 (bright white). These images are presented on the left panel of Fig. 1. We also use a noisy training dataset generated by adding a standardized independent Gaussian noise. These images are presented on the right panel of Fig. 1. We test five algorithms: the deterministic approximation of the EM algorithm (FAM-EM) presented by Allassonnière et al (2007), four MCMC-SAEM where the sampler is either the MALA, the adaptive MALA proposed by Atchadé (2006), the hybrid Gibbs sampler presented by Allassonnière et al (2010b) and our AMALA algorithm.

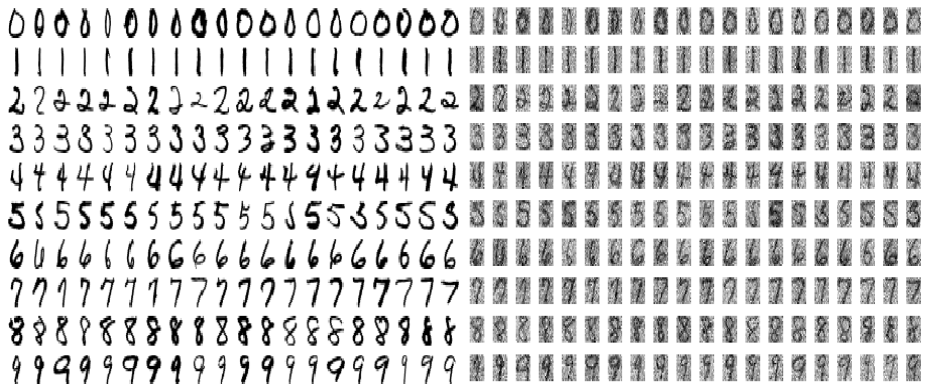

FIG 1. *Left: some images of the training set used for the estimation of the model parameters (inverse video). Right: same examples with additive noise.*

For these experiments the tuning parameters are chosen as follows: the threshold b is set to $1,000$, the scale $\delta = 10^{-3}$, the regularization $\varepsilon = 10^{-4}$. The other tuning parameters and hyper-parameters are chosen as in Allassonnière et al (2010b).

5.1. Computational performances

We compare first the computational performances of the algorithms. The computational time is smaller for the three MCMC-SAEM algorithms using "MALA-like" samplers compared to the

FAM. Indeed, a numerical convergence of that algorithm requires about 30 to 50 EM steps. Each of them requires a gradient descent which has 15 iterations in average. This implies to compute 15 times the gradient of the energy (which actually equals our gradient) for each image for each EM step. The "MALA-like"-SAEM require about 100 to 150 EM steps (depending on the digit) but only one gradient is computed for each image at each step. This reduces the computational time by a factor of at least 4 (up to 7 depending on the digit). No comparison can be done when the data are noisy since the FAM-EM does not converges toward the MAP estimator as mentioned above. Comparing to the hybrid Gibbs-SAEM, the computational time is 8 times lower with the AMALA-SAEM in this particular case of application. Indeed, the hybrid Gibbs sampler requires no computation of the gradient. However, it includes a loop over the coordinates of the hidden variable, here the deformation vector of size $2k_g = 72$. At each of these iterations, the candidate is straightforward to sample whereas the computational cost lies into the acceptance rate. When this becomes heavy, the less times you calculate it, the better. In the AMALA-SAEM, this acceptance rate only has to be calculated once for each image. Therefore, even when the dimension of the hidden variable increases, this is of constant cost. The main price to pay is the computation of the gradient. Therefore, a tradeoff has to be found between the computation of either one gradient or dk_g acceptance rates in order to select the algorithm to use in a given case.

5.2. Results on the template estimation

All the estimated templates obtained with the five algorithms and noise-free and noisy training data are presented in Fig. 2. As noticed by Allassonnière et al (2010b), the FAM-EM estimation is sharp when the training set is noise-free and is deteriorated while adding noise. This behavior is not surprising with regard to the theoretical bound established by Bigot and Charlier (2011) in the particular case of compact deformation group. Considering the adaptive sampler, it does not reach a good estimation of the templates which are still very blurry and noisy in both cases. The problem seems to come from the very low acceptation rate already at the beginning of the estimation. The bad initial guess we have about the covariance matrix of the proposal seems to block the chain. Moreover, the tuning parameters are difficult to calibrate along the iterations of the estimation algorithm. Concerning the estimated templates using the Gibbs, MALA and AMALA samplers, they look very similar to each other using the noise-free data as well as the noisy ones. This similarity confirms the convergence of all these algorithms toward the MAP estimator. In this case, the templates are as expected, noise free and sharp.

Nevertheless, when the dimension of the hidden variable increases, both the Gibbs and the MALA samplers show limitations. The Gibbs-SAEM would produce sharp estimations but explodes the computational time. For this reason, we did not run this algorithm on higher dimension experiments. We run the estimation on the same noisy USPS database, increasing the number k_g of geometrical control points. We choose the dimension of the deformation vector equal to 72, 128 and 200. The results are presented in Fig. 3. Concerning the MALA sampler, it does not seem to capture the whole variability of the population in such high dimension. This yields a poorly estimation of the templates. This phenomenon does not appear using our AMALA-SAEM algorithm. The templates still look sharp and the acceptation rate remains reasonable.

5.3. Results on the covariance matrix estimation

We keep considering the USPS database. Since we are provided with a generative model, once the parameters have been estimated, we can generate synthetic samples in order to evaluate the constrained on the deformations that have been learnt. Some of these samples are presented in Fig. 4. For each digit, 20 examples are generated with the deformation given by $+z$ and 20 others with $-z$ where z is simulated with respect to $\mathcal{N}(0, \Gamma_g)$. We recall that, as already noticed by Allassonnière et al (2010b), the Gaussian distribution is symmetric which may lead to strange samples in one direction whereas the other one looks like something present in the training set.

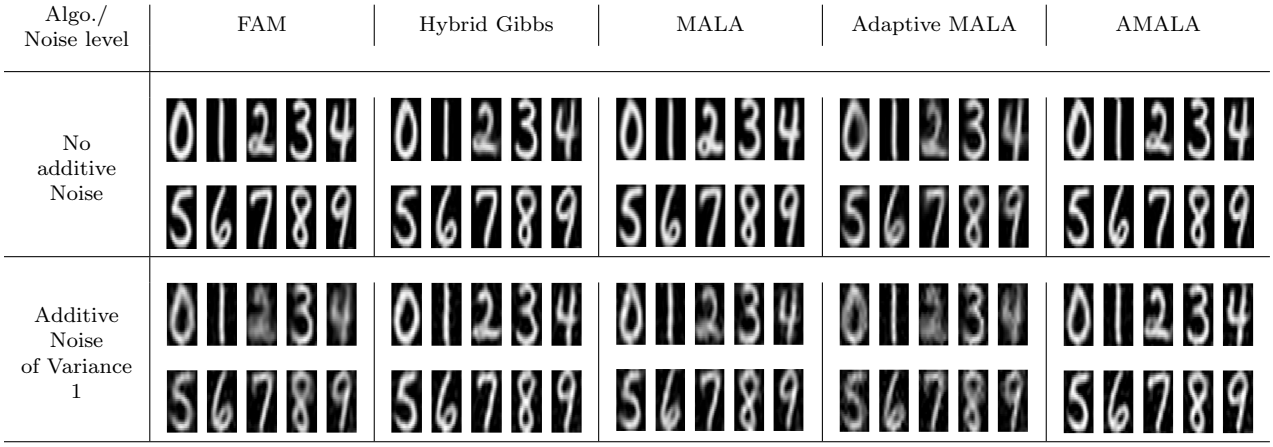

FIG 2. *Estimated templates using the five algorithms and noise free and noisy data. The training set includes* 20 *images per digit. The dimension of the hidden variable is* 72.

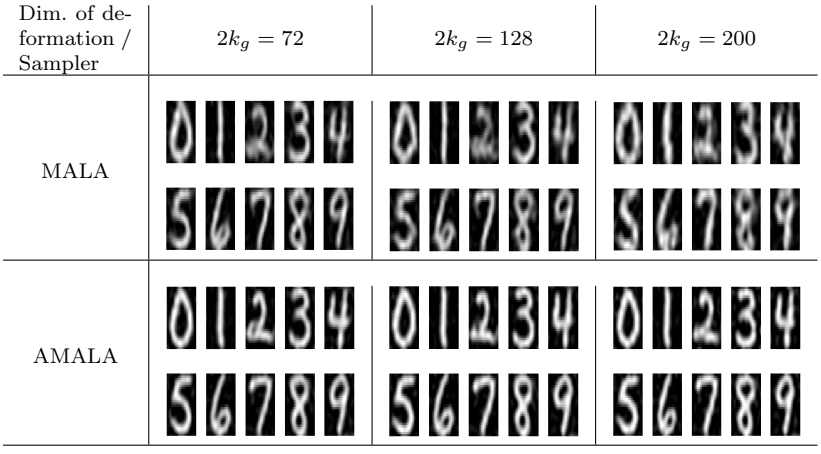

FIG 3. *Estimated templates using MALA and AMALA samplers in the stochastic EM algorithm on noisy training data. The training set includes* 20 *images per digit. The dimension of the hidden variable increases from* 72 *to* 200.

With regards to the above remarks concerning the computational time and the template estimations, we present in this subsection only the results obtained using MALA and AMALA-SAEM algorithms. We notice that the samples generated by both algorithms look alike in the case of hidden variable of dimension 72. Thus, we present the results of our AMALA-SAEM estimation. As we can see, the deformations are very well estimated in both cases (without or with noise) and even look similar. This tends to demonstrate that the noise has really been separated from the template as well as the geometric variability during the estimation process.

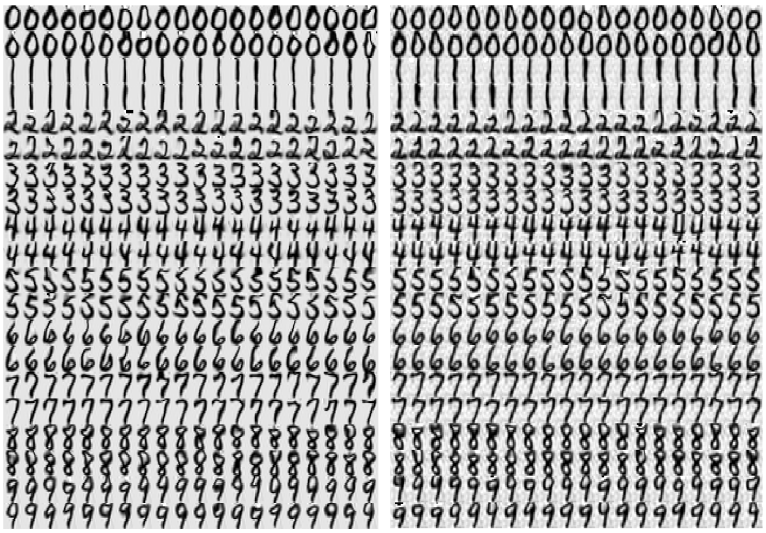

FIG 4. *Synthetic samples generated with respect to the BME template model using the estimated parameters with AMALA-SAEM. For each digit, the two lines represent the deformation using* $+$ *and* $-$ *the simulated deformation* z*. Left: data without noise. Right: data with noise variance* 1*. The number of geometric control points is* 36 *leading to a hidden variable of dimension* 72*.*

Increasing the dimension of the deformation to 128, we run both algorithms on the noisy dataset. We observe on Fig. 5 that the geometric variability of the samples remains similar to the one obtained in lower dimension using our AMALA-SAEM. However, the MALA-SAEM does not manage to capture the whole variability of the deformations which is related the results observed above on the template. This confirms the limitation of the use of MALA-SAEM in higher dimension.

5.4. Results on the noise variance estimation

The last check of the accuracy of the estimation relies in the noise variance estimation. The plots of their evolutions along the AMALA-SAEM iterations for each digit in both cases (without and with noise) are presented in Fig. 6. This variance is underestimated in particular in the noisy case, which is a well-known effect of the maximum likelihood (or in our case the MAP) estimator. We observe that the geometrically very constrained digits as 1 or 7 tend to converge very quickly whereas the digits 2 and 4 require more iterations to capture all the shape variability.

Since this is a real parameter, we used it to illustrate the Central Limit Theorem stated in Subsection 4.4. Fig. 7 and Fig. 8 show the histograms of $10,000$ runs of the algorithm with the same initial conditions. We use the digits 0 and 2 of the original data set as well as of the noisy data. As the iterations go along, the distribution of the estimates tends to look like a Gaussian distribution centered in the estimated noise variances which demonstrates empirically the Central Limit Theorem.

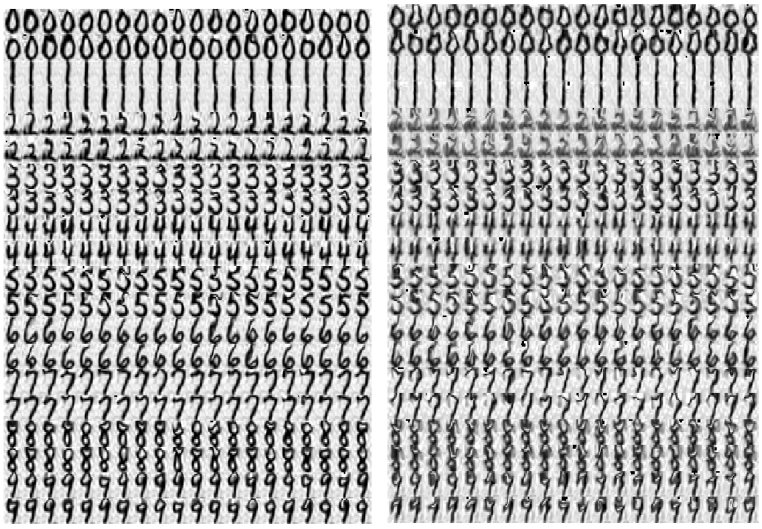

FIG 5. *Synthetic samples generated with respect to the BME template model using the estimated parameters with AMALA-SAEM (left) and MALA-SAEM (right). For each digit, the two lines represent the deformation using* $+$ *and* $-$ *the simulated deformation* z*. The number of geometric control points is* 64 *leading to a hidden variable of dimension* 128*.*

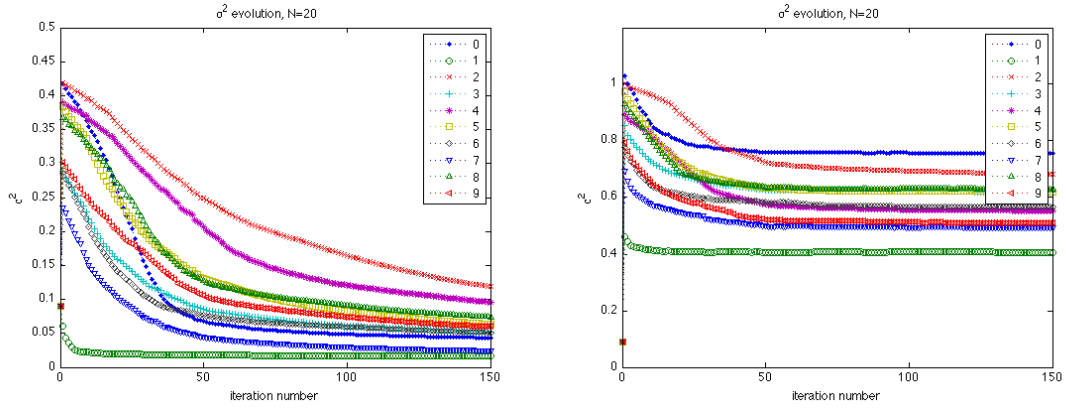

FIG 6. *Evolution of the estimation of the noise variance along the AMALA-SAEM iterations. Left: original data. Right: noisy data.*

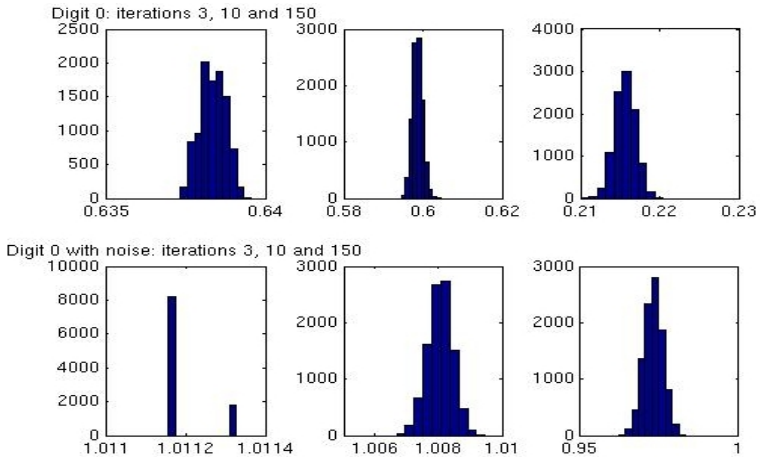

FIG 7. *Empirical convergence toward the Gaussian distribution of the estimated noise variance along the AMALA-SAEM iterations for digit* 0. *Left: original data. Right: noisy data.*

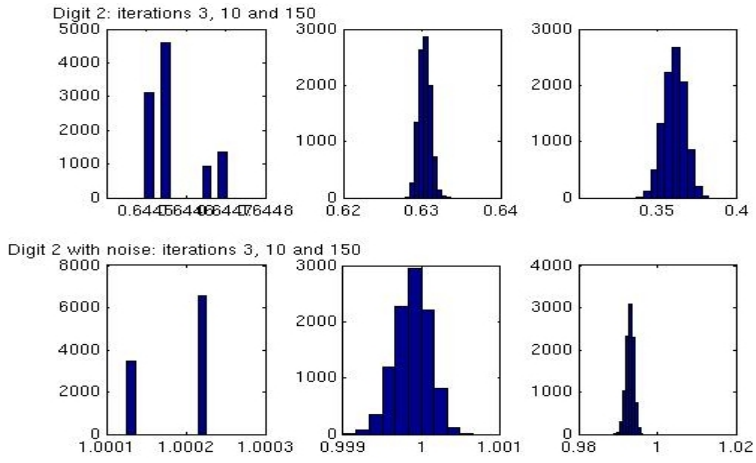

FIG 8. *Empirical convergence toward the Gaussian distribution of the estimated noise variance along the AMALA-SAEM iterations for digit* 2. *Left: original data. Right: noisy data.*

5.5. Classification results

The deformable template model enables to perform classification using the maximum likelihood of a new image to allocate it to one class, here the digit. We use the test USPS database (which contains 2007 digits) for classification while the training was done on the previous 20 noisy images. The results obtained with the hybrid Gibbs, MALA and AMALA-SAEM are presented in Table 1. In dimension 72, the best classification rate is performed by the hybrid Gibbs-SAEM. This is easily understandable since the sampling scheme enables to catch deformations which have been optimized control point by control point. Therefore, the estimated covariance matrix carries more local accuracy. The AMALA-SAEM leading in a much smaller computation time to estimates of the same quality provides also a very good classification rate. This confirms the good results observed on both the template estimates and the synthetic samples. Unfortunately, the MALA-SAEM shows again some limitations. Even if the templates look acceptable, the sampler does not manage to capture the whole class variability. Therefore, the classification rate falls down.

In order to evaluate the stability of our estimation algorithm with respect to the dimension, we perform the same classification with more control points. As expected, the MALA-SAEM classification rate is deteriorated whereas our AMALA-SAEM keeps very good performances. Note that the hybrid Gibbs sampler was not tested in dimension $2k_g = 128$ because of its very long computational time.

Sampler / Dim of Deformation	Hybrid Gibbs	MALA	AMALA
72	22.43	35.98	23.22
128	×	43.8	25.36

TABLE 1

Error rate using the estimations on the noisy training set with respect to the sampler used in the MCMC-SAEM algorithm and the dimension of the deformation $2k_g$. The classification is performed on the test set of the USPS database.

5.6. Medical image template estimation

A second database is used to illustrate our algorithm. As before, in order to make comparisons with existing algorithms, we use the same database presented by Allassonnière et al (2010a). It consists of 47 medical images, each of them is a $2D$ square zone around the end point of the corpus callosum. This box contains a part of this corpus callosum as well as a part of the cerebellum. Ten exemplars are presented in the top rows of Fig. 9.

The estimations are compared with these obtained with the FAM-EM and the hybrid Gibbs-SAEM algorithms and with the mean image (bottom row of Fig. 9). In this real situation, the Euclidean mean image (a) is very blurry. The estimated template using the FAM-EM (b) provides a first amelioration in particular leading to a sharper corpus callosum. However, the cerebellum still looks blurry in particular when comparing it to the shape which appears in the template estimated using the hybrid Gibbs SAEM (c). The result of our AMALA-SAEM is given in image (d). This template is very close to (c) as we could expect at a convergence point. Nevertheless the AMALA-SAEM has much lower computational time than the hybrid Gibbs-SAEM. This shows the advantage of using AMALA-SAEM in real cases of high dimension.

5.7. 3D medical images

We also test our algorithm in much higher dimension using the dataset of murine dendrite spines (see Aldridge et al (2005); Ceyhan et al (2007a,b)) already used by Allassonnière et al (2010a). The dataset consists of 50 binary images of microscopic structures, tiny protuberances found on many

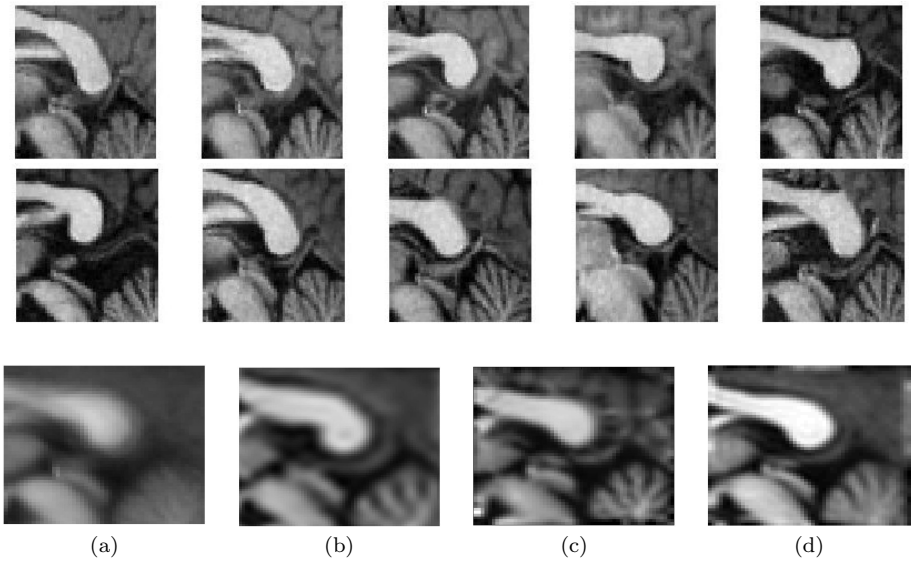

FIG 9. *Medical image template estimation. Top rows :* 10 *Corpus callosum and cerebellum training images among the* 47 *available. Bottom row : (a) mean image. (b) FAM-EM estimated template. (c) Hybrid Gibbs - SAEM estimated template. (d) AMALA-SAEM estimated template.*

types of neurons termed dendrite spines. The images are from control mice and knockout mice which have been genetically modified to mimic human neurological pathologies like Parkinson's disease. The acquisition process consisted of electron microscopy after injection of Lucifer yellow and subsequent photo-oxidation. The shapes were then manually segmented on the tomographic reconstruction of the neurons. Some of these binary images are presented in Fig. 10 which shows a 3D view of some exemplars among the training set. Each image is a binary (background = 0, object = 2) cubic volume of size 28^3. We can notice here the large geometrical variability of this population of images. Therefore we use a hidden variable of dimension $3k_g = 648$ to catch this complex structure.

The template estimated with either 30 or 50 observations are presented in Fig. 12. We obtain similar shapes which are coherent with what a mean shape could be regarding the training sample. To evaluate the estimated geometrical variability, we generate synthetic samples as done in Subsection 5.3. Eight of these are shown in Fig. 11. We observe different twistings which are all coherent with the shapes observed in the dataset. Note that the training shapes have very irregular boundaries whereas the parametric model used for the template leads to a smoother image. Thus, the synthetic samples do not reflect the local ruggedness of the segmented murine dendrite spines. If the aim was to capture these local bumps, the number of photometrical control points has to be increased. However, the goal of our study was to detect global shape deformations.

6. Conclusion

In this paper we have considered the deformable template estimation issue using the BME model. We were particularly interested in the high dimensional setting. To that purpose, we have proposed to optimize the sampling scheme in the MCMC-SAEM algorithm to get an efficient and accurate estimation process. We have exhibited a new MCMC method based on the classical Metropolis Adjusted Langevin Algorithm where we introduced an anisotropic covariance matrix in the proposal. This optimization takes into account the anisotropy of the target distribution. We proved that the generated Markov chain is geometrically ergodic uniformly on any compact set. We have also proved the almost sure convergence of the sequence of parameters generated by the estimation algorithm as well as its asymptotic normality. We have illustrated this estimation algorithm in

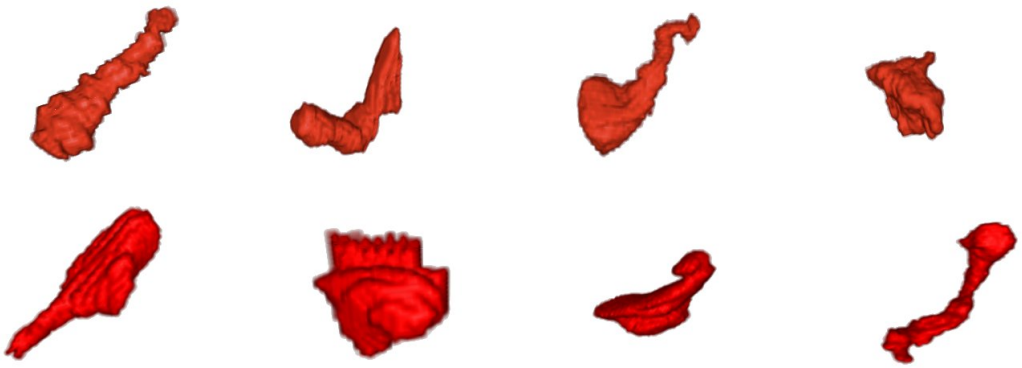

FIG 10. *3D views of eight samples of the data set of dendrite spines. Each image is a volume leading to a binary image.*

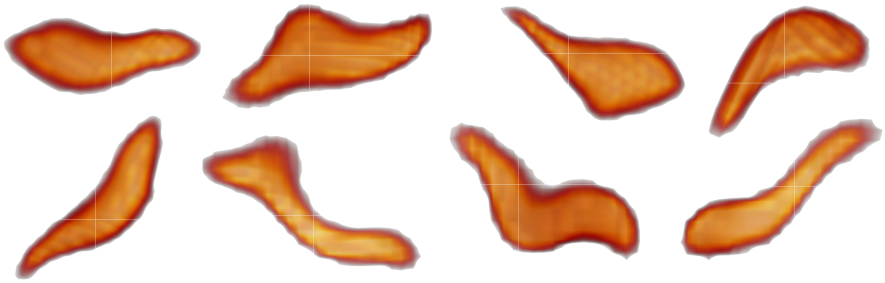

FIG 11. *3D views of eight synthetic data. The estimated template shown in Fig. 12 is randomly deformed with respect to the estimated covariance matrix.*

the BME model. We considered different datasets of the literature namely the USPS database, 2D medical images of corpus callosum and 3D medical images of murine dendrite excrescences. We have compared the results with previously published ones to highlight the gain in speed and accuracy of the proposed algorithm.

We emphasize that the proposed estimation scheme can be applied in a wide range of application fields involving missing data models in high dimensional setting. In particular, this method is promising when considering mixture models as by Allassonnière and Kuhn (2010). Indeed, it will enable to shorten the computation time of the simulation part which in that case requires the use of many auxiliary Markov chains. This also provides a good tool for this BME model when introducing a diffeomorphic constrain on the deformations. In this case, it is even more important to get an efficient estimation process since the computational cost of diffeomorphic deformation is

FIG 12. *Estimated templates of murine dendrite spines. The training set is either composed of* 30 *(left) or* 50 *(right) images.*

intrinsically large.

7. Appendix

7.1. Proof of Proposition 1

The idea of the proof is the same as the one of the geometric ergodicity of the random walk Metropolis algorithm developed by Jarner and Hansen (1998) and re-written by Atchadé (2006) for its adaptive version of the MALA with truncated drift. The fact that both the drift and the covariance matrix are bounded even depending on the gradient of $\log \pi$ enables some similar proofs.

Let us first recall the transition kernel:

$$\Pi(x, A) = \int_A \alpha(x, z) q_c(x, z) dz + \mathbb{1}_A(x) \int_{\mathcal{X}} (1 - \alpha(x, z)) q_c(x, z) dz \,, \tag{7.1}$$

where $\alpha(x, z) = \min(1, \rho(x, z))$ and $\rho(x, z) = \frac{\pi(z) q_c(z,x)}{q_c(x,z)\pi(x)}$.

Thanks to the bounded drift and covariance matrix, we can bound the proposal distribution q_c by two centered Gaussian distributions with covariance matrices $\epsilon_1 Id_l$ and $\epsilon_2 Id_l$ as follows: there exist two constants $0 < k_1 < k_2$ such that for all $(x, z) \in \mathcal{X}^2$

$$k_1 g_{\epsilon_1}(x - z) \leq q_c(x, z) \leq k_2 g_{\epsilon_2}(x - z) \,, \tag{7.2}$$

denoting by g_a the centered Gaussian probability density function in $\mathbb{R}^l$ with covariance matrix aId_l.

7.1.1. Proof of the existence of a small set $\mathtt{C}$

Let $K = B(0, R)$ be the ball in $\mathbb{R}^d$ centered at 0 and of radius $R > 0$. For any $x \in \mathcal{X}$ and for any $A \in \mathcal{B}$:

$$\Pi(x, A) \geq \int_{A \cap K} \alpha(x, z) q_c(x, z) dz \,.$$

Let $\mathtt{C}$ be a compact subset of $\mathcal{X}$ and
$\tau = \min_{x \in \mathtt{C},\ z \in K} \rho(x, z)$. Since ρ is a ratio of positive continuous functions in both variables, then $\tau > 0$. Therefore, for all $x \in \mathtt{C}$ and for any $A \in \mathcal{B}$:

$$\Pi(x, A) \geq \min(1, \tau) \int_{A \cap K} q_c(x, z) dz \,.$$

Moreover, thanks to (7.2),

$$\begin{aligned} \Pi(x, A) &\geq k_1 \min(1, \tau) \int_{A \cap K} g_{\epsilon_1}(z - x) dz \\ &\geq k_1 \min(1, \tau) \int_{\mathcal{X}} g_{\epsilon_1}(z) \mathbb{1}_{K_{\mathtt{C}}(A)}(z) dz \,, \end{aligned}$$

where since $\mathtt{C}$ is a compact subset of $\mathbb{R}^d$, it is bounded and we can define $K_{\mathtt{C}}(A)$ to be the intersection of all translations of $A \cap K$ by any element of $\mathtt{C}$. Note that we can choose R sufficiently large with respect to the size of $\mathtt{C}$ so that this intersection is not empty.

Let $\varepsilon = k_1 \min(1, \tau) Z$ and
$\nu(A) = \frac{1}{Z} \int_{\mathcal{X}} g_{\epsilon_1}(z) \mathbb{1}_{K_{\mathtt{C}}(A)}(z) dz$ where Z is the renormalisation constant. Thus, $\mathtt{C}$ is a small set for the transition kernel Π and (4.3) holds.

7.1.2. Proof of the drift condition

As already noticed by Atchadé (2006), we only need to prove the two following conditions:

$$\sup_{x\in\mathcal{X}} \frac{\Pi V(x)}{V(x)} < \infty \tag{7.3}$$

and

$$\limsup_{|x|\to\infty} \frac{\Pi V(x)}{V(x)} < 1\,. \tag{7.4}$$

We take the same path as Atchadé (2006) applied to our case and refer to Fig 13. for a 2D visualization of all the spaces introduced along the proof.

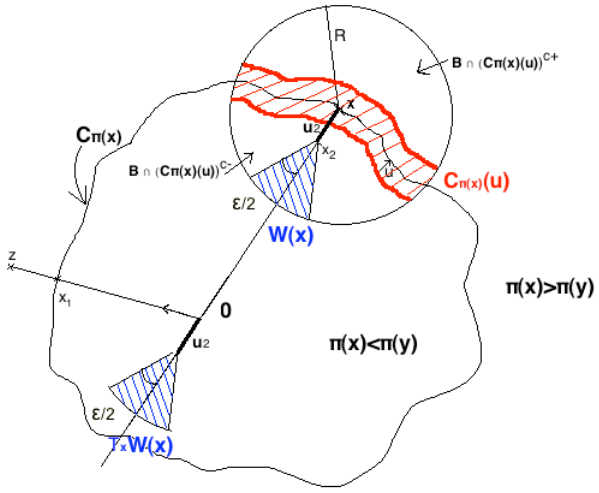

Fig 13. *2D representation of the sets used in the proof.*

For any $x \in \mathcal{X}$, we denote $A(x) = \{z \in \mathcal{X} \text{ such that } \rho(x,z) \geq 1\}$ the acceptance set and $R(x) = A(x)^c$ the complementary set of $A(x)$. Then, using $V(x) = c\pi(x)^{-\beta}$ for any $0 < \beta < 1$,

$$\begin{aligned}
\frac{\Pi V(x)}{V(x)} &= \int_{A(x)} q_c(x,z)\frac{V(z)}{V(x)}dz + \int_{R(x)} \frac{\pi(z)q_c(z,x)}{\pi(x)q_c(x,z)} q_c(x,z)\frac{V(z)}{V(x)}dz + \int_{R(x)} \left(1 - \frac{\pi(z)q_c(z,x)}{\pi(x)q_c(x,z)}\right) q_c(x,z)dz \\
&= \int_{A(x)} \underbrace{\frac{\pi(z)^{-\beta}}{\pi(x)^{-\beta}} q_c(x,z)}_{f_1(x,z)}\, dz + \int_{R(x)} \underbrace{\frac{\pi(z)^{1-\beta}}{\pi(x)^{1-\beta}} q_c(z,x)}_{f_2(x,z)}\, dz + \int_{R(x)} \underbrace{q_c(x,z)}_{f_3(x,z)} dz
\end{aligned}$$

On the acceptance set $A(x)$:

$$\frac{\pi(z)^{-\beta}}{\pi(x)^{-\beta}} q_c(x,z) \leq q_c(z,x)^{\beta} q_c(x,z)^{1-\beta}\,.$$

Thanks to Equation (7.2) one can bound this term by the following symmetric Gaussian distribution:

$$\frac{\pi(z)^{-\beta}}{\pi(x)^{-\beta}} q_c(x,z) \leq k_2 g_{\epsilon_2}(z-x) \tag{7.5}$$

which yields:

$$\int_{A(x)} f_1(x,y)dz \leq k_2 \int_{A(x)} g_{\epsilon_2}(z-x)dz\,. \tag{7.6}$$

Equivalently on $R(x)$, we have the following bound:

$$\frac{\pi(z)^{1-\beta}}{\pi(x)^{1-\beta}} q_c(z,x) \leq q_c(x,z)^{1-\beta} q_c(z,x)^{\beta} \leq k_2 g_{\epsilon_2}(z-x)\,. \tag{7.7}$$

Let fix $\varepsilon > 0$, there exists $a > 0$ such that $\int_{B(x,a)} g_{\epsilon_2}(z-x)dz \geq 1-\varepsilon$. This leads to:

$$\int_{A(x)\cap B(x,a)^c} f_1(x,z)dz \leq k_2\varepsilon$$

Let $C_{\pi(x)}$ be the level set of π in x: $C_{\pi(x)} = \{z \in \mathcal{X} \ : \ \pi(z) = \pi(x)\}$. We define a pipe around this level set as
$C_{\pi(x)}(u) = \{z + s\ n(z),\ |s| \leq u,\ z \in C_{\pi(x)}\}$.

Thanks to assumption (**B1**), there exists $r_1 > 0$ such that for all $x \in \mathcal{X}$ satisfying $|x| \geq r_1$ then 0 is inside the hyperspace defined by the level set $C_{\pi(x)}$ ($\pi(0) > \pi(x)$). Therefore, let $x \in \mathcal{X},\ |x| \geq r_1$, then for all $z \in \mathcal{X}$, $\exists x_1 \in C_{\pi(x)}$ such that $z = x_1 + s\ n(x_1)$.

Since $z \mapsto g_{\epsilon_2}(z-x)$ is a smooth density in the variable z, we can find $u > 0$ sufficiently small such that

$$\int_{B(x,a)\cap C_{\pi(x)}(u)} g_{\epsilon_2}(z-x)dz \leq \varepsilon\,, \tag{7.8}$$

leading to

$$\int_{A(x)\cap B(x,a)\cap C_{\pi(x)}(u)} f_1(x,z)dz \quad \leq \quad k_2\varepsilon\,.$$

Assumption (**B1**) implies that for any $r > 0$ and $s > 0$, $d_r(s) = \sup_{|x|\geq r} \frac{\pi(x+s\ n(x))}{\pi(x)}$ goes to 0 as r goes to ∞. Denote $C_{\pi(x)}(u)^{c+} = \{z \in C_{\pi(x)}(u)^c \ s.t.\ \pi(x) > \pi(z)\}$ and $C_{\pi(x)}(u)^{c-} = \{z \in C_{\pi(x)}(u)^c \ s.t.\ \pi(x) < \pi(z)\}$. Therefore there exists $r_2 > r_1 + a$ such that for any $x:\ |x| \geq r_2$ and using also the definition of the acceptance set $A(x)$:

$$\begin{aligned}\int_{A(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c+}} f_1(x,z)dz \quad &\leq \quad \int_{A(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c+}} \left(\frac{\pi(z)}{\pi(x)}\right)^{1-\beta} q_c(z,x)dz \\ &\leq \quad d_{r_2}(u)^{1-\beta} k_2 \int_{\mathcal{X}} g_{\epsilon_2}(z-x)dz \leq k_2 d_{r_2}(u)^{1-\beta}\,,\end{aligned}$$

using equation (4.2) which states that the stationary distribution is decreasing in the direction of the normal of x sufficiently large.

In the same way, one has:

$$\begin{aligned}\int_{A(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c-}} f_1(x,z)dz \quad &\leq \quad \int_{A(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c-}} \left(\frac{\pi(z)}{\pi(x)}\right)^{-\beta} q_c(x,z)dz \\ &\leq \quad k_2 d_{r_2}(u)^{\beta}\,.\end{aligned}$$

The same inequalities can be obtained for f_2 using the same arguments:

$$\begin{aligned}\int_{R(x)\cap B(x,a)^c} f_2(x,z)dz \quad &\leq \quad k_2\varepsilon \\ \int_{R(x)\cap B(x,a)\cap C_{\pi(x)}(u)} f_2(x,z)dz \quad &\leq \quad k_2\varepsilon \\ \int_{R(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c+}} f_2(x,z)dz \quad &\leq \quad k_2 d_{r_2}(u)^{1-\beta} \\ \int_{R(x)\cap B(x,a)\cap C_{\pi(x)}(u)^{c-}} f_2(x,z)dz \quad &\leq \quad k_2 d_{r_2}(u)^{\beta}\,.\end{aligned}$$

This yields

$$\limsup_{|x|\to\infty} \frac{\Pi V(x)}{V(x)} = \limsup_{|x|\to\infty} \int_{R(x)} q_c(x,z)dz. \tag{7.9}$$

Let $Q(x, A(x)) = \int_{A(x)} q_c(x,z)dz$, we get

$$\limsup_{|x|\to\infty} \frac{\Pi V(x)}{V(x)} = 1 - \liminf_{|x|\to\infty} Q(x, A(x)).$$

Let us now prove that $\liminf_{|x|\to\infty} Q(x, A(x)) \geq c > 0$ where c does not depend on x.

Let a fixed as above. Since q_c is an exponential function, there exists $c_0^a > 0$ such that

$$\inf_{z\in B(x,a)} \frac{q_c(z,x)}{q_c(x,z)} \geq c_0^a\,. \tag{7.10}$$

Moreover, thanks to assumption (**B1**) there exists $r_3 > 0$ such that for all $x \in \mathcal{X}$, $|x| \geq r_3$, there exists $0 < u_2 < a$ such that,

$$\frac{\pi(x)}{\pi(x - u_2\ n(x))} \leq c_0^a\,. \tag{7.11}$$

Hence, for $|x| \geq r_3$, any point $x_2 = x - u_2\ n(x)$ belongs to $A(x)$.

Let $W(x)$ be the cone defined as:

$$W(x) = \left\{x_2 - s\zeta,\ 0 < s < a - u_2,\ \zeta \in \mathcal{S}^{d-1}, |\zeta - n(x_2)| \leq \frac{\varepsilon}{2}\right\} \tag{7.12}$$

where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

Let us prove that $W(x) \subset A(x)$.

Using assumption (**B1**), we have for a sufficiently large x: $m(x).n(x) \leq -\varepsilon$. Besides, by construction of $W(x)$ for large x, for all $z \in W(x)$, $|n(z) - n(x)| \leq \varepsilon/2$ with $n(x) = n(x_2)$ (see Fig 13). This leads to for any sufficiently large x, for all $z \in W(x)$,

$$m(z).\zeta = m(z).(\zeta - n(x_2)) + m(z).(n(x_2) - n(z)) + m(z).n(z) \leq \varepsilon/2 + \varepsilon/2 - \varepsilon = 0\,. \tag{7.13}$$

Let now $z = x_2 - s\zeta \in W(x)$. Using the mean value theorem on the differentiable function π between x_2 and z, we get that there exists $\tau \in]0, s[$ such that $\pi(z) - \pi(x_2) = -s\zeta.\nabla\pi(x_2 - \tau\zeta)$. Using the definition of m, this implies that $\pi(z) - \pi(x_2) = -s\zeta.m(x_2 - \tau\zeta)|\nabla\pi(x_2 - \tau\zeta)| \geq 0$ thanks to Equation (7.13). Putting all these results together we finally get that for all $z \in W(x)$, $\pi(z) \geq \pi(x_2) \geq \frac{1}{c_0^a}\pi(x)$. Moreover, as $z \in B(x,a)$ as well, Equation (7.10) is satisfied, leading to $z \in A(x)$.

Then, we have

$$\begin{aligned} Q(x, A(x)) &= \int_{A(x)} q(x,z)dz \\ &\geq \int_{A(x)} k_1 g_{\epsilon_1}(z - x)dz \\ &\geq k_1 \int_{W(x)} g_{\epsilon_1}(z - x)dz \\ &= \int_{T_x(W(x))} g_{\epsilon_1}(z)dz \end{aligned}$$

where

$$T_x(W(x)) = \left\{-u_2\ n(x) - s\zeta,\ 0 < s < a - u_2,\ \zeta \in \mathcal{S}^{d-1},\ |\zeta - n(x)| \leq \frac{\varepsilon}{2}\right\} \tag{7.14}$$

is the translation of the set $W(x)$ by the vector x. But since g_{ϵ_1} is isotropic and $T_x(W(x))$ only depends on a fixed constant u_2 and $n(x)$, this last integral is independent of x, so there exists a positive constant c such that:

$$c = \int_{T_x(W(x))} g_{\epsilon_1}(z)dz\,. \tag{7.15}$$

Back to our limit,

$$\limsup_{|x|\to\infty} \frac{\Pi V(x)}{V(x)} \leq 1 - c \tag{7.16}$$

which ends the proof of the condition (7.4).

To prove (7.3), we use the previous result. Indeed, since $\frac{\Pi V(x)}{V(x)}$ is a smooth function on $\mathcal{X}$ it is bounded on every compact subset. Moreover since the lim sup is finite, then it is also bounded outside a fixed compact. This proves the results.

7.2. Proof of Theorem 1

We provide here the proof of the convergence of the stochastic approximation sequence. Proposition 1 ensures the existence of a small set for each transition probability Π_s which satisfies also a drift condition. Therefore, each of this transition probability generates a uniform ergodic chain which admits an unique stationary distribution (see Meyn and Tweedie (1993)). This result combined with assumption (**A2'**) implies assumption (**A2**) of Allassonnière et al (2010b).

We now prove assumption (**A3'**). Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. The proof of the existence of a small set for each transition probability detailed in Section 7.1.1 can be easily adapted to construct a common small set $\mathtt{C}$ for all transition probabilities obtained for all $s \in \mathcal{K}$ where $\mathcal{K}$ is compact.

Moreover, each transition probability satisfies a drift condition:

$$\forall s \in \mathcal{K},\ \forall z \in \mathcal{X} \quad \Pi_s V_s(z) \leq \lambda_s V_s(z) + b_s \mathbb{1}_{\mathtt{C}}(z), \tag{7.17}$$

where $\lambda_s \in]0,1[$, $b_s > 0$ and $V_s(z) = c_s^\beta \pi_s^{-\beta}(z)$ with $c_s = \max_{z\in\mathcal{X}} \pi_s(z)$. Note that assumption **(B1)** ensures that $0 < c_s < \infty$.

We now have to prove that there exist a function V, constants $m \in \mathbb{N}^*$, $0 < \lambda_m < 1$ and $b > 0$ independent of s such that

$$\forall z \in \mathcal{X} \quad \sup_{s\in\mathcal{K}} \Pi_s^m V(z) \leq \lambda_m V(z) + b\mathbb{1}_{\mathtt{C}}(z).$$

Thanks to assumption **(B2)**, there exists a constant c_0^a uniform in s with respect to the compact $\mathcal{K}$ such that Equations (7.10) and (7.11) still hold for all $s \in \mathcal{K}$. This implies that the set $T_x(W(x))$ is independent of $s \in \mathcal{K}$. Therefore, we can set $\lambda = 1 - c < 1$ where c is defined in Equation (7.15) and is also independent of any s in the compact $\mathcal{K}$. The same arguments hold for b.

Let $s_1 \in \mathcal{S}$ be fixed arbitrarily and define $V(z) = c_{s_1}^\beta \pi_{s_1}^{-\beta}(z)$. Since $\mathcal{K}$ is compact, there exist two constants (c_1, c_2) such that $\forall s \in \mathcal{K}\ \forall z \in \mathcal{X}\ c_1\pi_{s_1}(z) \leq \pi_s(z) \leq c_2\pi_{s_1}(z)$.

Then, we get

$$\begin{aligned}
\Pi_s^m V(z) &\leq c_{s_1}^\beta c_2^\beta c_s^{-\beta} \Pi_s^m V_s(z) \\
&\leq c_{s_1}^\beta c_2^\beta c_s^{-\beta} (\lambda^m V_s(z) + b/(1-\lambda)) \\
&\leq c_1^{-\beta} c_2^\beta \lambda^m V(z) + c_{s_1}^\beta c_2^\beta c_s^{-\beta} b/(1-\lambda) \\
&\leq (c_2/c_1)^\beta \lambda^m V(z) + c_{s_1}^\beta c_2^\beta c_{\mathcal{K}}^{-\beta} b/(1-\lambda)
\end{aligned}$$

where $c_{\mathcal{K}} = \min_{s\in\mathcal{K}} c_s > 0$. Choosing m large enough such that $\lambda_m = (c_2/c_1)^\beta \lambda^m < 1$ ensures the result.

The geometric ergodicity of the Markov chain also ensures the existence of a solution of the Poisson equation (see Meyn and Tweedie (1993)). Thus, assumption (**A3"**) implies assumption (**A3'**) in the context of Anisotropic Metropolis Adjusted Langevin transition probability and the stated convergence result holds.

7.3. Proof of Theorem 2

We provide here the proof of the convergence of the estimated sequence generated by Algorithm 1.

We apply Theorem 1 with the functions H_s equals to $H_s(z) = S(z) - s$, $\pi_s = p_{\hat{\theta}(s)}$ and

$$h(s) = \int (S(z) - s) p_{\hat{\theta}(s)}(z) \mu(dz) \, .$$

Let us first prove that assumption **(A1')** is satisfied. Assumptions **(M1)-(M6)** ensure that $\mathcal{S}$ is an open subset and that the function h is continuous on $\mathcal{S}$. Moreover defining $w(s) = -l(\hat{\theta}(s))$, we get that w is continuously differentiable on $\mathcal{S}$. Applying Lemma 2 of Delyon et al (1999), we get **(A1')(i)**, **(A1')(iii)** and **(A1')(iv)**.

To prove **(A1')(ii)**, we consider as absorbing set $\mathcal{S}_a$ the closure of the convex hull of $S(\mathbb{R}^l)$ denoted $\overline{Conv}(S(\mathbb{R}^l))$. So assumption **(M7)(ii)** is exactly equivalent to assumption **(A1')(ii)**.

This achieves the proof of assumption **(A1')**.

Assumption **(A2')** is directly implied by assumption**(M1)**.

Let us now prove assumption **(A3")**.

Let $s_1 \in \mathcal{S}$ be fixed arbitrarily and define $V(z) = c_{s_1}^{\beta} p_{\hat{\theta}(s_1)}^{-\beta}(z)$, with $c_{s_1} = \max_{z \in \mathcal{X}} p_{\hat{\theta}(s_1)}(z)$. Let also $\mathcal{K}$ be a compact subset of $\mathcal{S}$.

We first consider condition (**A3"(i)**).

Since $H_s(z) = S(z) - s$, assumptions **(M8)** and **(B1)** ensure that there exists a constant $C > 0$ such that $\sup_{s \in \mathcal{K}} \sup_{z \in \mathcal{X}} \frac{P(z)}{p_{\hat{\theta}(s_1)}^{-\beta}(z)} \leq C$ and inequality (4.8) of (**A3"(i)**) holds.

The uniform ergodicity of the family of Markov chains corresponding to the AMALA on $\mathcal{K}$ ensures that there exist constants $0 < \gamma_{\mathcal{K}} < 1$ and $C_{\mathcal{K}} > 0$ such that for all $s \in \mathcal{K}$

$$\begin{aligned} \sup_{s \in \mathcal{K}} \|g_{\hat{\theta}(s)}\|_V &= \sup_{s \in \mathcal{K}} \| \sum_{k \geq 0} (\Pi_{\hat{\theta}(s)}^k H_s - p_{\hat{\theta}(s)} H_s) \|_V \\ &\leq \sup_{s \in \mathcal{K}} \sum_{k \geq 0} C_{\mathcal{K}} \gamma_{\mathcal{K}}^k \|H_s\|_V < \infty \, . \end{aligned}$$

Thus $\forall s \in \mathcal{K}$, $g_{\hat{\theta}(s)}$ belongs to $\mathcal{L}_V = \{g \ : \mathbb{R}^l \to \mathbb{R}, \|g\|_V < \infty\}$.

Repeating the same calculation as above, it is immediate that $\sup_{s \in \mathcal{K}} |||\Pi_{\hat{\theta}(s)} g_{\hat{\theta}(s)}||$ is bounded. This ends the proof of inequality (4.9) of (**A3"(i)**).

We now move to the Hölder condition (4.11) of (**A3"(i)**). We will use the following lemmas which state Hölder conditions on the transition kernel and its iterates:

Lemma 1. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. There exists a constant $C_{\mathcal{K}}$ such that for all $1 \leq p < q$, for all function $f \in \mathcal{L}_{V^p}$ and for all $(s, s') \in \mathcal{K}^2$ we have :*

$$\|\Pi_{\hat{\theta}(s)} f - \Pi_{\hat{\theta}(s')} f\|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \, \|s - s'\| \, .$$

Proof. For any $f \in \mathcal{L}_{V^p}$, we have

$$\Pi_\theta f(z) = \int_{\mathbb{R}^l} f(z')\alpha(z,z',\theta)q_c(z,z';\theta)dz' + f(z)(1-\alpha(z,\theta)) \ ,$$

where $\alpha(z,z',\theta) = \min\left(1, \frac{p_\theta(z')q_c(z',z;\theta)}{q_c(z,z';\theta)p_\theta(z)}\right)$ and $\alpha(z,\theta) = \int \alpha(z,z',\theta)q_c(z,z';\theta)dz'$ is the average acceptance rate.

Let s and s' be two points in $\mathcal{K}$ and $s(\epsilon) = (1-\epsilon)s + \epsilon s'$ for $\epsilon \in [0,1]$ be a linear interpolation between s and s' (since $\mathcal{S}$ is convex, we can assume that $\mathcal{K}$ is a convex set so that $s(\epsilon) \in \mathcal{K}$ for any $\epsilon \in [0,1]$). We denote also by $\theta(\epsilon) \triangleq \hat{\theta}(s(\epsilon))$ the associated path in Θ which is a continuously differentiable function.

We introduce $\Pi^1_\theta f(z) \triangleq (1-\alpha(z,\theta))f(z)$ and $\Pi^2_\theta f(z) \triangleq \int_{\mathbb{R}^d} f(z')\alpha(z,z',\theta)q_c(z,z',\theta)dz'$. We start with the difference $|(\Pi^2_{\theta(1)}f - \Pi^2_{\theta(0)}f)(z)|$.

We have:

$$|(\Pi^2_{\theta(1)}f - \Pi^2_{\theta(0)}f)(z)| \leq \int_0^1 \int_{\mathbb{R}^l} |f(z')| \times \left|\frac{d}{d\epsilon}\left(\alpha(z,z',\theta(\epsilon))q_c(z,z';\theta(\epsilon))\right)\right| dz'd\epsilon \, .$$

Since $\alpha(z,z',\theta) = \min(\rho(z,z',\theta),1)$ where $\rho(z,z',\theta) \triangleq \frac{p_\theta(z')q_c(z',z;\theta)}{p_\theta(z)q_c(z,z';\theta)}$ is a smooth function in θ, we have almost every where:

$$\left|\frac{d}{d\epsilon}\left(\alpha(z,z',\theta(\epsilon))q_c(z,z';\theta(\epsilon))\right)\right| \leq q_c(z,z';\theta(\epsilon)) \times \left|\frac{d}{d\epsilon}\left(\log \pi_{\theta(\epsilon)}(z') - \log \pi_{\theta(\epsilon)}(z) + \ \log q_c(z,z';\theta(\epsilon))\right)\right|$$

$$\begin{aligned} &\leq \quad C_{\mathcal{K}} q_c(z,z';\theta(\epsilon))\|\frac{d}{d\epsilon}\nabla_z \log \pi_{\theta(\epsilon)}(z)\| \times \left[1 + \|S(z') - S(z)\|\|\frac{d}{d\epsilon}\phi(\theta(\epsilon))\| + \|z' - z - \delta D(z,\theta(\epsilon))\|^2\right] \\ &\leq \quad C_{\mathcal{K}} q_c(z,z';\theta(\epsilon))R(z,z') \, , \end{aligned}$$

where R is a polynomial function of both variables z and z', with degree 2 in z' (thanks to assumption (**M8**) and (**B3**)).

Since $|f(z')| \leq \|f\|_{V^p} V^p(z')$, there exists an $C_{\mathcal{K}}$ such that for all $(s,s') \in \mathcal{K}^2$, for all $z \in \mathbb{R}^l$,

$$\int_0^1 \int_{\mathbb{R}^l} |f(z')| \, q_c(z,z';\theta(\epsilon))R(z,z')dz'd\epsilon \leq C_{\mathcal{K}}\|f\|_{V^p} \int_0^1 \int_{\mathbb{R}^l} V^p(z')R(z,z')q_c(z,z';\theta(\epsilon))dz'd\epsilon \, .$$

Thanks to assumption **(M8)** and the normality of q_c, we have for all $0 < \beta < 1$ and $1 \leq p$ satisfying $0 < \beta p < 1$:

$$\sup_{s\in\mathcal{K}} \int_{\mathbb{R}^l} V^p(z')R(z,z')q_c(z,z';\hat{\theta}(s))dz' \leq C_{\mathcal{K}}R(z) \, ,$$

where R is a polynomial function.

Therefore we end up with:

$$|(\Pi^2_{\theta(1)}f - \Pi^2_{\theta(0)}f)(z)| \quad \leq \quad C_{\mathcal{K}}\|f\|_{V^p}\|s - s'\|R(z) \, .$$

Moreover, thanks to the super exponential property (**B1**) of p_θ, we have that for all $q > 0$:

$$\sup_{z\in\mathcal{X}} \frac{R(z)}{V^q(z)} < \infty \, .$$

This implies that: for all $f \in \mathcal{L}_{V^p}$,

$$\|(\Pi^2_{\hat{\theta}(s)}f - \Pi^2_{\hat{\theta}(s')}f)\|_{V^q} \quad \leq \quad C_{\mathcal{K}}\|f\|_{V^p}\|s - s'\| \, .$$

Now, looking at the first term,

$$\begin{aligned} |(\Pi^1_{\theta(1)} f - \Pi^1_{\theta(0)} f)(z)| &= |\alpha(z,\theta(1)) - \alpha(z,\theta(0))| \, |f(z)| \\ &= |(\Pi^2_{\theta(1)} \mathbf{1} - \Pi^2_{\theta(0)} \mathbf{1})(z)| \, |f(z)| \\ &\leq C_{\mathcal{K}} \|\mathbf{1}\|_{V^p} \|s - s'\| R(z) \, |f(z)| \,, \end{aligned}$$

where we define $\mathbf{1}$ to be the constant function equals to 1 everywhere.

Following the same lines as previously, we get

$$|(\Pi^1_{\theta(1)} f - \Pi^1_{\theta(0)} f)(z)| \leq C_{\mathcal{K}} \|\|s' - s\| \, R(z) \|f\|_{V^p} V^p(z) \,. \tag{7.18}$$

Since R is polynomial, we obtain for all $q > p$ and for all $f \in \mathcal{L}_{V^p}$,

$$\|(\Pi^1_{\hat\theta(s)} f - \Pi^1_{\hat\theta(s')} f)\|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\| \,.$$

Thus, adding both terms, we get (updating again $C_{\mathcal{K}}$) that

$$\|(\Pi_{\theta(1)} - \Pi_{\theta(0)}) f\|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s' - s\| \,, \tag{7.19}$$

which ends the proof of the lemma. □

Lemma 2. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. There exists a constant $C_{\mathcal{K}}$ such that for all $1 \leq p < q$, for all function $f \in \mathcal{L}_{V^p}$, for all $(s, s') \in \mathcal{K}^2$ and for all $k \geq 0$, we have:*

$$\|\Pi^k_{\hat\theta(s)} f - \Pi^k_{\hat\theta(s')} f\|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\| \,.$$

Proof. We use the usual decomposition of the difference denoting $\theta = \hat\theta(s)$ and $\theta' = \hat\theta(s')$:

$$\Pi^k_\theta f - \Pi^k_{\theta'} f = \sum_{i=1}^{k-1} \Pi^i_\theta (\Pi_\theta - \Pi_{\theta'})(\Pi^{k-i-1}_{\theta'} f - p_{\theta'}(f)) \,.$$

Using Lemma 1, the fact that $\|\Pi^k_\theta (f - p_\theta(f))\|_{V^p} \leq C_{\mathcal{K}} \gamma^k_{\mathcal{K}} \|f\|_{V^p}$ with $\gamma_{\mathcal{K}} < 1$ (geometric ergodicity) and $\sup_{j \geq 0} \sup_{s \in \mathcal{K}} \|\Pi^j_{\hat\theta(s)} V^q\|_{V^q} < \infty$ we get:

$$\begin{aligned} \|\Pi^k_\theta f - \Pi^k_{\theta'} f\|_{V^q} &\leq C_{\mathcal{K}} \sum_{i=1}^{k-1} \|(\Pi_\theta - \Pi_{\theta'}) \times (\Pi^{k-i-1}_{\theta'} f - p_{\theta'}(f))\|_{V^q} \\ &\leq C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\| \sum_{i=1}^{k-1} \gamma^{k-i+1}_{\mathcal{K}} \end{aligned}$$

and the lemma is proved. □

Thanks to the proofs of Allassonnière et al (2010b), we get that h is a Hölder function for any $0 < a < 1$ which leads to (**A3"(i)**).

We finally focus on the proof of (**A3"(ii)**).

Lemma 3. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \geq 1$. There exists $C_{\mathcal{K}} > 0$ such that for any $s, s' \in \mathcal{K}$, for any $z \in \mathcal{X}$,*

$$|V^p_{\hat\theta(s)}(z) - V^p_{\hat\theta(s')}(z)| \leq C_{\mathcal{K}} \|s - s'\| V^p_{\hat\theta(s)}(z) P(z) \,. \tag{7.20}$$

Proof. We calculate first the differential in s of the function $p_{\hat{\theta}(s)}^{-\beta p}$. We get for all $z \in \mathcal{X}$:

$$\partial_s p_{\hat{\theta}(s)}^{-\beta p}(z) = (-p\ \beta)\ p_{\hat{\theta}(s)}^{-\beta p-1}(z)\partial_s p_{\hat{\theta}(s)}(z).$$

and

$$|\partial_s p_{\hat{\theta}(s)}^{-\beta p}(z)| \leq p\ \beta\ p_{\hat{\theta}(s)}^{-\beta p}(z)|\partial_s \log(p_{\hat{\theta}(s)}(z))|\,.$$

We have for all $z \in \mathcal{X}$:

$$p_{\hat{\theta}(s)}(z) = \frac{\exp\left(\psi(\hat{\theta}(s)) + \langle S(z), \phi(\hat{\theta}(s))\rangle\right)}{\int \exp\left(\psi(\hat{\theta}(s)) + \langle S(z), \phi(\hat{\theta}(s))\rangle\right) dz}\,,$$

which leads to

$$\begin{aligned}\partial_s \log(p_{\hat{\theta}(s)}(z)) = \partial_s\psi(\hat{\theta}(s)) + \langle S(z), \partial_s\phi(\hat{\theta}(s))\rangle \\ - \frac{\int \exp\left(\psi(\hat{\theta}(s)) + \langle S(z), \phi(\hat{\theta}(s))\rangle\right)}{\int \exp\left(\psi(\hat{\theta}(s)) + \langle S(z), \phi(\hat{\theta}(s))\rangle\right) dz} \times \\ \frac{\left(\partial_s\psi(\hat{\theta}(s)) + \langle S(z), \partial_s\phi(\hat{\theta}(s))\rangle\right) dz}{\int \exp\left(\psi(\hat{\theta}(s)) + \langle S(z), \phi(\hat{\theta}(s))\rangle\right) dz}\,. \quad (7.21)\end{aligned}$$

We can deduce from assumptions (**M2**), (**M5**) and (**M8**) that there exists a constant $C_{\mathcal{K}}$ such that: for all $s \in \mathcal{K}$

$$|\partial_s \log(p_{\hat{\theta}(s)}(z))| \leq C_{\mathcal{K}} P(z)\,,$$

which implies that

$$|\partial_s p_{\hat{\theta}(s)}^{-\beta p}(z)| \leq C_{\mathcal{K}}\ p\ \beta\ p_{\hat{\theta}(s)}^{-\beta p}(z)P(z).$$

Introducing again $0 \leq c_1 \leq c_2$ such that

$$c_1 p_{\hat{\theta}(s_1)}^{-\beta}(z) \leq p_{\hat{\theta}(s)}^{-\beta}(z) \leq c_2 p_{\hat{\theta}(s_1)}^{-\beta}(z)$$

for any $(z, s) \in \mathcal{X} \times \mathcal{K}$ we deduce that

$$|\partial_s p_{\hat{\theta}(s)}^{-\beta p}(z)| \leq c_2^p C_{\mathcal{K}}\ p\ \beta\ p_{\hat{\theta}(s_1)}^{-\beta p}(z)P(z).$$

Using the mean value theorem, we get

$$|p_{\hat{\theta}(s)}^{-\beta p}(z) - p_{\hat{\theta}(s')}^{-\beta p}(z)| \leq \frac{c_2^p}{c_1^p} C_{\mathcal{K}}\ p\ \beta \|s - s'\|\ p_{\hat{\theta}(s)}^{-\beta p}(z)P(z). \quad (7.22)$$

We recall that $V_{\hat{\theta}(s)}(z) = c_s^{\beta} p_{\hat{\theta}(s)}^{-\beta}(z)$. This yields

$$|V_{\hat{\theta}(s')}^p(z) - V_{\hat{\theta}(s)}^p(z)| \leq |c_{s'}^{\beta p}||p_{\hat{\theta}(s)}^{-\beta p}(z) - p_{\hat{\theta}(s')}^{-\beta p}(z)| + |p_{\hat{\theta}(s)}^{-\beta p}(z)|\ |c_s^{\beta p} - c_{s'}^{\beta p}|\,.$$

Since $c_s = \max_{z \in \mathcal{X}} p_{\hat{\theta}(s)}(z)$ and using the smoothness of the target distribution with respect to both the parameter and the hidden variable, we can bound c_s on any compact subset $\mathcal{K}$ by $C_{\mathcal{K}}$ and using Equation (7.22) we have:

$$|V_{\hat{\theta}(s')}^p(z) - V_{\hat{\theta}(s)}^p(z)| \leq C'_{\mathcal{K}}\|s - s'\|p_{\hat{\theta}(s)}^{-\beta p}(z)P(z) + p_{\hat{\theta}(s)}^{-\beta p}(z)|c_s^{\beta p} - c_{s'}^{\beta p}|\,.$$

We now focus on the estimation of $|c_s^{\beta p} - c_{s'}^{\beta p}|$.

Let us note that:

$$\begin{aligned} |c_s^{\beta p} - c_{s'}^{\beta p}| &= |\max_{z\in\mathcal{X}} p_{\hat\theta(s)}^{\beta p}(z) - \max_{z\in\mathcal{X}} p_{\hat\theta(s')}^{\beta p}(z)| \\ &\leq \max_{z\in\mathcal{X}} |p_{\hat\theta(s)}^{\beta p}(z) - p_{\hat\theta(s')}^{\beta p}(z)| \,. \end{aligned}$$

We then use the mean value theorem as before:

$$\begin{aligned} |\partial_s p_{\hat\theta(s)}^{\beta p}(z)| &\leq C_{\mathcal{K}}\ p\ \beta\ p_{\hat\theta(s)}^{\beta p}(z)P(z) \\ &\leq C_{\mathcal{K}} c_1^{-p}\ p\ \beta\ p_{\hat\theta(s_1)}^{\beta p}(z)P(z)\,, \end{aligned}$$

and get:

$$\begin{aligned} |p_{\hat\theta(s)}^{\beta p}(z) - p_{\hat\theta(s')}^{\beta p}(z)| &\leq C_{\mathcal{K}}\ p\ \beta\|s-s'\|\ p_{\hat\theta(s_1)}^{\beta p}(z)P(z) \\ &\leq C_{\mathcal{K}}\ p\ \beta\|s-s'\| c_{s_1}^{\beta p} P(z)\,. \end{aligned} \tag{7.23}$$

This yields :

$$|c_s^{\beta p} - c_{s'}^{\beta p}| \leq C_{\mathcal{K}} P(z)\|s-s'\|$$

and thus

$$\begin{aligned} |V_{\hat\theta(s')}^p(z) - V_{\hat\theta(s')}^p(z)| &\leq C'_{\mathcal{K}}\|s-s'\| p_{\hat\theta(s)}^{-\beta p}(z)P(z) \\ &\leq C'_{\mathcal{K}} c_{\mathcal{K}}^{-\beta p}\|s-s'\| V_{\hat\theta(s)}^p(z)P(z)\,, \end{aligned}$$

which concludes the proof of Lemma 3. □

Lemma 4. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$, $p \geq 1$ and P be a polynomial function on $\mathcal{X}$. There exists $\mu_{\mathcal{K}} > 0$ such that for any $z \in \mathcal{X}$, for any $s \in \mathcal{K}$, we have*

$$\Pi_{\hat\theta(s)} V_{\hat\theta(s)}^p P(z) \leq \mu_{\mathcal{K}} V_{\hat\theta(s)}^p(z)\,. \tag{7.24}$$

Proof. This proof is similar to the proof of Proposition 1 and in particular of Equation (7.3). The key property leads in an equivalent inequalities sequence given in Equation (7.2) including the polynomial function P. As before, we have : thanks to the bounded drift and covariance matrix, there exist two constants $0 < k_1 < k_2$ such that for all $(x,z) \in \mathcal{X}^2$

$$k_1 g_{\epsilon_1}(x-z) \leq q_c(x,z)P(z) \leq k_2 g_{\epsilon_2}(x-z)\,, \tag{7.25}$$

denoting by g_a the centered Gaussian probability density function in $\mathbb{R}^l$ with covariance matrix aId_l. Therefore we get, with the condition that $\beta p < 1$:

$$\frac{\Pi V^p P(x)}{V^p(x)} = \int_{A(x)} \underbrace{\frac{\pi(z)^{-\beta p}}{\pi(x)^{-\beta p}} q_c(x,z)P(z)}_{f_1(x,y)P(z)}\, dz + \int_{R(x)} \underbrace{\frac{\pi(z)^{1-\beta p}}{\pi(x)^{1-\beta p}} q_c(z,x)P(z)}_{f_2(x,z)P(z)}\, dz + \int_{R(x)} \underbrace{q_c(x,z)P(z)}_{f_3(x,z)P(z)}\, dz\,.$$

Equation (7.25) enables to bound the two first elements as already done and thus they converge toward 0 as $|x| \to \infty$. Considering now the last term, we can use the same bound :

$$\int_{R(x)} q_c(x,z)P(z)dz \leq \int_{R(x)} k_2 g_{\epsilon_2}(x-z)dz \leq k_2\,.$$

Then, taking the lim sup over $|x|$ tending to infinity provides the result.

□

Lemma 5. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \geq 1$ such that $p\beta < 1$. For all sequences $\boldsymbol{\gamma} = (\gamma_k)_{k\geq 0}$ and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ satisfying $\varepsilon_k < \bar\epsilon$ for some $\bar\epsilon$ sufficiently small, there exists $C_{\mathcal{K}} > 0$, such that for any $z \in \mathcal{X}$, we have*

$$\sup_{s\in\mathcal{K}} \sup_{k\geq 0} \mathbb{E}_{z,s}^{\boldsymbol{\gamma}}[V^p(z_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] \leq C_{\mathcal{K}} V^p(z)\,.$$

Proof. Let K be a compact subset of Θ such that $\hat{\theta}(\mathcal{K}) \subset K$. We note in the sequel, $\theta_k = \hat{\theta}(s_k)$. We have for $k \geq 2$, using the Markov property and the drift property (7.17) for V_θ^p,

$$\begin{aligned}
\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &\leq \mathbb{E}_{z,s}^{\gamma}[\Pi_{\theta_{k-1}}V_{\theta_{k-1}}^p(z_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &(7.26)\\
&\leq \lambda\mathbb{E}_{z,s}^{\gamma}[(V_{\theta_{k-1}}^p(z_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] + C &(7.27)
\end{aligned}$$

Moreover, with the same arguments and Lemma 3,

$$\begin{aligned}
\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &= \mathbb{E}_{z,s}^{\gamma}[(V_{\theta_{k-1}}^p(z_{k-1}) - V_{\theta_{k-2}}^p(z_{k-1}))\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\\
&\quad +\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\\
&\leq C_{\mathcal{K}}\varepsilon_{k-1}\mathbb{E}_{z,s}^{\gamma}[\Pi_{\theta_{k-2}}V_{\theta_{k-2}}^p P(z_{k-2})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\\
&\quad +\mathbb{E}_{z,s}^{\gamma}[\Pi_{\theta_{k-2}}V_{\theta_{k-2}}^p\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\\
&\leq C_{\mathcal{K}}\varepsilon_{k-1}\mu_{\mathcal{K}}\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-2}}^p(z_{k-2})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\\
&\quad +\lambda\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-2}}^p(z_{k-2})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] + C\\
&\leq C + \lambda(\lambda + C_{\mathcal{K}}\varepsilon_{k-1}\mu_{\mathcal{K}}) \times \mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-2}}^p(z_{k-2})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\,.
\end{aligned}$$

Iterating this recursively leads to :

$$\begin{aligned}
\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &\leq C\left[1 + \sum_{j=2}^{k-1}\prod_{l=j}^{k-1}(\lambda + C_{\mathcal{K}}\varepsilon_l\mu_{\mathcal{K}})\right] + &(7.28)\\
&\quad \left[\prod_{j=1}^{k-1}(\lambda + C_{\mathcal{K}}\varepsilon_j\mu_{\mathcal{K}})\right]\mathbb{E}_{z,s}^{\gamma}[V_{\theta_0}^p(z)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}]\,, &(7.29)
\end{aligned}$$

which can be injected in Equation (7.26) and yields :

$$\begin{aligned}
\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &\leq \lambda\left[\prod_{j=1}^{k-1}(\lambda + C_{\mathcal{K}}\varepsilon_j\mu_{\mathcal{K}})\right]\mathbb{E}_{z,s}^{\gamma}\left[V_{\theta_0}^p(z)\right] +\\
&\quad C\left[1 + \lambda\sum_{j=2}^{k-1}\prod_{l=j}^{k-1}(\lambda + C_{\mathcal{K}}\varepsilon_l\mu_{\mathcal{K}})\right].
\end{aligned}$$

Since $\lambda < 1$, we can choose $\bar{\varepsilon}$ such that $\rho = \lambda + C_{\mathcal{K}}\bar{\varepsilon}\mu_{\mathcal{K}} < 1$ which ensures that for all $k \in \mathbb{N}$, we have :

$$\begin{aligned}
\mathbb{E}_{z,s}^{\gamma}[V_{\theta_{k-1}}^p(z_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}] &\leq \rho^{k-1}V_{\theta_0}^p(z) + C\sum_{l=0}^{k-2}\rho^l\\
&\leq V_{\theta_0}^p(z)\left(1 + \frac{C}{1-\rho}\right).
\end{aligned}$$

Using $0 \leq c_1 \leq c_2$ such that $c_1 p_{\hat{\theta}(s_1)}^{-\beta}(z) \leq p_\theta^{-\beta}(z) \leq c_2 p_{\hat{\theta}(s_1)}^{-\beta}(z)$ for any $(z,\theta) \in \mathbb{R}^l \times K$ and the definitions of V and V_θ we get the result. □

This yields (**A3"(ii)**) under (**A4**) and allows finally to apply Theorem 1.

7.4. Proof of the Central Limit Theorem for the Estimated Sequence

To prove this result we consider Theorem 24 of Delyon (2000): Consider the stochastic approximation with a noise of the form:

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k\xi_k + \nu_k - \nu_{k-1} + r_k \tag{7.30}$$

and assumption **(C)** given by

(C) The algorithm defined by Equation (7.30) is A-stable, the sequence (s_k) converges toward some limit s^*. The function h is C^1 in some neighborhood of s^* with first derivatives Lipschitz and J the Jacobean matrix of the mean field h in s^* has all its eigenvalues with negative real part.

Theorem 4 (Theorem 24 Delyon (2000))**.** *Let assumptions* ***(C)*** *and* ***(N3)*** *be satisfied, and if* $\lim_{k\to\infty} k\gamma_k = 1$ *we require that* $\gamma_k = 1/k$. *Furthermore assume that for some matrix* U, *some* $\varepsilon > 0$ *and some positive random variables* X, X', X'':

$$\text{The sequence } (\xi_i) \text{ is a } \mathcal{F} - \text{martingale} \tag{7.31}$$

$$\sup_{i\in\mathbb{N}} \|\xi_i\|_{2+\varepsilon} < \infty \tag{7.32}$$

$$\lim_{k\to\infty} \gamma_k^{-1/2} \|X r_k\|_1 = 0 \tag{7.33}$$

$$\lim_{k\to\infty} \gamma_k^{1/2} \|X' \nu_k\|_1 = 0 \tag{7.34}$$

$$\lim_{k\to\infty} \gamma_k \|X'' \sum_{i=1}^{k} (\xi_i \xi_i' - U)\|_1 = 0 \tag{7.35}$$

where $\mathcal{F} = (\mathcal{F}_i)_{i\in\mathbb{N}}$ *is the increasing family of* $\sigma-$*algebra generated by the random variables* $(s_0, z_1, ..., z_i)$. *Then*

$$\frac{s_k - s^*}{\sqrt{\gamma_k}} \to_{\mathcal{L}} \mathcal{N}(0, V) \tag{7.36}$$

where V *is the solution of the following Lyapunov equation depending on the cases of assumption* ***(N3)****:*

Case 1: $U + JV + VJ' = 0$,
Case 2: $U + (J + I/2)V + V(J + I/2)' = 0$.

Unfortunately, we are not able to prove the A-stability required in assumption **(C)** and to prove our convergence we use a relaxed version of this condition. This new assumption is a sufficient condition to prove the required result as can be seen when following the proofs of Delyon (2000). We therefore define a new assumption **(C')**.

(C') The sequence s_k converges to some limit s^*, the function h is C^1 in some neighborhood of s^* with first derivatives Lipschitz and J the Jacobean matrix of the mean field h in s^* has all its eigenvalues with negative real part.

The result of Theorem 24 still holds under this alleviated assumption. Indeed, it is sufficient to establish that the random variable $\gamma_k^{-1/2} \sum_{i=0}^{k} \exp((t_k - t_i)J)\gamma_i r_i$ converges toward 0 in probability where $t_i = \sum_{j=1}^{i} \gamma_j$. Theorem 19 and Proposition 39 of Delyon (2000) can be applied in expectation. Theorems 23 and 20 of Delyon (2000) also still hold with assumption **(C')** instead of **(C)**.

We have now to prove that the assumptions of Theorem 24 of Delyon (2000) replacing **(C)** by **(C')**.

We decompose the remainder term as follows:

$$\eta_k = \xi_k + \nu_k - \nu_{k-1} + r_k \tag{7.37}$$

with

$$\xi_k = g_{s_{k-1}}(z_k) - \Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1}) \tag{7.38}$$

$$\nu_k = -\Pi_{s_k} g_{s_k}(z_k) \tag{7.39}$$

$$r_k = \Pi_{s_k} g_{s_k}(z_k) - \Pi_{s_{k-1}} g_{s_{k-1}}(z_k) \tag{7.40}$$

where for any $s \in \mathcal{S}$, g_s is a solution of the Poisson equation $g - \Pi_s g = H_s - p_s(H_s)$.

Thanks to the assumptions of Theorem 2 which provides the convergence of the sequence (s_k) and assumption **(N1)** we get **(C')**.

By definition of ξ_i it is obvious that (7.31) is true. Moreover, the following lemma proves that there exists $\varepsilon > 0$ such that (7.32) holds with $X = 1$.

Lemma 6. *For all $\varepsilon \geq 0$ such that $\beta(2+\varepsilon) < 1$, the sequence (ξ_k) is in $L^{2+\varepsilon}$.*

Proof. We use the convexity of the function $x \mapsto x^{2+\delta}$. Indeed, we have

$$\begin{aligned} |g_{s_{k-1}}(z_k) - \Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1})|^{2+\varepsilon} &\leq (|g_{s_{k-1}}(z_k)| + |\Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1})|)^{2+\varepsilon} \\ &\leq C_\varepsilon (|g_{s_{k-1}}(z_k)|^{2+\varepsilon} + |\Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1})|^{2+\varepsilon}) , \end{aligned}$$

where $C_\varepsilon = \frac{1}{2^{3+\varepsilon}}$.

Using assumptions **(M8)** and **(B1)**, we get applying the drift condition:

$$\begin{aligned} \mathbb{E}(||\xi_k||^{2+\varepsilon} | \mathcal{F}_{k-1}) &\leq C_\varepsilon \left(\mathbb{E}(|g_{s_{k-1}}(z_k)|^{2+\varepsilon} \mid \mathcal{F}_{k-1}) + |\mathbb{E}(\Pi_{s_{k-1}} g_{s_{k-1}}(z_{k-1}) | \mathcal{F}_{k-1})|^{2+\varepsilon} \right) \\ &\leq C \, \mathbb{E}(V(z_k)^{2+\varepsilon} + V(z_{k-1})^{2+\varepsilon} | \mathcal{F}_{k-1})) \leq C \left(\lambda V^{2+\varepsilon}(z_{k-1}) + 1 \right) . \end{aligned}$$

Finally taking the expectation after induction leads to:

$$\mathbb{E}(||\xi_k^{2+\varepsilon}||) \leq C V^{2+\varepsilon}(z_0) < +\infty .$$

□

Let us now focus on Equation (7.33). Thanks to the Hölder property of our kernel and the fact that H_{s_k} belongs to $\mathcal{L}_V$:

$$\begin{aligned} \|r_n\|_1 &= \mathbb{E}[|\Pi_{s_n} g_{s_n}(z_n) - \Pi_{s_{n-1}} g_{s_{n-1}}(z_n)|] \\ &\leq C \mathbb{E}[V^q(z_n) |s_n - s_{n-1}|^a] \\ &\leq C \mathbb{E}[V^{q+1}(z_n)] \gamma_n^a \\ &\leq C \gamma_n^a \end{aligned}$$

where the last inequality comes from the drift property. Since the Hölder property is true for any $0 < a < 1$, we can choose $a > 1/2$ which leads to the conclusion.

To prove Equation (7.34), we note that using the drift condition as in the previous lemma, $\mathbb{E}[|\nu_k|]$ is uniformly bounded in k. Since the step-size sequence (γ_k) tends to zero, the result follows with $X'' = 1$.

Equation (7.35) is an assumption of our asymptotic normality theorem and Theorem 24 applies. The Delta method enables to get the result on the sequence (θ_k) achieving the proof of our Central Limit theorem.

References

Aldridge G, Ratnanather J, Martone M, Terada M, Beg M, Fong L, Ceyhan E, Kolasny A, Brown T, Cochran E, Tang S, Pisano D, Vaillant M, Hurdal M, Churchill J, Greenough W, Miller M, Ellisman M (2005) Semi-automated shape analysis of dendrite spines from animal models of fragilex and parkinson's disease using large deformation diffeomorphic metric mapping. Society for Neuroscience Annual Meeting, Washington DC

Allassonnière S, Kuhn E (2010) Stochastic algorithm for bayesian mixture effect template estimation. ESAIM Probab Stat 14:382–408

Allassonnière S, Amit Y, Trouvé A (2007) Toward a coherent statistical framework for dense deformable template estimation. JRSS 69:3–29

Allassonnière S, Kuhn E, Trouvé A (2010a) Bayesian consistent estimation in deformable models using stochastic algorithms: Applications to medical images. Journal de la Société Française de Statistique 151(1):1–16

Allassonnière S, Kuhn E, Trouvé A (2010b) Bayesian deformable models building via stochastic approximation algorithm: A convergence study. Bernoulli J 16(3):641–678

Andrieu C, Moulines E, Priouret P (2005) Stability of stochastic approximation under verifiable conditions. SIAM J Control Optim 44(1):283–312 (electronic)

Atchadé Y (2006) An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. Methodol Comput Appl Probab 8:235–254

Bigot J, Charlier B (2011) On the consistency of Fréchet means in deformable models for curve and image analysis. Electron J Stat 5:1054–1089, , URL http://dx.doi.org/10.1214/11-EJS633

Ceyhan E, Fong L, Tasky T, Hurdal M, Beg M MFand Martone, Ratnanather J (2007a) Type-specific analysis of morphometry of dendrite spines of mice. 5th Int Symp Image Signal Proc Analysis, ISPA pp 7–12

Ceyhan E, Ölken R, Fong L, Tasky T, Hurdal M, Beg M, Martone M, Ratnanather J (2007b) Modeling metric distances of dendrite spines of mice based on morphometric measures. Int Symp on Health Informatics and Bioinformatics

Cootes T, Taylor C, Cooper D, Graham J (1995) Active shape models: their training and application. Comp Vis and Image Understanding 61(1):38–59

Delyon B (2000) Stochastic approximation with decreasing gain: convergence and asymptotic theory. Technical Report: Publication interne 952, IRISA

Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of the EM algorithm. Ann Statist 27(1):94–128

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 1:1–22

Gilks W, Richardson S, Spiegelhalter D (1996) Markov Chain Monte Carlo in Practice. Chapman & Hall

Girolami M, Calderhead B (2011) Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B 73(2):1–37

Glasbey CA, Mardia KV (2001) A penalised likelihood approach to image warping. Journal of the Royal Statistical Society, Series B 63:465–492

Grenander U, Miller MI (1998) Computational anatomy: An emerging discipline. Quarterly of Applied Mathematics LVI(4):617–694

Jarner S, Hansen E (1998) Geometric ergodicity of metropolis algorithms. In: Stochastic Processes and Their Applications, pp 341–361

Joshi S, Davis B, Jomier M, Gerig G (2004) Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23:S151–S160

Kuhn E, Lavielle M (2004) Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM Probab Stat 8:115–131 (electronic)

Maire F, Lefebvre S, Moulines E, Douc R (2011) Aircraft classification with low infrared sensor. Statistical Signal Processing Workshop (SSP), IEEE

Marshall T, Roberts G (2012) An adaptive approach to langevin MCMC. Statistics and Computing 22 (5):1041–1057

Marsland S, Twining C (2004) Constructing diffeomorphic representations for the groupewise analysis of non-rigid registrations of medical images. IEEE Transactions on Medical Imaging 23

Meyn SP, Tweedie RL (1993) Markov chains and stochastic stability. Communications and Control Engineering Series, Springer-Verlag London Ltd., London

Micheli M, Michor PW, Mumford DB (2012) Sectional curvature in terms of the cometric, with applications to the riemannian manifolds of landmarks. SIAM Journal on Imaging Sciences 5(1):394–433

Miller M, Priebe C, Qiu A, Fischl B, Kolasny A, Brown T, Park Y, Ratnanather J, Busa E, Jovicich J, Yu P, Dickerson B, Buckner R (2009) Morphometry BIRN. collaborative computational

anatomy: An MRI morphometry study of the human brain via diffeomorphic metric mapping. Human Brain Mapping 30(7):2132–2141

Miller MI, Trouvé A, Younes L (2002) On the metrics and Euler-Lagrange equations of computational anatomy. Annual Review of biomedical Engineering 4

Richard F, Samson A, Cuenod CA (2009) A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. Statistics and Computing 19:465–478

Roberts GO, Tweedie RL (1996) Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2(4):341–363, , URL http://dx.doi.org/10.2307/3318418

Stramer O, Tweedie R (1999a) Langevin-type models i: Diffusions with given stationary distributions, and their discretizations. Methodol Comput Appl Probab 1(3):283–306

Stramer O, Tweedie R (1999b) Langevin-type models ii: self-targeting candidates for mcmc algorithms. Methodol Comput Appl Probab 1(3):307–328

Vercauteren T, Pennec X, Perchant A, Ayache N (2009) Diffeomorphic demons: Efficient non-parametric image registration. Neuroimage 45:61–72