ESAIM: PS December 2010, Vol. 14, p. 382–408
DOI: 10.1051/ps/2009001

ESAIM: Probability and Statistics
www.esaim-ps.org

STOCHASTIC ALGORITHM FOR BAYESIAN MIXTURE EFFECT TEMPLATE ESTIMATION

Stéphanie Allassonnière[1] and Estelle Kuhn[2]

Abstract. The estimation of probabilistic deformable template models in computer vision or of probabilistic atlases in Computational Anatomy are core issues in both fields. A first coherent statistical framework where the geometrical variability is modelled as a hidden random variable has been given by [S. Allassonnière *et al.*, *J. Roy. Stat. Soc.* **69** (2007) 3–29]. They introduce a Bayesian approach and mixture of them to estimate deformable template models. A consistent stochastic algorithm has been introduced in [S. Allassonnière *et al.* (in revision)] to face the problem encountered in [S. Allassonnière *et al.*, *J. Roy. Stat. Soc.* **69** (2007) 3–29] for the convergence of the estimation algorithm for the one component model in the presence of noise. We propose here to go on in this direction of using some "SAEM-like" algorithm to approximate the MAP estimator in the general Bayesian setting of mixture of deformable template models. We also prove the convergence of our algorithm toward a critical point of the penalised likelihood of the observations and illustrate this with handwritten digit images and medical images.

Mathematics Subject Classification. 60J22, 62F10, 62F15, 62M40.

Received June 3, 2008. Revised January 16, 2009 and March 19, 2009.

1. Introduction

The issue of representing and analysing some geometrical structures upon which some deformations can act is a challenging question in applied mathematics as well as in Computational Anatomy. One central point is the modelisation of varying objects, and the quantification of this variability with respect to one or several reference models which will be called templates. This is known as "Deformable Templates" [12]. To our best knowledge, the problem of constructing probabilistic models of variable shapes in order to statistically quantify this variability has not been successfully addressed yet in spite of its importance. For example, modelling the anatomical variability of organs around an ideal shape is of a crucial interest in the medical domain in order to find some characteristic differences between populations (pathological and control), or to exhibit some pathological kind of deformations or shapes of an organ.

Many solutions have been proposed to face the problem of the template definition. They go from some generalised Procruste's means with a variational [11] or statistical [10] point of view to some statistical models

Keywords and phrases. Stochastic approximations, non rigid-deformable templates, shapes statistics, MAP estimation, Bayesian method, mixture models.

[1] CMAP - École polytechnique, Route de Saclay, 91128 Palaiseau, France; Allassoniere@gmail.com

[2] LAGA - Université Paris 13, 99 av. J.-B. Clément, 93430 Villetaneuse, France and INRA - Unité MIA, Domaine de Vilvert, 78352 Jouy-en-Josas, France.

Article published by EDP Sciences

like Active Appearance Model [6] or Minimum Description Length methods [16]. Unfortunately, all these methods only focus on the template whereas the geometrical variability is computed afterwards (using PCA). This contradicts with the fact that a metric is required to compute the template through the computation of deformations. Moreover, they do not really differ from the variational point of view since they consider the deformations as some nuisance parameters which have to be estimated and not as some unobserved random variables.

The main goal of this paper is to propose a coherent estimation of both photometric model and geometrical distribution in a given population. Another issue addressed here is the clustering problem. Given a set of images, the statistical estimation of the component weights and of the image labels is usually supervised, at least the number of components is fixed. The templates of each component and the label are estimated iteratively (for example in methods like K-means) but the geometry, and related to this the metric used to compute the distances between elements, is still fixed. Moreover, the label, which is not observed, is, as the deformations, considered as a parameter and not as a hidden random variable. These methods do not lead to a statistical coherent framework for the understanding of deformable template estimation and all these iterative algorithms derived from those approaches do not have a statistical interpretation as the parameter optimisation of a generative model describing the data.

In this paper we consider the statistical framework for dense deformable templates developed by Allassonnière *et al.* in [1] in the generalised case of mixture model for multicomponent estimation. Each image taken from a database is supposed to be generated from a noisy random deformation of a template image picked randomly among a given set of possible templates. All the templates are assumed to be drawn from a common prior distribution on the template image space. To propose a generative model, each deformation and each image label have to be considered as *hidden* variables. The templates, the parameters of the deformation laws and the components weights are the parameters of interest. This generative model allows to automatically decompose the database into components and, at the same time, estimates the parameters corresponding to each component while increasing the likelihood of the observations.

Given this parametric statistical Bayesian model, the parameter estimation is performed in [1] by a Maximum A Posteriori (MAP). The authors carry out this estimation problem using a deterministic and iterative scheme based on the EM (Expectation Maximisation) algorithm where the posterior distribution is approximated by a Dirac measure on its mode. Unfortunately, this gives an algorithm whose convergence toward the MAP estimator cannot be proved. Moreover, as shown by the experiments in that paper, the convergence is lost within a noisy setting.

Our goal in this paper is to propose some stochastic iterative method to reach the MAP estimator for which we will be able to get a convergence result as already done for the one component case in [3]. We propose to use a stochastic version of the EM algorithm to reach the maximum of the posterior distribution.We use the Stochastic Approximation EM (SAEM) algorithm introduced by Delyon *et al.* in [7] coupled with a Monte Carlo Markov Chain (MCMC) method. This coupling algorithm has been introduced by Kuhn and Lavielle in [14] in the case where the missing variables had a compact support. Contrary to the one component model where we can couple the iteration of the SAEM algorithm with the Markov chain evolution (*cf.* [3]), we show here that it cannot be driven numerically. We need to consider an alternative method. We propose to simulate the hidden variables using some auxiliary Markov chains, one per component, to approach the posterior distribution. We prove the convergence of our algorithm for a non compact setting by adapting Delyon's theorem about general stochastic approximations and introducing truncation on random boundaries as in [5].

The paper is organised as follows: in Section 2 we first recall the observation mixture model proposed by Allassonnière *et al.* in [1]. In Section 3, we describe the stochastic algorithm used in our particular setting. Section 4 is devoted to the convergence theorem. Illustrative experiments on 2D real data sets are presented in Section 5. The proofs of the convergence of the algorithm are postponed in Section 6 whereas conclusion and discussion are given in Section 7.

2. The observation model

We consider the framework of multicomponent model introduced in [1]. Given a sample of gray level images $(y_i)_{1\leq i\leq n}$ observed on a grid of pixels $\{v_u \in D \subset \mathbb{R}^2, u \in \Lambda\}$ where D is a continuous domain and Λ the pixel network, we are looking for some template images which explain the population. Each of these images is a real function $I_0 : \mathbb{R}^2 \to \mathbb{R}$ defined on the whole plane. An observation y is supposed to be a discretisation on Λ of a deformation of one of the templates plus an independent additive noise. This leads to assume the existence of an unobserved deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that for $u \in \Lambda$:

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

where ϵ denotes an additive noise.

2.1. Models for templates and deformations

We use the same framework as chosen in [1] to describe both the templates I_0 and the deformation fields z. Our model takes into account two complementary sides: photometric – indexed by p – corresponding to the templates and the noise variances, and geometric – indexed by g – corresponding to the deformations. The templates I_0 and the deformations z are assumed to belong to some finite dimensional subspaces of two reproducing kernels Hilbert spaces V_p and V_g (determined by their respective kernels K_p and K_g). We choose a representation of both of them by finite linear combinations of the kernels centred at some fixed landmark points in the domain D: $(v_{p,j})_{1\leq j\leq k_p}$ respectively $(v_{g,j})_{1\leq j\leq k_g}$. They are therefore parametrised by the coefficients $\alpha \in \mathbb{R}^{k_p}$ and $\beta \in (\mathbb{R}^{k_g})^2$ which yield: $\forall v \in D$,

$$\begin{aligned} I_\alpha(v) &\triangleq (\mathbf{K_p}\alpha)(v) \triangleq \sum_{j=1}^{k_p} K_p(v, v_{p,j})\alpha^j \,, \\ z_\beta(v) &\triangleq (\mathbf{K_g}\beta)(v) \triangleq \sum_{j=1}^{k_g} K_g(v, v_{g,j})\beta^j. \end{aligned}$$

2.2. Parametrical model

In this paper, we consider a mixture of the deformable template models which allows a fixed number τ_m of components in each training set. This means that the data will be separated in τ_m (at most) different components by the algorithm.

Therefore, for each observation y_i, we consider the pair (β_i, τ_i) of unobserved variables which correspond respectively to the deformation field and to the label of image i. We denote below by $\boldsymbol{y}^t \triangleq (y_1^t, \ldots, y_n^t)$, by $\boldsymbol{\beta}^t \triangleq (\beta_1^t, \ldots, \beta_n^t)$ and by $\boldsymbol{\tau}^t \triangleq (\tau_1, \ldots, \tau_n)$. The generative model is:

$$\left\{ \begin{array}{l} \boldsymbol{\tau} \sim \otimes_{i=1}^n \sum_{t=1}^{\tau_m} \rho_t \delta_t \mid (\rho_t)_{1\leq t\leq \tau_m} \,, \\ \boldsymbol{\beta} \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_{g,\tau_i}) \mid \boldsymbol{\tau},\ (\Gamma_{g,t})_{1\leq t\leq \tau_m}, \\ \boldsymbol{y} \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_{\alpha_{\tau_i}}, \sigma^2_{\tau_i} Id_{|\Lambda|}) \mid \boldsymbol{\beta},\ \boldsymbol{\tau},\ (\alpha_t, \sigma_t^2)_{1\leq t\leq \tau_m}, \end{array} \right. \tag{2.1}$$

where $z_\beta I_\alpha(u) = I_\alpha(v_u - z_\beta(v_u))$ is the action of the deformation on the template I_α, for u in Λ and δ_t is the Dirac function on t. The parameters of interest are the vectors $(\alpha_t)_{1\leq t\leq \tau_m}$ coding the templates, the variances $(\sigma_t^2)_{1\leq t\leq \tau_m}$ of the additive noises, the covariance matrices $(\Gamma_{g,t})_{1\leq t\leq \tau_m}$ of the deformation fields and the component weights $(\rho_t)_{1\leq t\leq \tau_m}$. We denote by $(\theta_t, \rho_t)_{1\leq t\leq \tau_m}$ the parameters so that θ_t corresponds to the parameters composed of the photometric part (α_t, σ_t^2) and the geometric part $\Gamma_{g,t}$ for component t.

We assume that for all $1 \leq t \leq \tau_m$, the parameter $\theta_t = (\alpha_t, \sigma_t^2, \Gamma_{g,t})$ belongs to the open space Θ defined as $\Theta = \{ (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p},\ |\alpha| < R,\ \sigma > 0,\ \Gamma_g \in \mathrm{Sym}^+_{2k_g,*}(\mathbb{R}) \}$, where R is an arbitrary positive constant and $\mathrm{Sym}^+_{2k_g,*}(\mathbb{R})$ is the set of strictly positive symmetric matrices. Concerning the weights $(\rho_t)_{1\leq t\leq \tau_m}$, we assume that they belong to the set $\varrho = \left\{ (\rho_t)_{1\leq t\leq \tau_m} \in]0,1[^{\tau_m} \mid \sum_{t=1}^{\tau_m} \rho_t = 1 \right\}$.

Remark 2.1. This yields a generative model: given the parameters of the model, to get a realisation of an image, we first draw a label τ with respect to the probability law $\sum_{t=1}^{\tau_m} \rho_t \delta_t$. Then, we simulate a deformation field β using the covariance matrix corresponding to component τ according to $\mathcal{N}(0, \Gamma_{g,\tau})$. We apply it to the template of the τth component. Last, we add an independent Gaussian noise of variance σ_τ^2.

We choose a normal distribution for the unobserved deformation variable because of the background we have in image analysis. Indeed, the registration problem is an issue that has been studied deeply for the past two decades. The goal is, given two images, to find the best deformation that will match one image close to the other. Such methods require to choose the kind of deformations that are allowed (smooth, diffeomorphic, etc.). These conditions are equivalent, for some of these methods, to choose a covariance matrix that enables to define an inner product between two deformations coded by a vector β (*cf.* [4,18]). The regularisation term of the matching energy in the small deformation framework treated in this paper can be written as: $\beta^t \Gamma_g^{-1} \beta$. This looks like the logarithm of the density of a Gaussian distribution on β with 0 mean and a covariance matrix Γ_g. The link between these two points of view has been given in [1]; the mode of the posterior distribution equals the solution of a general matching problem. This is why we therefore set on the deformation vector β such a distribution. Moreover, many experiments have been run using a large variety of such a matrix which gives us now a good initial guess for our parameter. This leads us to consider a Bayesian approach with a weakly informative prior.

2.3. **The Bayesian approach**

The information given by the image analysis background is here introduced mathematically in terms of prior laws on the parameters of model (2.1). As already mentioned in the previous paragraph, this background knowledge enables to determine a good initial guess for the laws and the values of the hyper-parameters. As well as for the covariance matrix Γ_g, the same arguments are true for the noise variance σ^2. In the registration viewpoint, this variance is the tradeoff between the deformation cost and the data attachment term that compose the energy to minimise. An empirical good initial guess is therefore known as well.

On another hand, the high dimensionality of the parameters can lead to degenerated maximum likelihood estimator when the training sample is small. While introducing prior distributions, the estimation with small samples is still possible. The importance of these prior distributions in the estimation problem has been shown in [1]. The solution of the estimation equation can be interpreted as barycenters between the hyper-parameters of the priors and the empirical values. This ensures easy computations and other theoretical properties as for example, the invertibility of the covariance matrix Γ_g. The role of the other hyper-parameters are discussed in the experiments.

We use a generative model which includes natural standard conjugate prior distributions with *fixed* hyper-parameters. These distributions are an inverse-Wishart priors on each $\Gamma_{g,t}$ and σ_t^2 and a normal prior on each α_t, for all $1 \leq t \leq \tau_m$. All priors are assumed independent. Then,

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

where $\langle A, B\rangle_F \triangleq \mathrm{tr}(A^t B)$ is the scalar product of two matrices A and B and tr stands for the trace.

For the prior law ν_ρ, we choose the Dirichlet distribution, $\mathcal{D}(a_\rho)$, with density

$$\nu_\rho(\rho) \propto \left(\prod_{t=1}^{\tau_m} \rho_t\right)^{a_\rho}, \text{with fixed parameter } a_\rho.$$

The choice of the hyper-parameters in practice will be discussed in Section 5.1.1.

3. Parameter estimation using a stochastic version of the EM algorithm

For the sake of simplicity, let us denote by $N \triangleq 2nk_g$ and by $\mathcal{T} \triangleq \{1, \ldots, \tau_m\}^n$ so that the missing deformation variables take their values in $\mathbb{R}^N$ and the missing labels in $\mathcal{T}$. We also introduce the following notations: $\eta = (\theta, \rho)$ with $\theta = (\theta_t)_{1 \leq t \leq \tau_m}$ and $\rho = (\rho_t)_{1 \leq t \leq \tau_m}$.

In our Bayesian framework, we choose the MAP estimator to estimate the parameters:

$$\tilde{\eta}_n = \operatorname*{argmax}_\eta q_B(\eta|\boldsymbol{y}), \tag{3.1}$$

where $q_B(\eta|\boldsymbol{y})$ denotes the distribution of η conditionally to $\boldsymbol{y}$.

Remark 3.1. Even if we are working in a Bayesian framework, we do not want to estimate the distributions of our parameters. Knowing the distribution of the template image and its possible deformations is not of great interest from an image analysis point of view. Indeed, people are more interested, in particular in the medical imaging community, in an atlas which characterises the populations of shapes that they consider rather than its distribution. Moreover, the distribution of the deformation law makes even less sense. This is the reason why we focus on the MAP.

In practice, to reach this estimator, we maximise this posterior distribution using a Stochastic Approximation EM (SAEM) algorithm coupled with a Monte Carlo Markov Chain (MCMC) method. Indeed, due to the intractable computation of the E step of the EM algorithm introduced by [8] encountered in this complex non linear setting, we follow a stochastic variation called SAEM proposed in [7]. However, again due to the expression of our model, the simulation required in this algorithm cannot be performed directly. Therefore, we propose to use some MCMC methods to reach this simulation as proposed by Kuhn and Lavielle in [14] and done for the one component model in [3]. Unfortunately, the direct generalisation of the algorithm presented in [3] paper turns out to be of no use in practice because of some trapping state problems (*cf.* Sect. 3.2.). This suggests to go back to some other extension of the SAEM procedure.

3.1. The SAEM algorithm using MCMC methods

Let us first recall the SAEM algorithm. It generates a sequence of estimated parameters $(\eta_k)_k$ which converges towards a critical point of $\eta \mapsto \log q(\boldsymbol{y}, \eta)$ under some mild assumptions (*cf.* [7]). These critical points coincide with the critical points of $\eta \mapsto \log q_B(\eta|\boldsymbol{y})$. The kth iteration consists in three steps:

Simulation step: the missing data, here the deformation parameters and the labels, $(\boldsymbol{\beta}, \boldsymbol{\tau})$, are drawn with respect to the distribution of $(\boldsymbol{\beta}, \boldsymbol{\tau})$ conditionally to $\boldsymbol{y}$ denoted by π_η, using the current parameter η_{k-1}

$$(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) \sim \pi_{\eta_{k-1}}, \tag{3.2}$$

Stochastic approximation step: given $(\Delta_k)_k$ a decreasing sequence of positive step-sizes, a stochastic approximation is done on the quantity $\log q(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \eta)$, using the simulated value of the missing data:

$$Q_k(\eta) = Q_{k-1}(\eta) + \Delta_k[\log q(\boldsymbol{y}, \boldsymbol{\beta}_k, \boldsymbol{\tau}_k, \eta) - Q_{k-1}(\eta)], \tag{3.3}$$

Maximisation step: the parameters are updated in the M-step,

$$\eta_k = \underset{\eta}{\operatorname{argmax}}\, Q_k(\eta). \tag{3.4}$$

Initial values Q_0 and η_0 are arbitrarily chosen.

We notice that the density function of the model proposed in paragraphs 2.2 and 2.3 belongs to the curved exponential family. That is to say that the complete likelihood can be written as: $q(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \eta) = \exp\left[-\psi(\eta) + \langle S(\boldsymbol{\beta}, \boldsymbol{\tau}), \phi(\eta)\rangle\right]$, where the sufficient statistic S is a Borel function on $\mathbb{R}^N \times \mathcal{T}$ taking its values in an open subset $\mathcal{S}$ of $\mathbb{R}^m$ and ψ, ϕ two Borel functions on $\Theta \times \varrho$. (Note that S, ϕ and ψ may depend also on $\boldsymbol{y}$, but since $\boldsymbol{y}$ will stay fixed in the sequel, we omit this dependency.) Thanks to this property of our model, it is equivalent to do the stochastic approximation on the complete log-likelihood as well as on the sufficient statistics. This yields equation (3.3) to be replaced by the following stochastic approximation s of the sufficient statistics S:

$$s_k = s_{k-1} + \Delta_k(S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - s_{k-1}). \tag{3.5}$$

We now introduce the following function: $L : \mathcal{S} \times \Theta \times \varrho \to \mathbb{R}$ as $L(s; \eta) = -\psi(\eta) + \langle s, \phi(\eta)\rangle$. It has been proved in [1] that there exists a critical function $\hat{\eta} : \mathcal{S} \to \Theta \times \varrho$ which is a zero of ∇L. It is straightforward to prove that this function satisfies: $\forall \eta \in \Theta \times \varrho, \forall s \in \mathcal{S}, L(s; \hat{\eta}(s)) \geq L(s; \eta)$ so that the maximisation step (3.4) becomes:

$$\eta_k = \hat{\eta}(s_k).$$

Concerning the simulation step, in our model, the simulation of the missing variables with respect to the conditional distribution π_η cannot be carried out. Indeed, its probability density function (pdf) has a close form but rather complicated; it does not correspond to some usual pdf. One solution proposed in [14] for such cases is to couple the SAEM algorithm with Monte Carlo Markov Chain (MCMC) method. However, we do not fit exactly into their requirements since the missing variable $\boldsymbol{\beta}$ does not have a compact support. We introduce an ergodic Markov chain whose stationary distribution is the conditional distribution π_η. We denote its transition kernel by Π_η. The simulation step (3.2) is thus replaced by the following step:

$$(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) \sim \Pi_{\eta_{k-1}}((\boldsymbol{\beta}_{k-1}, \boldsymbol{\tau}_{k-1}), .). \tag{3.6}$$

The most common choice of kernel is an accept-reject step which is carried out through a Metropolis-Hastings algorithm. Unfortunately, in our particular setting, we deal with large dimensions for the missing variables. This made us move to some other kind of MCMC methods, like a Gibbs sampler, to simulate our missing variables.

3.2. The transition of the MCMC method: a hybrid Gibbs sampler

If we consider the full vector $(\boldsymbol{\beta}, \boldsymbol{\tau})$ as a single vector of missing data, we can use the hybrid Gibbs sampler on $\mathbb{R}^N \times \mathcal{T}$ as follows. For any $b \in \mathbb{R}$ and $1 \leq j \leq N$, let us denote by $\boldsymbol{\beta}_{b \to j}$ the unique configuration which is equal to $\boldsymbol{\beta}$ everywhere except the coordinate j where $\boldsymbol{\beta}^j_{b \to j} = b$ and by $\boldsymbol{\beta}^{-j}$ the vector $\boldsymbol{\beta}$ without the coordinate j. Each coordinate of the deformation field $\boldsymbol{\beta}^j$ is updated using a Metropolis-Hastings step where the proposal is given by the conditional distribution of $\boldsymbol{\beta}^j | \boldsymbol{\beta}^{-j}, \boldsymbol{\tau}$ coming from the current Gaussian distribution with the corresponding parameters (pointed by $\boldsymbol{\tau}$). Then, the last coordinates corresponding to the missing variable $\boldsymbol{\tau}$ are drawn with respect to $q(\boldsymbol{\tau} | \boldsymbol{\beta}, \boldsymbol{y}, \eta)$.

Even if this procedure provides an estimated parameter sequence which would theoretically converge toward the MAP estimator, in practice, as mentioned in [19], it would take a quite long time to reach its limit because of the trapping state problem: when a small number of observations are assigned to a component, the estimation of the component parameters is hardly concentrated and the probability of changing the label of an image to this component or from this component to another is really small (most of the time under the computer precision).

We can interpret this from an image analysis viewpoint: the first iteration of the algorithm gives a random label to the training set and computes the corresponding maximiser $\eta = (\theta, \rho)$. Then, for each image, according

to its current label, it simulates a deformation field which only takes into account the parameters of this given component. Indeed, the simulation of $\boldsymbol{\beta}$ through the Gibbs sampler involves a proposal whose corresponding Markov chain has $q(\boldsymbol{\beta}|\boldsymbol{\tau}, \boldsymbol{y}, \eta)$ as stationary distribution. Therefore, the deformation tries to match $\boldsymbol{y}$ to the deformed template of the given component $\boldsymbol{\tau}$. The deformation field tries to get a better connection between the component parameters and the observation, and there is only small probability that the observation given *this* deformation field will be closer to another component. The update of the label $\boldsymbol{\tau}$ is therefore conditional to this deformation which would not leave much chance to switch component.

To overcome the trapping state problem, we will simulate the optimal label, using as many Markov chains in $\boldsymbol{\beta}$ as the number of components so that each component has a corresponding deformation which "computes" its distance to the observation. Then we can simulate the optimal deformation corresponding to that optimal label.

Since we aim to simulate $(\boldsymbol{\beta}, \boldsymbol{\tau})$ through a transition kernel that has $q(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{y}, \eta)$ as stationary distribution, we simulate $\boldsymbol{\tau}$ with a kernel whose stationary distribution is $q(\boldsymbol{\tau}|\boldsymbol{y}, \eta)$ and then $\boldsymbol{\beta}$ through a transition kernel that has $q(\boldsymbol{\beta}|\boldsymbol{\tau}, \boldsymbol{y}, \eta)$ as stationary distribution.

For the first step, we need to compute the weights $q(t|y_i, \eta) \propto q(t, y_i|\eta)$ for all $1 \leq t \leq \tau_m$ and all $1 \leq i \leq n$ which cannot be easily reached. However, for any density function f, for any image y_i and for any $1 \leq t \leq \tau_m$, we have

$$q(t, y_i|\eta) = \left(\mathbb{E}_{q(\beta|y_i,t,\eta)} \left[\frac{f(\beta)}{q(y_i, \beta, t|\eta)} \right] \right)^{-1}. \tag{3.7}$$

Obviously the computation of this expectation w.r.t. the posterior distribution is not tractable either but we can approximate it by a Monte Carlo sum. However, we cannot easily simulate variables through the posterior distribution $q(\cdot|y_i, t, \eta)$ as well, so we use some realisations of an ergodic Markov chain having $q(\cdot|y_i, t, \eta)$ as stationary distribution instead of some independent realisations of this distribution.

The solution we propose is the following: suppose we are at the kth iteration of the algorithm and let η be the current parameters. Given any initial deformation field $\xi_0 \in \mathbb{R}^{2k_g}$, we run, for each component t, the hybrid Gibbs sampler $\Pi_{\eta,t}$ on $\mathbb{R}^{2k_g}$ J times so that we get J elements $\xi_{t,i} = (\xi_{t,i}^{(l)})_{1 \leq l \leq J}$ of an ergodic homogeneous Markov chain whose stationary distribution is $q(\cdot|y_i, t, \eta)$. Let us denote by $\xi_i = (\xi_{t,i})_{1 \leq t \leq \tau_m}$ the matrix of all the auxiliary variables. We then use these elements for the computation of the weights $p_J(t|\xi_i, y_i, \eta)$ through a Monte Carlo sum:

$$p_J(t|\xi_i, y_i, \eta) \propto \left(\frac{1}{J} \sum_{l=1}^{J} \left[\frac{f(\xi_{t,i}^{(l)})}{q(y_i, \xi_{t,i}^{(l)}, t|\eta)} \right] \right)^{-1}, \tag{3.8}$$

where the normalisation is done such that their sum over t equals one, involving the dependence on all the auxiliary variables ξ_i. The ergodic theorem ensures the convergence of our approximation toward the expected value. We then simulate $\boldsymbol{\tau}$ through $\otimes_{i=1}^{n} \sum_{t=1}^{\tau_m} p_J(t|\xi_i, y_i, \eta)\delta_t$.

Concerning the second step, we update $\boldsymbol{\beta}$ by re-running J times the hybrid Gibbs sampler $\Pi_{\eta,\boldsymbol{\tau}}$ on $\mathbb{R}^N$ starting from a random initial point $\boldsymbol{\beta}_0$ in a compact subset of $\mathbb{R}^N$. The size of J will depend on the iteration k of the SAEM algorithm in a sense that will be precised later, thus we now index it by k.

The density function f involved in the Monte Carlo sum above needs to be specified to get the convergence result proved in the last section of this paper. We show that using the prior on the deformation field enables to get the sufficient conditions for convergence. This density is the Gaussian density function and depends on the component we are working with:

$$f_t(\xi) = \frac{1}{\sqrt{2\pi}^{2k_g} \sqrt{|\Gamma_{g,t}|}} \exp\left(-\frac{1}{2} \xi^t \Gamma_{g,t}^{-1} \xi \right). \tag{3.9}$$

Algorithm 2 shows the detailed iteration.

Remark 3.2. The use of one simulation of $\boldsymbol{\beta}$ per component is a point that was already used in [1] while computing the best matching, β^*, for all components by minimising the corresponding energies. This gives as many β^* as components for each image. Then, according to these best matchings, it computed the best component which therefore pointed the matching to consider.

3.3. **Truncation on random boundaries**

Since our missing data have a non-compact support, some of the convergence assumptions of such algorithms [14] are not satisfied. This leads to consider a truncation algorithm as suggested in [7] and extended in [3].

Let $(\mathcal{K}_q)_{q\geq 0}$ be an increasing sequence of compact subsets of $\mathcal{S}$ such as $\cup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}), \forall q \geq 0$. Let K be a compact subset of $\mathbb{R}^N$. Let Π_η be a transition kernel of an ergodic Markov chain on $\mathbb{R}^N$ having π_η as stationary distribution. We construct the homogeneous Markov chain $((\boldsymbol{\beta}_k, \boldsymbol{\tau}_k, s_k, \kappa_k))_{k\geq 0}$ as explained in Algorithm 1. As long as the stochastic approximation does not wander out the current compact set, we run our "SAEM-MCMC" algorithm. As soon as this previous condition is not satisfied, we reinitialise the sequences of s and $(\boldsymbol{\beta}, \boldsymbol{\tau})$ using a projection (for more details see [7]). The current compact is then enlarge. To point toward the current compact, we use a counter sequence $(\kappa_k)_k$ which remains unchanged when the previous condition is satisfied and increases to point toward a bigger compact when re-projecting.

Algorithm 1 Stochastic approximation with truncation on random boundaries

Set $\boldsymbol{\beta}_0 \in \text{K}$, $\boldsymbol{\tau}_0 \in \mathcal{T}$, $s_0 \in \mathcal{K}_0$ and $\kappa_0 = 0$.
for all $k \geq 1$ **do**
  compute $\bar{s} = s_{k-1} + \Delta_k(S(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\tau}}) - s_{k-1})$
  where $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\tau}})$ are sampled from a transition kernel $\Pi_{\eta_{k-1}}$ (see Algorithm 2).
  if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ **then**
    set $(s_k, \boldsymbol{\beta}_k, \boldsymbol{\tau}_k) = (\bar{s}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\tau}})$ and $\kappa_k = \kappa_{k-1}$
  else
    set $(s_k, \boldsymbol{\beta}_k, \boldsymbol{\tau}_k) = (\tilde{s}, \tilde{\boldsymbol{\beta}}, \bar{\boldsymbol{\tau}}) \in \mathcal{K}_0 \times \text{K} \times \mathcal{T}$ and $\kappa_k = \kappa_{k-1} + 1$
    and $(\tilde{s}, \tilde{\boldsymbol{\beta}})$ can be chosen through different ways (*cf.* [7]).
  end if
  $\eta_k = \underset{\eta}{\text{argmax}}\, \hat{\eta}(s_k)$.
end for

4. Convergence theorem of the multicomponent procedure

In this particular section the variances of the components $(\sigma_t^2)_{1\leq t\leq \tau_m}$ are fixed. Alleviating this condition is not straightforward and is an issue of our future work.

To prove the convergence of our parameter estimate toward the MAP, we have to go back to a convergence theorem which deals with general stochastic approximations. Indeed, the SAEM-MCMC algorithm introduced and detailed above is a Robbins-Monro type stochastic approximation procedure. One common tool to prove the w.p.1 convergence of such a stochastic approximation has been introduced by Kushner and Clark in [15]. However, some of the assumptions they require are intractable with our procedure (in particular concerning the mean field defined below). This leads us to slightly adapt the convergence theorem for stochastic approximations given in [7].

We consider the following Robbins-Monro stochastic approximation procedure:

$$s_k = s_{k-1} + \Delta_k(h(s_{k-1}) + e_k + r_k), \tag{4.1}$$

Algorithm 2 Transition step $k \to k+1$ using a hybrid Gibbs sampler on $(\boldsymbol{\beta}, \boldsymbol{\tau})$

Require: $\eta = \eta_k$, $J = J_k$
  for all $i = 1 : n$ **do**
    for all $t = 1 : \tau_m$ **do**
      $\xi_{t,i}^{(0)} = \xi_0$
      for all $l = 1 : J$ **do**
        $\xi = \xi_{t,i}^{(l-1)}$
        `Gibbs sampler` $\Pi_{\eta,t}$`:`
        for all $j = 1 : 2k_g$ **do**
          `Metropolis-Hastings procedure:`
          $b \sim q(b|\xi^{-j}, t, \eta)$
          Compute $r_j(\xi^j, b; \xi^{-j}, \eta, t) = \left[\frac{q(y_i|\xi_{b\to j}, t, \eta)}{q(y_i|\xi, t, \eta)} \wedge 1\right]$
          $u \sim \mathcal{U}[0,1]$
          if $u < r_j(\xi^j, b; \xi^{-j}, \eta, t)$ **then**
            $\xi^j = b$
          end if
        end for
        $\xi_{t,i}^{(l)} = \xi$
      end for

$$p_J(t|\xi_i, y_i, \eta) \propto \left(\frac{1}{J}\sum_{l=1}^{J}\left[\frac{f_t(\xi_{t,i}^{(l)})}{q(y_i, \xi_{t,i}^{(l)}, t|\eta)}\right]\right)^{-1}$$

    end for
  end for

$$\boldsymbol{\tau}_{k+1} \sim \otimes_{i=1}^{n} \sum_{t=1}^{\tau_m} p_J(t|\xi_i, y_i, \eta)\delta_t \quad \text{and} \quad \boldsymbol{\beta}_{k+1} \sim \Pi^J_{\eta, \boldsymbol{\tau}_{k+1}}(\boldsymbol{\beta}_0).$$

where $(e_k)_{k\geq 1}$ and $(r_k)_{k\geq 1}$ are random processes defined on the same probability space taking their values in an open subset $\mathcal{S}$ of $\mathbb{R}^{n_s}$; h is referred to as the mean field of the algorithm; $(r_k)_{k\geq 1}$ is a remainder term and $(e_k)_{k\geq 1}$ is a stochastic excitation.

To be able to get a convergence result, we consider the truncated sequence $(s_k)_k$ defined as follow: let $\mathcal{S}_a \subset \mathcal{S}$ and $\bar{s}_k = s_{k-1} + \Delta_k h(s_{k-1}) + \Delta_k e_k + \Delta_k r_k$, where

$$\begin{cases} \text{if } \bar{s}_k \in \mathcal{K}_{\kappa_{k-1}} \begin{cases} s_k = \bar{s}_k\,, \\ \kappa_k = \kappa_{k-1}\,, \end{cases} \\ \text{if } \bar{s}_k \notin \mathcal{K}_{\kappa_{k-1}} \begin{cases} s_k = \tilde{s}_k \in \mathcal{S}_a\,, \\ \kappa_k = \kappa_{k-1} + 1\,. \end{cases} \end{cases} \tag{4.2}$$

The projection $\tilde{s}_k$ can be made through different ways (*cf.* [7]).

We will use Delyon's theorem which gives sufficient conditions for the sequence $(s_k)_{k\geq 0}$ truncated on random boundaries to converge with probability one. The theorem we state here is a generalisation of the theorem presented in [7]. Indeed, we have add the existence of an absorbing set for the stochastic approximation sequence. The proof of this theorem can be carried out the same way Delyon et al. do theirs adding the absorbing set. This is why it is not detailed here.

Theorem 4.1. *We consider the sequence $(s_k)_{k\geq 0}$ given by the truncated procedure (4.2). Assume that:*

(SA0') *There exists a closed convex set $\mathcal{S}_a \subset \mathcal{S}$ such that for all $k \geq 0$, $s_k \in \mathcal{S}_a$ w.p.1.*

(SA1) $(\Delta_k)_{k\geq 1}$ *is a decreasing sequence of positive numbers such that* $\sum_{k=1}^{\infty} \Delta_k = \infty$.

(SA2) *The vector field* h *is continuous on* $\mathcal{S}$ *ou* $\mathcal{S}_a$ *and there exists a continuously differentiable function* $w : \mathcal{S} \to \mathbb{R}$ *such that*

(i) *for all* $s \in \mathcal{S}, F(s) = \langle \partial_s w(s), h(s) \rangle \leq 0$.

(ii) $int(w(\mathcal{L}')) = \emptyset$, *where* $\mathcal{L}' \triangleq \{s \in \mathcal{S} : F(s) = 0\}$.

(STAB1') *There exist a continuous differentiable function* $W : \mathbb{R}^N \to \mathbb{R}$ *and a compact set* $\mathcal{K}$ *such that*

(i) *For all* $c \geq 0$, *we have* $\mathcal{W}_c \cap \mathcal{S}_a$ *is a compact subset of* $\mathcal{S}$ *where* $\mathcal{W}_c = \{s \in \mathcal{S} : \ W(s) \leq c\}$ *is a level set.*

(ii) $\langle \partial_s W(s), h(s) \rangle < 0$, *for all* $s \in \mathcal{S} \setminus \mathcal{K}$.

(STAB2) *For any positive integer* M, *w.p.1* $\lim_{p\to\infty} \sum_{k=1}^{p} \Delta_k e_k \mathbb{1}_{W(s_{k-1})\leq M}$ *exists and is finite and w.p.1* $\limsup_{k\to\infty} |r_k| \mathbb{1}_{W(s_{k-1})\leq M} = 0$.

Then, w.p.1, $\limsup_{k\to\infty} d(s_k, \mathcal{L}') = 0$.

Assumption (SA2), which involves a Lyapunov function w, replaces the usual condition that, w.p.1, the sequence comes back infinitely often in a compact set which is not easy to check in practice. In addition, assumptions (STAB1') and (STAB2) give a recurrent condition introducing a Lyapunov function W which controls the excursions outside the compact sets. The two Lyapunov functions w and W do not need to be the same. Another interesting point is that the truncation does not change the mean field and therefore the stationary points of the sequence.

This theorem does not ensure the convergence of the sequence to a maximum of the likelihood but to one of its critical points. To ensure that the critical point reached is a maximum, we would have to satisfy two other conditions (called (LOC1-2) in [7]) which are typical conditions. That is to say, it requires that the critical points are isolated and for every stationary points $s^* \in \mathcal{L}'$ the Hessian matrix of the observed log-likelihood is negative definite.

We now want to apply this theorem to prove the convergence of our "SAEM like" procedure where the missing variables are not simulated through the posterior density function but by a kernel which can be as close as wanted -increasing J_k- to this posterior law (generalising Thm. 3 in [7]).

Let us consider the following stochastic approximation: $(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k)$ are simulated by the transition kernel described in the previous section and

$$s_k = s_{k-1} + \Delta_k (S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - s_{k-1}) \,,$$

which can be connected to the Robbins-Monro procedure using the notations introduced in [7]: let $\mathcal{F} = (\mathcal{F}_k)_{k\geq 1}$ be the filtration where $\mathcal{F}_k$ is the σ-algebra generated by the random variables $(S_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_k)$, $\mathbb{E}_{\pi_\eta}$ is the expectation with respect to the posterior distribution π_η and

$$\begin{aligned} h(s_{k-1}) &= \mathbb{E}_{\pi_{\hat{\eta}(s_{k-1})}} [S(\boldsymbol{\beta}, \boldsymbol{\tau})] - s_{k-1}, \\ e_k &= S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - \mathbb{E} [S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) | \mathcal{F}_{k-1}] \,, \\ r_k &= \mathbb{E} [S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) | \mathcal{F}_{k-1}] - \mathbb{E}_{\pi_{\hat{\eta}(s_{k-1})}} [S(\boldsymbol{\beta}, \boldsymbol{\tau})] \,. \end{aligned}$$

Theorem 4.2. *Let* $w(s) = -l(\hat{\eta}(s))$ *where* $l(\eta) = \log \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} q(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \eta) \mathrm{d}\boldsymbol{\beta}$ *and* $h(s) = \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} (S(\boldsymbol{\beta}, \boldsymbol{\tau}) - s) \pi_{\hat{\eta}(s)}(\boldsymbol{\beta}, \boldsymbol{\tau}) \mathrm{d}\boldsymbol{\beta}$ *for* $s \in \mathcal{S}$. *Assume that:*

(A1) *the sequences* $(\Delta_k)_{k\geq 1}$ *and* $(J_k)_{k\geq 1}$ *satisfy:*

(i) $(\Delta_k)_{k\geq 1}$ *is non-increasing, positive,* $\sum_{k=1}^{\infty} \Delta_k = \infty$ *and* $\sum_{k=1}^{\infty} \Delta_k^2 < \infty$.

(ii) $(J_k)_{k\geq 1}$ *takes its values in the set of positive integers and* $\lim_{k\to\infty} J_k = \infty$.

(A2) $\mathcal{L}' \triangleq \{s \in \mathcal{S}, \langle \partial_s w(s), h(s) \rangle = 0\}$ *is included in a level set of* w.

Let $(s_k)_{k\geq 0}$ *be the truncated sequence defined in equation (4.2),* K *a compact set of* $\mathbb{R}^N$ *and* $\mathcal{K}_0 \subset S(\mathbb{R}^N)$ *a compact subset of* $\mathcal{S}$. *Then, for all* $\boldsymbol{\beta}_0 \in \mathrm{K}$, $\boldsymbol{\tau}_0 \in \mathcal{T}$ *and* $s_0 \in \mathcal{K}_0$, *we have*

$$\lim_{k\to\infty} \mathrm{d}(s_k, \mathcal{L}') = 0 \ \bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0} \text{ -a.s.},$$

where $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0}$ *is the probability measure associated with the chain* $(Z_k = (\boldsymbol{\beta}_k, \boldsymbol{\tau}_k, s_k, \kappa_k))_{k\geq 0}$ *starting at* $(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0, s_0, 0)$.

The first assumption which concerns the two sequences involved in the algorithm, is not restrictive at all since these sequences can be chosen arbitrarily.

The second assumption, however, is more complex. This is required to satisfy the assumptions of Theorem 4.1. This is a condition we have not proved yet and is part of our future work.

Proof. The proof of this theorem is given in Section 6. We give here a quick sketch to emphasise the main difficulties and differences between our proof and the convergence proof of the SAEM algorithm given in [7].

Even if the only algorithmic difference between our algorithm and the SAEM algorithm is the simulation of the missing data which is not done with respect to the posterior law $q(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{y}, \eta)$ but through an approximation which can be arbitrarily close, this yields a very different proof. Indeed, whereas for the SAEM algorithm, the stochastic approximation leads to a Robbins-Monro type equation (4.1) with no residual term r_k, our method induces one. The first difficulty is therefore to prove that this residual term tends to 0 while the number of iterations k tends to infinity. Our proof is decomposed into two part, the first one concerning the deformation variable $\boldsymbol{\beta}$ and the second one the label $\boldsymbol{\tau}$. The first term requires to prove the geometric ergodicity of the Markov chain in $\boldsymbol{\beta}$ generated through our kernel. For this purpose, we prove some typical sufficient conditions which include the existence of a small set for the transition kernel and a drift condition. Then, we use for the second term some concentration inequalities for non stationary Markov chains to prove that the kernel associated with the label distribution converges toward the posterior distribution $q(\boldsymbol{\tau}|\boldsymbol{y}, \eta)$.

The second difficulty is to prove the convergence of the excitation term e_k. This can be carried out as in [7] using the properties of our Markov chain and some martingale limits properties. □

Corollary 4.1. *Under the assumptions of Theorem 4.2 we have for all* $\boldsymbol{\beta}_0 \in \mathrm{K}$, $\boldsymbol{\tau}_0 \in \mathcal{T}$, $s_0 \in \mathcal{S}_a$ *and* $\eta_0 \in \Theta \times \varrho$,

$$\lim_{k\to\infty} \mathrm{d}(\eta_k, \mathcal{L}) = 0 \quad \bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0}\text{-a.s},$$

where $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0,}$ *is the probability measure associated with the chain* $(Z_k = (\boldsymbol{\beta}_k, \boldsymbol{\tau}_k, s_k, \kappa_k))_{k\geq 0}$ *starting at* $(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0, s_0, 0)$ *and* $\mathcal{L} \triangleq \{\eta \in \hat{\eta}(\mathcal{S}), \frac{\partial l}{\partial \eta}(\eta) = 0\}$.

Proof. This is a direct consequence of the smoothness of the function $s \mapsto \hat{\eta}(s)$ on $\mathcal{S}$ and of Lemma 2 of [7] which links the sets $\mathcal{L}$ and $\mathcal{L}'$. □

5. Illustrative experiments

5.1. **USPS database**

To illustrate the previous algorithm for the deformable template model, we are considering handwritten digit images. For each digit, referred as class later, we learn two templates, the corresponding noise variances and the geometric covariance matrices. We use the USPS database which contains a training set of around 7000 images. Each picture is a (16×16) gray level image with intensity in $[0, 2]$ where 0 corresponds to the black background. In Figure 1 we show some of the training images used for the statistical estimation.

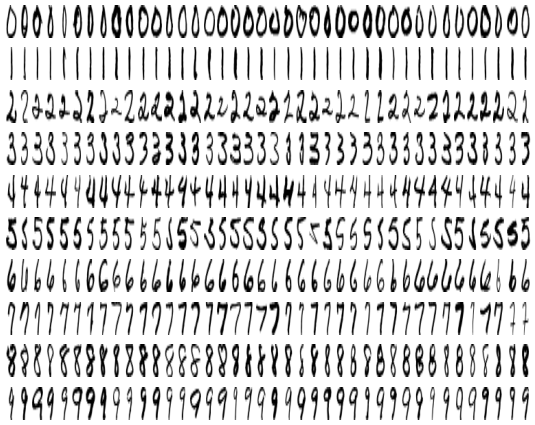


FIGURE 1. Some examples of the training set: 40 images per class (inverse video).

5.1.1. *General setting of the hierarchical model*

A natural choice for the prior laws on α and Γ_g is to set 0 for the mean on α and to induce the two covariance matrices by the metric of the spaces V_p and V_g involving the correlation between the landmarks through the kernels: define the square matrices $M_p(k,k') = K_p(v_{p,j}, v_{p,j'})\ \forall 1 \leq k,k' \leq k_p$, and $M_g(k,k') = K_g(v_{g,j}, v_{g,j'})\ \forall 1 \leq k,k' \leq k_g$. Then $\Sigma_p = M_p^{-1}$ and $\Sigma_g = M_g^{-1}$. In our experiments, we have chosen Gaussian kernels for both K_p and K_g. Their respective standard deviations are fixed: $\sigma_p = 0.2$ and $\sigma_g = 0.12$ for an estimation on $[-1.5, 1.5]^2$ and $[-1, 1]$ respectively. Indeed, too large values of these standard deviations lead to smooth images and deformations. The registration becomes less accurate and yields blurry estimated templates and non optimal estimated covariance matrices. On the other hand, too small values of σ_p and σ_g concentrate the information in a small neighbourhoods around the landmarks; for example the templates would be composed of small balls centred on the landmarks.

Another issue is the calibration of the parameter a_g. Indeed, to satisfy the theorical requirement a_g is supposed to be larger than $4k_g$ which equals 144 in our case. However, looking at the geometrical update formula (6.2), this hyper-parameter corresponds to the weight of the prior matrix in the barycenter expression between the prior and the empirical terms. Since we only have few images, this prior would predominate over the empirical term. This would prevent the geometrical covariance from moving far away from the prior. So we choose $a_g = 0.5$ which leads to a well defined posterior distribution even if it does not satisfy the theoretical assumption.

The influence of the other hyper-parameters a_p and a_ρ has been tested on simulation and is negligeable. We choose $a_p = 200$ and $a_\rho = 2$.

5.1.2. *General setting of the algorithm*

For the stochastic approximation step-size, we allow a heating period k_h which corresponds to the absence of memory for the first iterations. This allows the Markov chain to reach an area of interest in the posterior probability density function $q(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{y}, \eta)$ before exploring this particular region. In the experiments presented, the heating time k_h lasts up to 150 iterations and the whole algorithm is stopped at, at most, 200 iterations depending on the data set (noisy or not). This number of iterations corresponds to a point when the convergence is reached. We choose, as suggested in [7] the step-size sequence as $\Delta_k = 1$ for all $1 \leq k \leq k_h$ and $\Delta_k = (k - k_h)^{-0.6}$ otherwise.

The multicomponent case has to face the problem of its computational time. Indeed, as we have to approximate the posterior density by running J_k elements of τ_m independent Markov chains, the computation time increases linearly with J_k. In our experiments, we have chosen a fixed J for every EM iteration, $J = 50$. This is enough to get a good approximation thanks to the coupling between the iterations of the SAEM algorithm and

FIGURE 2. Estimated prototypes of the two components model for each digit (40 images per class; 100 iterations; two components per class).

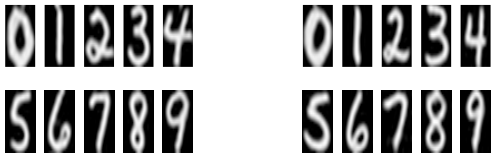


FIGURE 3. Estimated prototypes of the two components model for each digit (40 images per class, second random sample).

the iterations of the Markov chains. Indeed, even if the parameter η is modified along the SAEM iterations, its successive jumps are small enough to ensure the convergence of the MCMC method. Intuitively speaking, it is equivalent to consider not only 50 iterations of the MCMC method but 50 times the number of SAEM iterations.

5.1.3. *The estimated templates*

We are showing here the results of the statistical learning algorithm for our generative model. To avoid the problems shown in [3], we choose the same initialisation of the template parameter α as they did, that is to say, we set the initial value of α such that the corresponding I_α is the mean of the gray-level training images.

In Figure 2, we show the two estimated templates obtained by the multicomponent procedure with 40 training examples per class. It appears that, as for the mode approximation algorithm which results are presented on this database in [1], the two components reached are meaningful, such as the 2 with and without loop or American and European 7. They even look alike.

In Figure 3, we show a second run of the algorithm with a different database, the training images are randomly selected in the whole USPS training set. We can see that there are some variability, in particular for digit 7 where there were no European 7 in this training set. This generates two different clusters still relevant for this digit. The other digits are quite stable, in particular the strongly constrained ones (like 3, 5, 8 or 9).

5.1.4. *The photometric noise variance*

Even if we prove the convergence result for fixed component noise variances, we still try to learn them in the experiments. The same behaviour for our stochastic EM as for the mode approximation EM algorithm done in [1] is observed for the noise variances. Indeed, allowing the decomposition of the class into components enables the model to better fit the data yielding a lower residual noise. In addition, the stochastic algorithm enables to look around the whole posterior distribution and not only focusing on its mode which increases the accuracy of the geometric covariance and the template estimation. This yields lower noise required to explain the gap between the model and the truth. The evolution of the estimated variances for the two components of each digits are presented in Figure 4.

The convergence of this variance for some very constrained digits like digit 1 is faster. This is due to the well defined templates and geometric variability in the class which can be easily captured. Therefore, a very low level of noise is required very quickly. On the other hand, some very variable digits like digit 2 are slower

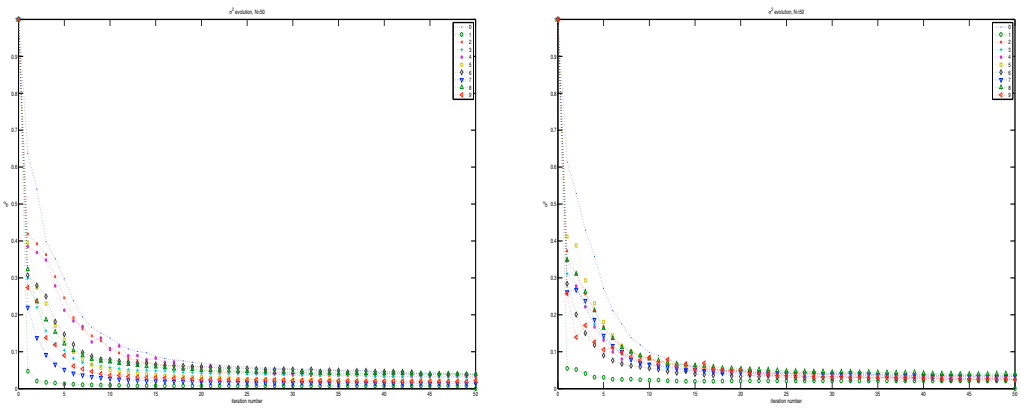


FIGURE 4. Evolution of the two cluster variances along the iterations.

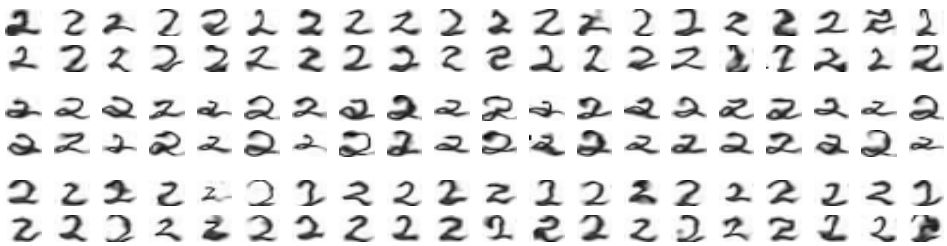


FIGURE 5. Some synthetic examples of the components of digit 2: first four rows: templates of the two components deformed through some deformation field β and $-\beta$ drawn from their respective geometric covariance. Two last row: template of the first component from Figure 2 with deformations drawn with respect to the second component covariance matrix.

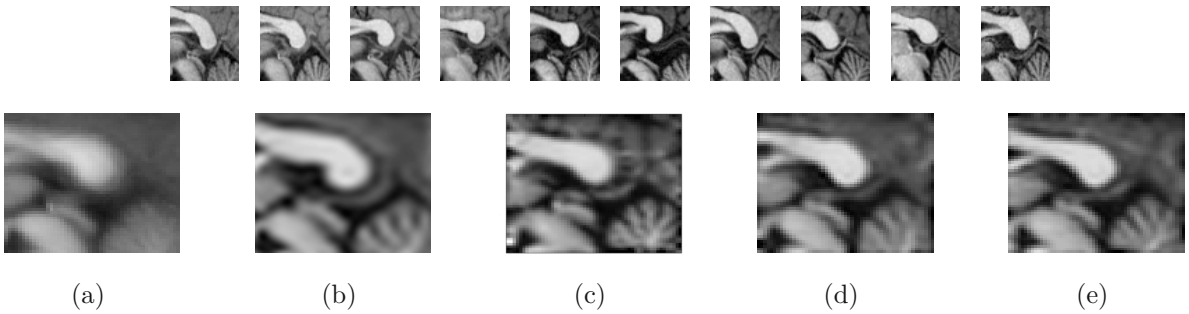
to converge. The huge geometric variability adding to very complex shapes for the templates lead to a more difficult estimation and therefore more iterations before convergence.

Last point that can be noticed is the convergence of the European 7 which looks slower than the other component (American 7). The reason of this behaviour is that there are only two images of such a 7 in the training set and it takes a longer time for the algorithm to put together and only together these two shapes so that the clustering is better with respect to the likelihood. The other 7 does not suffer from this problem and converges faster.

5.1.5. *The estimated geometric distribution*

To be able to compare the learnt geometry, we draw some synthetic examples using the mixture model with the learnt parameters. Even when the templates look similar, the separation between two components can be justified by the different geometry distributions. To show the effects of the geometry on the components, we have drawn some "2" with their respective parameters in the four top rows of Figure 5.

For each component, we have drawn the deformation given by the variable β and its opposite $-\beta$ since, as soon as one is learnt, because of the symmetry of the centred Gaussian distribution, the opposite deformation is learnt at the same time. This is why sometimes, one of the two looks strange whereas the other looks like some element of the training set.



(a) (b) (c) (d) (e)

FIGURE 6. First row: ten images of the training set representing the splenium and a part of the cerebellum. Second row: Results from the template estimation. (a) gray level mean image of the 47 images. Templates estimated (b) with the FAM (c) with the stochastic algorithm on the simple model (d,e) on the two component model.

The simulation is done using a common standard Gaussian distribution which is then multiplied by a square root of the covariance matrix we want to apply. We can see the effects of the covariance matrix on both templates and the large variability learnt. This has to be compared with the bottom rows of Figure 5, where the two samples are drawn on the one template but with the covariance matrix of the other one. Even if these six lines represent some "2"s, the bottom ones suffer from the geometrical tendency of the other cluster and do not look as natural. This shows the variability of the models into classes.

5.2. **Medical images**

We also test the algorithm on a database which consists in 47 2D medical images. Each of them represents the splenium (back of the corpus calosum) and a part of the cerebellum. Some of the training images are shown in Figure 6 first row.

The results of the estimation are presented in Figure 6, second row. The first picture presented, (a), is the gray level mean of the 47 images. The second one, (b), shows the estimated template computed with the Fast Approximation with Mode Algorithm presented in [1] for a single component model. This algorithm is an EM-like algorithm where the E step is simplified. The posterior distribution of the hidden variable is approximated by a Dirac distribution on its mode. This yields a deterministic algorithm, quite simple to implement but with no theoretical convergence properties. It shows a well contrasted splenium whereas the cerebellum remains a little bit blurry (note that it is still much better that the simple mean (a)).

This picture has to be compared with picture (c) which gives the estimated template computed with our algorithm with $\tau_m = 1$. The great improvement from the gray level mean of the images (a) or the FAM estimation (b) to our estimations is obvious. In particular, the splenium is still very contrasted, better localised and the cerebellum is reconstructed with several branches. The background presents several structures whereas the other estimates are blurry. The two anatomical shapes are relevant representants of the ones observed in the training set.

The estimation has been done while enabling the decomposition of the database into two components with our SAEM-MCMC algorithm presented here. The two estimated templates are shown in Figure 6(d) and 6(e). The differences can be seen in particular on the shape of the splenium where the boundaries are more or less curved. The thickness of the splenium varies as well between the two estimates. The position of the fornix is also different, being closer to the boundary of the image. The number of branches in the two cerebella also tends to be different from one template to the other (4 in the first component and 5 in the second one).

The estimation suffers from the small number of images we have. This can be seen in the estimation of the background which is blurry in both images. To be able to explain the huge variability of the two

anatomical shapes, more components would be interesting but at the same time more images required so that the components will not end up empty.

6. Proof of Theorem 4.2

We recall that in this section the variances of the components are fixed. This reduces the parameters θ_t to $(\alpha_t, \Gamma_{g,t})$ for all $1 \leq t \leq \tau_m$.

First let exhibit sufficient statistics for the model. The complete log-likelihood equals:

$$\begin{aligned}\log q(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}|\eta) &= \sum_{i=1}^{n} \left\{ \log \left[\left(\frac{1}{2\pi\sigma_{\tau_i}^2} \right)^{|\Lambda|/2} \exp\left(-\frac{1}{2\sigma_{\tau_i}^2} \|y_i - K_p^{\beta_i} \alpha_{\tau_i}\|^2 \right) \right] \right. \\ &\quad \left. + \log \left[\left(\frac{1}{2\pi} \right)^{k_g} |\Gamma_{g,\tau_i}|^{-1/2} \exp\left(-\frac{1}{2} \beta_i^t \Gamma_{g,\tau_i}^{-1} \beta_i \right) \right] + \log(\rho_{\tau_i}) \right\},\end{aligned}$$

where $K_p^{\beta}\alpha = z_\beta I_\alpha$ and $\|.\|$ denotes the Euclidean norm. This emphasises five sufficient statistics given in their matricial form for all $1 \leq t \leq \tau_m$,

$$\begin{aligned}S_{0,t}(\boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\tau_i = t}\,, \\ S_{1,t}(\boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\tau_i = t} \left(K_p^{\beta_i}\right)^t y_i\,, \\ S_{2,t}(\boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\tau_i = t} \left(K_p^{\beta_i}\right)^t \left(K_p^{\beta_i}\right)\,, \\ S_{3,t}(\boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\tau_i = t} \beta_i^t \beta_i\,, \mathbb{P} \\ S_{4,t}(\boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\tau_i = t} \|y_i\|^2.\end{aligned}$$

Thus we apply the stochastic approximation at iteration k of the algorithm leading to:

$$s_{k,m,t} = s_{k-1,m,t} + \Delta_k (S_{m,t}(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - s_{k-1,m,t})$$

for $0 \leq m \leq 4$ and rewrite the maximisation step. The weights and the covariance matrix are updated as follows:

$$\rho_{\tau,k} = \frac{s_{k,0,\tau} + a_\rho}{n + \tau_m a_\rho}\,, \tag{6.1}$$

$$\Gamma_{g,\tau,k} = \frac{1}{s_{k,0,\tau} + a_g} (s_{k,0,\tau} s_{k,3,\tau} + a_g \Sigma_g)\,. \tag{6.2}$$

The photometric parameters are solution of the following system:

$$\begin{cases} \alpha_{\tau,k} = \left(s_{k,0,\tau} s_{k,2,\tau} + \sigma_{\tau,k}^2 (\Sigma_p)^{-1} \right)^{-1} \left(s_{k,0,\tau} s_{k,1,\tau} + \sigma_{\tau,k}^2 (\Sigma_p)^{-1} \mu_p \right)\,, \\ \sigma_{\tau,k}^2 = \frac{1}{s_{k,0,\tau}|\Lambda| + a_p} \left(s_{k,0,\tau} \left(s_{k,4,\tau} + (\alpha_{\tau,k})^t s_{k,2,\tau} \alpha_{\tau,k} - 2(\alpha_{\tau,k})^t s_{k,1,\tau} \right) + a_p \sigma_0^2 \right)\,, \end{cases} \tag{6.3}$$

which can be solved iteratively for each component τ starting with the previous values. In this part, all $\sigma_{\tau,k}^2$ are fixed and this leads to an explicit form for the parameters $\alpha_{\tau,k}$.

We will now apply Theorem 4.1 to prove Theorem 4.2.

(SA0') is satisfied with the set $\mathcal{S}_a$ defined by

$$\mathcal{S}_a \triangleq \{ S \in \mathcal{S} \mid 0 \leq S_{0,t} \leq n,\ \|S_{1,t}\| \leq \|\boldsymbol{y}\|,\ \|S_{2,t}\| \leq n,\ 0 \leq S_{3,t},\ 0 \leq S_{4,t} \leq \|\boldsymbol{y}\|^2,\ \forall 1 \leq t \leq \tau_m \} .$$

Thanks to the convexity of this set, the new value s_k defined as a barycenter remains in $\mathcal{S}_a$.

Assumption (SA1) is trivially satisfied since we can choose our step-size sequence $(\Delta_k)_k$.

(SA2) holds as already proved in [3] for the one component case with $w(s) = -l(\hat{\eta}(s))$ such as (STAB1'(ii)) with the same function $W(s) = -l(\hat{\eta}(s))$. These conditions imply the contraction property of the Lyapunov function w and the convergence of the stochastic approximation under some conditions on the perturbations.

We need to suppose, like in the one component case [3], that the critical points of our model are in a compact subset of $\mathcal{S}$ which stands for (STAB1'(i)). This is an assumption which has to be considered in a future work.

We will now focus on (STAB2) which is the assumption which gives the control of the perturbations required for the convergence.

We first show the convergence to zero of the remainder term $|r_k| \mathbb{1}_{W(s_{k-1}) \leq M}$ for any positive integer M. We denote by $\pi_k = \pi_{\hat{\eta}(s_k)}$ for any $k \geq 0$. We have $r_k = \mathbb{E}\left[S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) | \mathcal{F}_{k-1}\right] - \mathbb{E}_{\pi_{k-1}}\left[S(\boldsymbol{\beta}, \boldsymbol{\tau})\right]$ thus,

$$\begin{aligned} r_k &= \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \prod_{i=1}^{n} \int p_{J_k}(\tau_i | \xi_i, y_i, \eta_{k-1}) \prod_{t=1}^{\tau_m} \prod_{l=1}^{J_k} \Pi_{\eta,t}(\xi_{t,i}^{(l-1)}, \xi_{t,i}^{(l)}) \mathrm{d}\xi_{t,i}^{(l)} \mathrm{d}\boldsymbol{\beta} \\ &\quad - \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) \pi_{\eta_{k-1}}(\boldsymbol{\beta}, \boldsymbol{\tau}) \mathrm{d}\boldsymbol{\beta}. \end{aligned}$$

We denote by $Q(\xi_i)\mathrm{d}\xi_i = \prod_{t=1}^{\tau_m} \prod_{l=1}^{J_k} \Pi_{\eta,t}(\xi_{t,i}^{(l-1)}, \xi_{t,i}^{(l)})\ \mathrm{d}\xi_{t,i}^{(l)}$ and by
$R_{J_k}(\boldsymbol{\tau} | y, \eta_{k-1}) = \prod_{i=1}^{n} \int p_{J_k}(\tau_i | \xi_i, y_i, \eta_{k-1})\ Q(\xi_i)\ \mathrm{d}\xi_i$. We can now rewrite

$$\begin{aligned} |r_k| &\leq \left| \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) \left[\Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) R_{J_k}(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) \mathrm{d}\boldsymbol{\beta} - \pi_{\eta_{k-1}}(\boldsymbol{\beta}, \boldsymbol{\tau}) \right] \mathrm{d}\boldsymbol{\beta} \right| \\ &\leq \sum_{\boldsymbol{\tau}} \left| \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) \left[\Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - q(\boldsymbol{\beta} | \boldsymbol{\tau}, \boldsymbol{y}, \eta_{k-1}) \right] \mathrm{d}\boldsymbol{\beta} \right| \left| R_{J_k}(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) \right| \\ &\quad + \sum_{\boldsymbol{\tau}} \left| \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) q(\boldsymbol{\beta} | \boldsymbol{\tau}, \boldsymbol{y}, \eta_{k-1}) \mathrm{d}\boldsymbol{\beta} \right| \left| R_{J_k}(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) - q(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) \right| . \end{aligned}$$

Denoting $\mathcal{M}_{\eta_{k-1}} = \max_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} |S(\boldsymbol{\beta}, \boldsymbol{\tau})| q(\boldsymbol{\beta} | \boldsymbol{\tau}, \boldsymbol{y}, \eta_{k-1}) \mathrm{d}\boldsymbol{\beta}$, we obtain finally

$$|r_k| \mathbb{1}_{W(s_{k-1}) \leq M} \leq \sum_{\boldsymbol{\tau}} \left| \int_{\mathbb{R}^N} S(\boldsymbol{\beta}, \boldsymbol{\tau}) \left[\Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - q(\boldsymbol{\beta} | \boldsymbol{\tau}, \boldsymbol{y}, \eta_{k-1}) \right] \mathrm{d}\boldsymbol{\beta} \right| \mathbb{1}_{W(s_{k-1}) \leq M} \tag{6.4}$$

$$+ \mathcal{M}_{\eta_{k-1}} \sum_{\boldsymbol{\tau}} \left| R_{J_k}(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) - q(\boldsymbol{\tau} | \boldsymbol{y}, \eta_{k-1}) \right| \mathbb{1}_{W(s_{k-1}) \leq M}. \tag{6.5}$$

We will first show that the Gibbs sampler kernel $\Pi_{\eta, \boldsymbol{\tau}}$ satisfies a lower bound condition and a Drift condition (MDRI) to get its geometric ergodicity (as it has been done in [3]).

(MDRI): For any $s \in \mathcal{S}$ and any $\boldsymbol{\tau} \in \mathcal{T}$, $\Pi_{\hat{\eta}(s), \boldsymbol{\tau}}$ is irreducible and aperiodic. In addition there exists a function $V : \mathbb{R}^N \to [1, \infty[$ such that for any $p \geq 1$ and any compact subset $\mathcal{K} \subset \mathcal{S}$, there exist a set $\mathtt{C}$,

an integer m, constants $0 < \kappa < 1$, $B > 0$, $\delta > 0$ and a probability measure ν such that

$$\inf_{s\in\mathcal{K},\boldsymbol{\tau}\in\mathcal{T}} \Pi^m_{\hat{\eta}(s),\boldsymbol{\tau}}(\boldsymbol{\beta}, A) \geq \delta\nu(A) \quad \forall \boldsymbol{\beta} \in \mathtt{C}, \forall A \in \mathcal{B}(\mathbb{R}^N)\,, \tag{6.6}$$

$$\sup_{s\in\mathcal{K},\boldsymbol{\tau}\in\mathcal{T}} \Pi^m_{\hat{\eta}(s),\boldsymbol{\tau}} V^p(\boldsymbol{\beta}) \leq \kappa V^p(\boldsymbol{\beta}) + B\mathbb{1}_{\mathtt{C}}(\boldsymbol{\beta}). \tag{6.7}$$

Notation 6.1. Let $(e_j)_{1\leq j\leq N}$ be the canonical basis of the $\boldsymbol{\beta}$-space and for any $1 \leq j \leq N$, let $E_{\eta,\boldsymbol{\tau},j} \triangleq \{\, \boldsymbol{\beta} \in \mathbb{R}^N \mid \langle \boldsymbol{\beta}, e_j\rangle_{\eta,\boldsymbol{\tau}} = 0\}$ be the orthogonal of $\mathrm{Span}\{e_j\}$ and $p_{\eta,\boldsymbol{\tau},j}$ be the orthogonal projection on $E_{\eta,\boldsymbol{\tau},j}$ *i.e.*

$$p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) \triangleq \boldsymbol{\beta} - \frac{\langle \boldsymbol{\beta}, e_j\rangle_{\eta,\boldsymbol{\tau}}}{\|e_j\|^2_{\eta,\boldsymbol{\tau}}} e_j\,,$$

where $\langle \boldsymbol{\beta}, \boldsymbol{\beta}'\rangle_{\eta,\boldsymbol{\tau}} = \sum_{i=1}^n \beta_i^t \Gamma^{-1}_{g,\tau_i} \beta'_i$ for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ in $\mathbb{R}^N$ (*i.e.* the natural dot product associated with the covariance matrices $(\Gamma_{g,t})_t$) and $\|.\|_{\eta,\tau}$ is the corresponding norm.

We denote for any $1 \leq j \leq N$, $\eta \in \Theta \times \varrho$ and $\boldsymbol{\tau} \in \mathcal{T}$, by $\Pi_{\eta,\boldsymbol{\tau},j}$ the Markov kernel on $\mathbb{R}^N$ associated with the j-th Metropolis-Hastings step of the Gibbs sampler on $\mathbb{R}^N$. We have $\Pi_{\eta,\boldsymbol{\tau}} = \Pi_{\eta,\boldsymbol{\tau},N} \circ \cdots \circ \Pi_{\eta,\boldsymbol{\tau},1}$.

Inequality (6.6) is equivalent to the existence of a small set $\mathtt{C}$ for the kernel $\Pi_{\hat{\eta}(s),\tau}$ independent of $s \in \mathcal{K}$. We recall here the definition of a small set:

Definition 6.1 (*cf.* [17])**.** A set $\mathcal{E} \in \mathcal{B}(\mathbb{R}^N)$ is called a **small set** for the kernel Π if there exist an integer $m > 0$ and a non trivial measure ν_m on $\mathcal{B}(\mathbb{R}^N)$, such that for all $\boldsymbol{\beta} \in \mathcal{E}$, $B \in \mathcal{B}(\mathbb{R}^N)$, $\Pi^m(\boldsymbol{\beta}, B) \geq \nu_m(B)$.

When this holds, we say that $\mathcal{E}$ is ν_m-small.

We now prove the following lemma:

Lemma 6.1. *Let $\mathcal{E}$ be a compact subset of $\mathbb{R}^N$ and $\mathcal{K}$ be a compact subset of $\mathcal{S}$, then $\mathcal{E}$ is a small set of $\mathbb{R}^N$ for $(\Pi_{\hat{\eta}(s),\boldsymbol{\tau}})_{s\in\mathcal{K},\boldsymbol{\tau}\in\mathcal{T}}$.*

Proof. The transition probability kernel of our Markov chain on $\boldsymbol{\beta}$ is defined as follows: for coordinate j, the kernel is

$$\Pi_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}, \mathrm{d}\boldsymbol{z}) = (\otimes_{m\neq j}\delta_{\boldsymbol{\beta}^m}(\mathrm{d}\boldsymbol{z}^m))\, \mathbb{P}\left[q_j(\mathrm{d}\boldsymbol{z}^j|\boldsymbol{\beta}^{-j}, \eta, \boldsymbol{\tau}) r_j(\boldsymbol{\beta}^j, \mathrm{d}\boldsymbol{z}^j; \boldsymbol{\beta}^{-j}, \eta, \boldsymbol{\tau}) \right.$$
$$\left. + \delta_{\boldsymbol{\beta}^j}(\mathrm{d}\boldsymbol{z}^j) \int (1 - r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \eta, \boldsymbol{\tau}))q_j(b|\boldsymbol{\beta}^{-j}, \eta, \boldsymbol{\tau})\mathrm{d}b\right]. \tag{6.8}$$

Then note that there exists $a_c > 0$ such that for any $\eta \in \Theta \times \varrho$, any $\boldsymbol{\beta} \in \mathbb{R}^N$ and any $b \in \mathbb{R}$, the acceptance rate $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \eta, \boldsymbol{\tau})$ is uniformly lower bounded by a_c so that for any $1 \leq j \leq N$ and any non-negative function f,

$$\Pi_{\eta,\boldsymbol{\tau},j} f(\boldsymbol{\beta}) \geq a_c \int_{\mathbb{R}} f(\boldsymbol{\beta}^{-j} + be_j) q_j(b|\boldsymbol{\beta}^{-j}, \boldsymbol{\tau}, \eta)\mathrm{d}b = a_c \int_{\mathbb{R}} f(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) + ze_j/\|e_j\|_{\eta,\boldsymbol{\tau}}) g_{0,1}(z)\mathrm{d}z,$$

where $g_{0,1}$ is the probability density function of the standard $\mathcal{N}(0, 1)$. By induction, we have

$$\Pi_{\eta,\boldsymbol{\tau}} f(\boldsymbol{\beta}) \geq a_c^N \int_{\mathbb{R}^N} f\left(p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}) + \sum_{j=1}^N z_j p_{\eta,\boldsymbol{\tau},j+1,N}(e_j)/\|e_j\|_{\eta,\boldsymbol{\tau}}\right) \prod_{j=1}^N g_{0,1}(z_j)\mathrm{d}z_j\,, \tag{6.9}$$

where $p_{\eta,\boldsymbol{\tau},q,r} = p_{\eta,\boldsymbol{\tau},r} \circ p_{\eta,\boldsymbol{\tau},r-1} \circ \ldots \circ p_{\eta,\boldsymbol{\tau},q}$ for any integer $q \leq r$ and $p_{\eta,\boldsymbol{\tau},N+1,N} = Id_{\mathbb{R}^N}$. Let $A_{\eta,\boldsymbol{\tau}} \in \mathcal{L}(\mathbb{R}^N)$ be the linear mapping on $z_1^N = (z_1, \ldots, z_N)$ defined by

$$A_{\eta,\boldsymbol{\tau}} z_1^N = \sum_{j=1}^N z_j p_{\eta,\boldsymbol{\tau},j+1,N}(e_j)/\|e_j\|_{\eta,\boldsymbol{\tau}}.$$

One easily checks that for any $1 \leq k \leq N$, $\mathrm{Span}\{\ p_{\eta,\boldsymbol{\tau},j+1,N}(e_j),\ k \leq j \leq N\} = \mathrm{Span}\{e_j, k \leq j \leq N\}$ so that $A_{\eta,\boldsymbol{\tau}}$ is an invertible mapping. By a change of variable, we get

$$\int_{\mathbb{R}^N} f\left(p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}) + A_{\eta,\boldsymbol{\tau}} z_1^N\right) \prod_{j=1}^N g_{0,1}(z_j) \mathrm{d}z_j = \int_{\mathbb{R}^N} f(u) g_{p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}), A_{\eta,\boldsymbol{\tau}} A_{\eta,\boldsymbol{\tau}}^t}(u) \mathrm{d}u,$$

where $g_{\mu,\Sigma}$ stands for the probability density function of the normal law $\mathcal{N}(\mu, \Sigma)$.

Since $(\eta, \boldsymbol{\tau}) \to A_{\eta,\boldsymbol{\tau}}$ is smooth on the set of invertible mappings in $(\eta, \boldsymbol{\tau})$, we deduce that there exist $c_{\mathcal{K}} > 0$ and $C_{\mathcal{K}} > 0$ such that $c_{\mathcal{K}} \mathrm{Id} \leq A_{\eta,\boldsymbol{\tau}} A_{\eta,\boldsymbol{\tau}}^t \leq \mathrm{Id}/c_{\mathcal{K}}$ and $g_{p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}), A_{\eta,\boldsymbol{\tau}} A_{\eta,\boldsymbol{\tau}}^t}(u) \geq C_{\mathcal{K}} g_{p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}), \mathrm{Id}/c}(u)$ uniformly for $\eta = \hat{\eta}(s)$ with $s \in \mathcal{K}$ and $\boldsymbol{\tau} \in \mathcal{T}$. Assuming that $\boldsymbol{\beta} \in \mathcal{E}$, since $\eta \to p_{\eta,\boldsymbol{\tau},1,N}$ is smooth and $\mathcal{E}$ is compact, we have $\sup_{\boldsymbol{\beta} \in \mathcal{E}, \eta = \hat{\eta}(s),\ s \in \mathcal{K}, \boldsymbol{\tau} \in \mathcal{T}} \|p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta})\| < \infty$ so that there exist other constants $C_{\mathcal{K}} > 0$ and $c_{\mathcal{K}} > 0$ such that for any $(u, \boldsymbol{\beta}) \in \mathbb{R}^N \times \mathcal{E}$ and any $\eta = \hat{\eta}(s),\ s \in \mathcal{K}, \boldsymbol{\tau} \in \mathcal{T}$

$$g_{p_{\eta,\boldsymbol{\tau},1,N}(\boldsymbol{\beta}), A_{\eta,\boldsymbol{\tau}} A_{\eta,\boldsymbol{\tau}}^t}(u) \geq C_{\mathcal{K}} g_{0, \mathrm{Id}/c_{\mathcal{K}}}(u). \tag{6.10}$$

Using (6.9) and (6.10), we deduce that for any A $\Pi_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}, A) \geq C_{\mathcal{K}} a_c^N \nu(A)$, with ν equal to the density of the normal law $\mathcal{N}(0, \mathrm{Id}/c_{\mathcal{K}})$. This yields the existence of the small set as well as equation (6.6). □

This property also implies the ϕ-irreducibility of the Markov Chain generated by $\Pi_{\eta,\boldsymbol{\tau}}$. Moreover, the existence of a ν_1-small set implies the aperiodicity of the chain (*cf.* [17]).

Now consider the Drift condition (6.7).

We set $V : \mathbb{R}^N \to [1, +\infty[$ as the following function $V(\boldsymbol{\beta}) = 1 + \|\boldsymbol{\beta}\|^2$,where $\|\cdot\|$ denotes the Euclidean norm. Define for any $g : \mathbb{R}^N \to \mathbb{R}^{n_s}$ the norm $\|g\|_V = \sup_{\boldsymbol{\beta} \in \mathbb{R}^N} \frac{\|g(\boldsymbol{\beta})\|}{V(\boldsymbol{\beta})}$ and the functional space $\mathcal{L}_V = \{g : \mathbb{R}^N \to \mathbb{R}^{n_s} \mid \|g\|_V < +\infty\}$. For any $\eta \in \Theta \times \varrho$ and any $\boldsymbol{\tau} \in \mathcal{T}$, we introduce a $(\eta, \boldsymbol{\tau})$ dependent function $V_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \triangleq 1 + \|\boldsymbol{\beta}\|_{\eta,\boldsymbol{\tau}}^2$.

Lemma 6.2. *Let K be a compact subset of $\Theta \times \varrho$. For any integer $p \geq 1$, there exist $0 < \rho < 1$ and $C > 0$ such that for any $\eta \in K$, any $\boldsymbol{\tau} \in \mathcal{T}$, any $\boldsymbol{\beta} \in \mathbb{R}^N$ we have*

$$\Pi_{\eta,\boldsymbol{\tau}} V_{\eta,\boldsymbol{\tau}}^p(\boldsymbol{\beta}) \leq \rho V_{\eta,\boldsymbol{\tau}}^p(\boldsymbol{\beta}) + C\,.$$

Proof. The proposal distribution for $\Pi_{\eta,\boldsymbol{\tau},j}$ is given by $q_j(\boldsymbol{\beta}^j \mid \boldsymbol{\beta}^{-j}, \boldsymbol{\tau}, \boldsymbol{y}, \eta) \overset{\text{law}}{=} p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) + \frac{z}{\|e_j\|_{\eta,\tau}} e_j$, where $z \sim \mathcal{N}(0,1)$. Then, for any $\boldsymbol{\beta} \in \mathbb{R}^N$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^N)$, there exists $a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})$ uniformly bounded from below by $a_c > 0$ such that

$$\Pi_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}, A) = (1 - a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) \mathbb{1}_A(\boldsymbol{\beta}) + a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) \int_{\mathbb{R}} \mathbb{1}_A\left(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) + \frac{z}{\|e_j\|_{\eta,\tau}} e_j\right) g_{0,1}(\mathrm{d}z),$$

Since $\langle p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}), e_j\rangle_{\eta,\boldsymbol{\tau}} = 0$, we get $V_{\eta,\boldsymbol{\tau}}\left(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) + \frac{z}{\|e_j\|_{\eta,\tau}} e_j\right) = V_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) + z^2$. We deduce that there exists C such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$:

$$\begin{aligned}
\Pi_{\eta,\boldsymbol{\tau},j} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) &= (1 - a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}))V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) \int_{\mathbb{R}} \left(V_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) + z^2\right)^p g_{0,1}(\mathrm{d}z) \\
&\leq (1 - a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}))V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})\mathbb{P} \\
&\quad \left(V^p_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) + V^{p-1}_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) \int_{\mathbb{R}} (1 + z^2)^p g_{0,1}(\mathrm{d}z)\right) \\
&\leq (1 - a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}))V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})V^p_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) + CV^{p-1}_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})).
\end{aligned}$$

We have used in the last inequality the fact that a Gaussian variable has bounded moment of any order. Since $a_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) \geq a_c$ and $\|p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})\|_{\eta,\boldsymbol{\tau}} \leq \|\boldsymbol{\beta}\|_{\eta,\boldsymbol{\tau}}$ ($p_{\eta,\boldsymbol{\tau},j}$ is an orthonormal projection for the dot product $\langle\cdot,\cdot\rangle_{\eta,\boldsymbol{\tau}}$), we get that for any $\varepsilon > 0$, there exists $C_{K,\varepsilon}$ such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$ and $\eta \in K, \boldsymbol{\tau} \in \mathcal{T}$

$$\Pi_{\eta,\boldsymbol{\tau},j} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq (1 - a_c)V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + (a_c + \varepsilon)V^p_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta})) + C_{K,\varepsilon}.$$

By induction, we get

$$\Pi_{\eta,\boldsymbol{\tau}} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq \sum_{u\in\{0,1\}^N} \prod_{j=1}^{N} (1 - a_c)^{1-u_j}(a_c + \varepsilon)^{u_j} V^p_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau},u}(\boldsymbol{\beta})) + \frac{C_{K,\varepsilon}}{\varepsilon}\left((1 + \varepsilon)^{N+1} - 1\right),$$

where $p_{\eta,\boldsymbol{\tau},u} = ((1 - u_N)\mathrm{Id} + u_N p_{\eta,\boldsymbol{\tau},N}) \circ \ldots \circ ((1 - u_1)\mathrm{Id} + u_1 p_{\eta,\boldsymbol{\tau},1})$.

Let $p_{\eta,\boldsymbol{\tau}} = p_{\eta,\boldsymbol{\tau},N} \circ \cdots \circ p_{\eta,\boldsymbol{\tau},1}$ and note that $p_{\eta,\boldsymbol{\tau},u}$ is contracting so that

$$\Pi_{\eta,\boldsymbol{\tau}} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq b_{c,\varepsilon} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + (a_c + \varepsilon)^N V^p_{\eta,\boldsymbol{\tau}}(p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta})) + \frac{C_{K,\varepsilon}}{\varepsilon}\left((1 + \varepsilon)^{N+1}\right)$$

for $b_{c,\varepsilon} = \left(\sum_{u\in\{0,1\}^N,\ u\neq\mathbf{1}} \prod_{j=1}^{N} (1 - a_c)^{1-u_j}(a_c + \varepsilon)^{u_j}\right)$. To end the proof, we need to check that $p_{\eta,\boldsymbol{\tau}}$ is strictly contracting uniformly on K. Indeed, $\|p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta})\|_{\eta,\boldsymbol{\tau}} = \|\boldsymbol{\beta}\|_{\eta,\boldsymbol{\tau}}$ implies that $p_{\eta,\boldsymbol{\tau},j}(\boldsymbol{\beta}) = \boldsymbol{\beta}$ for any $1 \leq j \leq N$ so that $\langle\boldsymbol{\beta}, e_j\rangle_{\eta,\boldsymbol{\tau}} = 0$ and $\boldsymbol{\beta} = 0$ since $(e_j)_{1\leq j\leq N}$ is a basis. Using the continuity of the norm of $p_{\eta,\boldsymbol{\tau}}$ and the compactness of K, we deduce that there exists $0 < \rho_K < 1$ such that $\|p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta})\|_{\eta,\boldsymbol{\tau}} \leq \rho_K \|\boldsymbol{\beta}\|_{\eta,\boldsymbol{\tau}}$ for any $\boldsymbol{\beta} \in \mathbb{R}^N$, $\eta \in K$ and any $\boldsymbol{\tau} \in \mathcal{T}$. Changing ρ_K for $1 > \rho'_K > \rho_K$ we get $(1 + \rho_K^2\|\boldsymbol{\beta}\|^2_{\eta,\boldsymbol{\tau}})^p \leq \rho'^{2p}_K(1 + \|\boldsymbol{\beta}\|^2_{\eta,\boldsymbol{\tau}})^p + C_K$ for some uniform constant C_K so that

$$\Pi_{\eta,\boldsymbol{\tau}} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq b_{c,\varepsilon} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + \rho'^{2p}_K (a_c + \varepsilon)^N V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + C_{K,\varepsilon}.$$

Since we have $\inf_{\varepsilon>0}\left\{b_{c,\varepsilon} + \rho'^{2p}_K (a_c + \varepsilon)^N\right\} < 1$ the result is straightforward. □

Lemma 6.3. *For any compact set $K \subset \Theta\mathbb{P}\varrho$, any integer $p \geq 0$, there exist $0 < \rho < 1$, $C > 0$ and a positive integer m_0 such that $\forall m \geq m_0$, $\forall \eta \in K$, $\forall \boldsymbol{\beta} \in \mathcal{T}$*

$$\Pi^m_{\eta,\boldsymbol{\tau}} V^p(\boldsymbol{\beta}) \leq \rho V^p(\boldsymbol{\beta}) + C.$$

Proof. Indeed, there exist $0 \leq c_1 \leq c_2$ such that $c_1 V(\boldsymbol{\beta}) \leq V_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq c_2 V(\boldsymbol{\beta})$ for any $(\boldsymbol{\beta}, \eta, \boldsymbol{\tau}) \in \mathbb{R}^N \times K \times \mathcal{T}$. Then, using the previous lemma, we have $\Pi^m_{\eta,\boldsymbol{\tau}} V^p(\boldsymbol{\beta}) \leq c_1^{-p}\Pi^m_{\eta,\boldsymbol{\tau}} V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) \leq c_1^{-p}(\rho^m V^p_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}) + C/(1 - \rho)) \leq (c_2/c_1)^p(\rho^m V^p(\boldsymbol{\beta}) + C/(1 - \rho))$. Choosing m large enough for $(c_2/c_1)^p\rho^m < 1$ gives the result. □

This finishes the proof of (6.7) and in the same time the (MDRI).

Thanks to this property we can use the following proposition (*cf.* [5,17] Prop. B1) and lemma applied to every sequence $(\xi_{t,i}^{(l)})_l$ with stationary distribution $q(\cdot|y_i,t,\eta)$ for all $1 \leq t \leq \tau_m$ and all $1 \leq i \leq n$.

Proposition 6.1. *Suppose that* Π *is irreducible and aperiodic and that* $\Pi^m(\boldsymbol{\beta}_0,.) \geq \mathbb{1}_{\mathcal{C}}(\boldsymbol{\beta}_0)\delta\nu(.)$ *for a set* $\mathcal{C} \in \mathcal{B}(\mathbb{R}^N)$*, some integer* m *and* $\delta > 0$ *and that there is a Drift condition to* $\mathcal{C}$ *in the sense that, for some* $0 < \kappa < 1$*,* $B > 0$ *and a function* $V : \mathbb{R}^N \to [1, +\infty[$*,*

$$\Pi V(\boldsymbol{\beta}_0) \leq \kappa V(\boldsymbol{\beta}_0)\ \forall \boldsymbol{\beta}_0 \notin \mathcal{C} \text{ and } \sup_{\boldsymbol{\beta}_0 \in \mathcal{C}} (V(\boldsymbol{\beta}_0) + \Pi V(\boldsymbol{\beta}_0)) \leq B.$$

Then, there exist constants K *and* $0 < \rho < 1$*, depending only upon* m, δ, κ, B*, such that, for all* $\boldsymbol{\beta}_0 \in \mathbb{R}^N$*, and all* $g \in \mathcal{L}_V$

$$\|\Pi^n g(\boldsymbol{\beta}_0) - \pi(g)\|_V \leq K\rho^n \|g\|_V.$$

Lemma 6.4. *Assume that there exist an integer* m *and constants* $0 < \kappa < 1$ *and* $\varsigma > 0$ *and a set* $\mathcal{C}$ *such that*

$$\Pi^m V(\boldsymbol{\beta}_0) \leq \kappa V(\boldsymbol{\beta}_0)\ \forall \boldsymbol{\beta}_0 \notin \mathcal{C} \text{ and } \Pi V(\boldsymbol{\beta}_0) \leq \varsigma V(\boldsymbol{\beta}_0)\ \forall \boldsymbol{\beta}_0 \in \mathbb{R}^N$$

for some function $V : \mathbb{R}^N \to [1, +\infty[$*. Then there exists a function* $\tilde{V}$ *and constants* $0 < \rho < 1, c > 0$ *and* $C > 0$*, depending only upon* m, κ, ς*, such that,*

$$\Pi\tilde{V}(\boldsymbol{\beta}_0) \leq \rho\tilde{V}(\boldsymbol{\beta}_0)\ \forall \boldsymbol{\beta}_0 \notin \mathcal{C} \text{ and } cV \leq \tilde{V} \leq CV\,.$$

Proof. Define

$$\tilde{V} = \sum_{j=1}^{m} \kappa^{1-j/m} \Pi^{j-1} V\,.$$

For $\boldsymbol{\beta}_0 \notin \mathtt{C}$, we have

$$\begin{aligned} \Pi\tilde{V}(\boldsymbol{\beta}_0) &\leq \sum_{j=1}^{m-1} \kappa^{1-j/m} \Pi^j V(\boldsymbol{\beta}_0) + \kappa V(\boldsymbol{\beta}_0) \\ &\leq \kappa^{1/m} \tilde{V}(\boldsymbol{\beta}_0)\,. \end{aligned}$$

Therefore we obtain:

$$\kappa^{1-1/m} V \leq \tilde{V} \leq \left(\sum_{j=1}^{m} \kappa^{1-j/m} \varsigma^{j-1} \right) V\,.$$

This ends the proof of Lemma 6.4. □

Thus, applying the Proposition 6.1 and Lemma 6.4 to the Drift conditions of Lemmas 6.2 and 6.3, we get that each Gibbs sampler kernel $\Pi_{\eta,\boldsymbol{\tau}}$ is geometrically ergodic.

Let us now go back to the convergence of the first part of the residual term (6.4) towards 0.

We use the term $\mathbb{1}_{W(s_{k-1}) \leq M}$ to show that the parameters η_{k-1} are constrained to move in a compact set of $\Theta \times \varrho$. We show first that the observed log-likelihood l tends to minus infinity as the parameters tend to the boundary of $\Theta \times \varrho$. Equation (2.1) implies that for any $\theta \in \Theta$ we have:

$$q(y_i|\beta_i, \tau_i, \alpha, \sigma^2) q(\beta_i|\Gamma_{g,\tau_i}) \leq (2\pi\sigma^2)^{-|\Lambda|/2} (2\pi)^{-k_g} |\Gamma_{g,\tau_i}|^{-1/2} \exp\left(-\frac{1}{2} \beta_i^t \Gamma_{g,\tau_i}^{-1} \beta_i \right),$$

so that denoting C as a constant:

$$\log(q(\boldsymbol{y},\eta)) \leq \sum_{i=1}^{n}\left[-\frac{a_g}{2}\langle\Gamma_{g,\tau_i}^{-1},\Sigma_g\rangle_F + \frac{1+a_g}{2}\log|\Gamma_{g,\tau_i}^{-1}| - \frac{a_p\sigma_0^2}{2\sigma_{\tau_i}^2} - \frac{|\Lambda|+a_p}{2}\log(\sigma_{\tau_i}^2)\right.$$
$$\left. -\frac{1}{2}(\alpha_{\tau_i}-\mu_p)^t\Sigma_p^{-1}(\alpha_{\tau_i}-\mu_p) - a_\rho\log\rho_{\tau_i}\right] + C.$$

It was shown in [1] that we have $\lim_{||\Gamma||+||\Gamma^{-1}||\to\infty} -\frac{a_g}{2}\langle\Gamma^{-1},\Sigma_g\rangle_F + \frac{1+a_g}{2}\log|\Gamma^{-1}| = -\infty$ and $\lim_{\|\alpha\|\to\infty} -\frac{1}{2}(\alpha-\mu_p)^t\Sigma_p^{-1}(\alpha-\mu_p) = -\infty$. Moreover, we have $\lim_{\rho\to 0}\log(\rho) = -\infty$, so we get $\lim_{\eta\to\partial(\Theta\times\varrho)}\log q(\boldsymbol{y},\eta) = -\infty$, which ensures that for all $M>0$ there exists $\ell>0$ such that $\|\alpha_t\|\geq\ell$ or $||\Gamma_t||+||\Gamma_t^{-1}||\geq\ell$ or $\rho_t\leq\frac{1}{\ell}$ implies $-l(\eta)\geq M$.

So $W(s_{k-1})\leq M$ implies that for all $1\leq t\leq\tau_m$ we have $\|\alpha_t\|\leq\ell$, $||\Gamma_t||+||\Gamma_t^{-1}||\leq\ell$ and $\frac{1}{\ell}\leq\rho_t\leq 1-\frac{1}{\ell}$ because $\sum_{t=1}^{\tau_m}\rho_t = 1$.

Let us denote by $\mathcal{V}_\ell = \Theta_\ell^{\tau_m}\times\left\{(\rho_t)_{1\leq t\leq\tau_m}\in\left[\frac{1}{\ell},1-\frac{1}{\ell}\right]^{\tau_m} \mid \sum_{t=1}^{\tau_m}\rho_t = 1\right\}$, where

$$\Theta_\ell = \left\{\theta=(\alpha,\Gamma_g)\mid\alpha\in\mathbb{R}^{k_p},\ \Gamma_g\in\mathrm{Sym}^+_{2k_g,*}(\mathbb{R})\mid\|\alpha\|\leq\ell,\frac{1}{\ell}\leq||\Gamma_g||\leq\ell\right\}.$$

So there exists a compact set $\mathcal{V}_\ell$ of $\Theta\times\varrho$ such that $W(s_{k-1})\leq M$ implies $\hat{\eta}(s_{k-1})\in\ \mathcal{V}_\ell$ and the first term (6.4) can be bounded as follows:

$$\sum_{\boldsymbol{\tau}}\left|\int_{\mathbb{R}^N} S(\boldsymbol{\beta},\boldsymbol{\tau})\left[\Pi^{J_k}_{\eta_{k-1},\boldsymbol{\tau}}(\boldsymbol{\beta}_0,\boldsymbol{\beta}) - q(\boldsymbol{\beta}|\boldsymbol{\tau},\boldsymbol{y},\eta_{k-1})\right]\mathrm{d}\boldsymbol{\beta}\right|\mathbb{1}_{W(s_{k-1})\leq M}$$
$$\leq\sum_{\boldsymbol{\tau}}\sup_{\eta\in\mathcal{V}_\ell}\left|\int_{\mathbb{R}^N} S(\boldsymbol{\beta},\boldsymbol{\tau})\left[\Pi^{J_k}_{\eta,\boldsymbol{\tau}}(\boldsymbol{\beta}_0,\boldsymbol{\beta}) - q(\boldsymbol{\beta}|\boldsymbol{\tau},\boldsymbol{y},\eta)\right]\mathrm{d}\boldsymbol{\beta}\right|.$$

Since for each $\boldsymbol{\tau}$ the function $\boldsymbol{\beta}\to S(\boldsymbol{\beta},\boldsymbol{\tau})$ belongs to $\mathcal{L}_V$, since we have proved that each transition kernel $\Pi_{\eta,\boldsymbol{\tau}}$ is geometrically ergodic and since the set $\mathcal{V}_\ell$ is compact, we can deduce that the first term (6.4) converges to zero as J_k tends to infinity.

We now consider the second term (6.5). We first need to prove that $\mathcal{M}_{\eta_k}\mathbb{1}_{W(s_{k-1})\leq M}$ is uniformly bounded that is to say the integral of the sufficient statistics are uniformly bounded on $\{W(s_{k-1})\leq M\}$; we only need to focus on the sufficient statistic which is not bounded itself: let $(j,m)\in\{1,...,2k_g\}^2$:

$$\begin{aligned}\int|\boldsymbol{\beta}^j\boldsymbol{\beta}^m|q(\boldsymbol{\beta}|\boldsymbol{\tau},\boldsymbol{y},\eta_{k-1})\mathrm{d}\boldsymbol{\beta}\mathbb{1}_{\eta_{k-1}\in\mathcal{V}_\ell} &\leq \int|\boldsymbol{\beta}^j\boldsymbol{\beta}^m|\frac{q(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{y},\eta_{k-1})}{q(\boldsymbol{\tau},\boldsymbol{y},\eta_{k-1})}\mathrm{d}\boldsymbol{\beta}\mathbb{1}_{\eta_{k-1}\in\mathcal{V}_\ell}\\ &\leq \frac{C(\mathcal{V}_\ell)}{q(\boldsymbol{\tau},\boldsymbol{y},\eta_{k-1})}\int|\boldsymbol{\beta}^j\boldsymbol{\beta}^m|\exp\left(-\frac{1}{2}\boldsymbol{\beta}^t\hat{\Gamma}^{-1}_{g,\boldsymbol{\tau},k-1}\boldsymbol{\beta}\right)\mathrm{d}\boldsymbol{\beta}\\ &\leq C(\mathcal{V}_\ell)\int Q(\boldsymbol{\beta},\hat{\Gamma}_{g,\boldsymbol{\tau},k-1})\exp\left(-\frac{1}{2}\|\boldsymbol{\beta}\|^2\right)\mathrm{d}\boldsymbol{\beta}<\infty,\end{aligned}$$

where $C(\mathcal{V}_\ell)$ is a constant depending only on the set $\mathcal{V}_\ell$, $\hat{\Gamma}_{g,\boldsymbol{\tau}}$ is the diagonal block matrix with all the Γ_{g,τ_i} given by the label vector $\boldsymbol{\tau}$ and we have changed the variable in the last inequality and Q is a quadratic form in $\boldsymbol{\beta}$ whose coefficients are continuous functions of elements of the matrix Γ_g. So we obtain that for all $M>0$ there exists $\ell>0$ such that for all integer k we have: $\mathcal{M}_{\eta_k}\mathbb{1}_{W(s_{k-1})\leq M}\leq C(\mathcal{V}_\ell)$.

We now prove the convergence to 0 of the second term of the product involved in (6.5). Let us denote by $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k}$ for the term $|R_{J_k}(\boldsymbol{\tau}|\boldsymbol{y},\eta_{k-1}) - q(\boldsymbol{\tau}|\boldsymbol{y},\eta_{k-1})|$. Thus we have:

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k} &= \left|\prod_{i=1}^{n}\int p_{J_k}(\tau_i|\xi_i,y_i,\eta_{k-1})Q(\xi_i)\mathrm{d}\xi_i - \prod_{i=1}^{n} q(\tau_i|y_i,\eta_{k-1})\right| \\
&\leq \sum_{i=1}^{n}\left|\int p_{J_k}(\tau_i|\xi_i,y_i,\eta_{k-1})Q(\xi_i)\mathrm{d}\xi_i - q(\tau_i|y_i,\eta_{k-1})\right| \\
&\leq \sum_{i=1}^{n}\int \left|p_{J_k}(\tau_i|\xi_i,y_i,\eta_{k-1}) - q(\tau_i|y_i,\eta_{k-1})\right| Q(\xi_i)\mathrm{d}\xi_i \\
&\leq \sum_{i=1}^{n}\int \left|\frac{S_{J_k}(\tau_i,y_i|\xi_{\tau_i,i},\eta_{k-1})}{\sum_s S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})} - \frac{q(\tau_i,y_i|\eta_{k-1})}{q(y_i|\eta_{k-1})}\right| Q(\xi_i)\mathrm{d}\xi_i,
\end{aligned}$$

where we denote by $S_J(t,y_i|\xi_{t,i},\eta)$ the quantity $\left(\frac{1}{J}\sum_{l=1}^{J}\left[\frac{f(\xi_{t,i}^{(l)})}{q(y_i,\xi_{t,i}^{(l)},t|\eta)}\right]\right)^{-1}$.

We write each term of this sum as follows:

$$\begin{aligned}
&\frac{S_{J_k}(\tau_i,y_i|\xi_{\tau_i,i},\eta_{k-1})}{\sum_{s=1}^{\tau_m} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})} - \frac{q(\tau_i,y_i|\eta_{k-1})}{q(y_i|\eta_{k-1})} = \\
&\qquad \frac{S_{J_k}(\tau_i,y_i|\xi_{\tau_i,i},\eta_{k-1})(q(y_i|\eta_{k-1}) - \sum_{s=1}^{\tau_m} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1}))}{q(y_i|\eta_{k-1})\sum_{s=1}^{\tau_m} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})} \\
&\qquad\qquad + \frac{(S_{J_k}(\tau_i,y_i|\xi_{\tau_i,i},\eta_{k-1}) - q(\tau_i,y_i|\eta_{k-1}))\sum_{s=1}^{\tau_m} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})}{q(y_i|\eta_{k-1})\sum_{s=1}^{\tau_m} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})}.
\end{aligned}$$

Denoting by $\mathcal{T}_i$ the set of τ_m+1 integers $\{1,\ldots,\tau_m\}\cup\{\tau_i\}$, we obtain finally:

$$\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k} \leq \sum_{i=1}^{n}\frac{1}{q(y_i|\eta_{k-1})}\sum_{s\in\mathcal{T}_i}\int |S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1}) - q(s,y_i|\eta_{k-1})|\, Q(\xi_i)\mathrm{d}\xi_i.$$

Defining the event $A_{k,i,t} = \{|S_{J_k}(t,y_i|\xi_{t,i},\eta_{k-1}) - q(t,y_i|\eta_{k-1})| > \zeta_k\}$ for some positive sequence $(\zeta_k)_k$, we get:

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k} \leq\ & \sum_{i=1}^{n}\frac{1}{q(y_i|\eta_{k-1})}\sum_{s\in\mathcal{T}_i}\int_{A_{k,i,s}} |S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1}) - q(s,y_i|\eta_{k-1})|\, Q(\xi_i)\mathrm{d}\xi_i \\
&+ \sum_{i=1}^{n}\frac{1}{q(y_i|\eta_{k-1})}\sum_{s\in\mathcal{T}_i}\int_{A_{k,i,s}^c} |S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1}) - q(s,y_i|\eta_{k-1})|\, Q(\xi_i)\mathrm{d}\xi_i.
\end{aligned}$$

So we deduced that:

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k} &\leq \sum_{i=1}^{n} \frac{1}{q(y_i|\eta_{k-1})} \sum_{s\in\mathcal{T}_i} (\sup_{\xi} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1}) + q(s,y_i|\eta_{k-1}))P(A_{k,i,s}) \\
&\quad + \left(\sum_{i=1}^{n} \frac{1}{q(y_i|\eta_{k-1})}\right)(\tau_m+1)\zeta_k \\
&\leq \sum_{i=1}^{n} \left(\frac{\sup_{\xi,s} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})}{q(y_i|\eta_{k-1})} + 1\right)\left(\sum_{s\in\mathcal{T}_i} P(A_{k,i,s}) + P(A_{k,i,\tau_i})\right) \\
&\quad + \left(\sum_{i=1}^{n} \frac{1}{q(y_i|\eta_{k-1})}\right)(\tau_m+1)\zeta_k.
\end{aligned}$$

Assuming $\zeta_k < \min_{i,t} q(t,y_i|\eta_{k-1})$, we obtain:

$$\begin{aligned}
P(A^c_{k,i,t}) &= P(|S_{J_k}(t,y_i|\xi_{t,i},\eta_{k-1}) - q(t,y_i|\eta_{k-1})| \leq \zeta_k) \\
&\geq P\left(\left|\frac{1}{S_{J_k}(t,y_i|\xi_{t,i},\eta_{k-1})} - \frac{1}{q(t,y_i|\eta_{k-1})}\right| \leq \frac{\zeta_k}{q(t,y_i|\eta_{k-1})(q(t,y_i|\eta_{k-1})+\zeta_k)}\right) \\
&\geq P\left(\left|\frac{1}{S_{J_k}(t,y_i|\xi_{t,i},\eta_{k-1})} - \frac{1}{q(t,y_i|\eta_{k-1})}\right| \leq \frac{\zeta_k}{2q(t,y_i|\eta_{k-1})^2}\right).
\end{aligned}$$

Using the first inequality of Theorem 2 of [9], we get: $P(A_{k,i,t}) \leq c_1 \exp\left(-c_2 \frac{J_k\zeta_k^2}{q(t,y_i|\eta_{k-1})^4}\right)$, where c_1 and c_2 are independent of k since (η_k) only moves in a compact set $\mathcal{V}_\ell$ thanks to the condition $\mathbb{1}_{W(s_{k-1}\leq M)}$. This yields:

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{y},k} \leq c_1 \sum_{i=1}^{n} \left(\frac{\sup_{\xi,s} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})}{q(y_i|\eta_{k-1})} + 1\right)(\tau_m+1)\exp\left(-c_2\frac{J_k\zeta_k^2}{\max_i q(y_i|\eta_{k-1})^4}\right) \\
+ \sup_{\eta_{k-1}\in\mathcal{L}_m}\left(\sum_{i=1}^{n}\frac{1}{q(y_i|\eta_{k-1})}\right)(\tau_m+1)\zeta_k.
\end{aligned}$$

We have to prove that the Monte Carlo sum involved in $S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})$ does not equal zero everywhere, so that $\sup_{\xi,s} S_{J_k}(s,y_i|\xi_{s,i},\eta_{k-1})$ is finite. For this purpose, we can choose a particular probability density function f. Indeed, if we set f to be the prior density function on the simulated deformation fields ξ, we have for all $\eta \in \mathcal{V}_\ell$:

$$\begin{aligned}
\frac{1}{J}\sum_{l=1}^{J}\left[\frac{f(\xi_{t,i}^{(l)})}{q(y_i,\xi_{t,i}^{(l)},t|\eta)}\right] &= \frac{1}{J}\sum_{l=1}^{J}\left[\frac{1}{q(y_i|\xi_{t,i}^{(l)},t,\eta)q(t|\eta)}\right] \\
&\geq \frac{1}{J}\sum_{l=1}^{J}\left[\frac{1}{\frac{1}{(2\pi\sigma_t^2)^{|\Lambda|}}\exp\left(-\frac{1}{2\sigma_t^2}\|y_i - K_p^{\xi^{(l)}}\alpha_t\|^2\right)}\right] \geq (2\pi\sigma^2)^{|\Lambda|},
\end{aligned}$$

where σ is the lower bound of the variances (σ_t).

We choose the sequence $(\zeta_k)_k$ depending upon $(J_k)_k$ such that $\lim_{k\to\infty}\zeta_k = 0$ and $\lim_{k\to\infty} J_k\zeta_k^2 = +\infty$. We can take for example $\zeta_k = J_k^{-1/3}$ for all $k \geq 1$.

We will now prove the convergence of the sequence of excitation terms.

For any $M > 0$ we define $M_n = \sum_{k=1}^{n} \Delta_k e_k \mathbb{1}_{W(s_{k-1}) \leq M}$ and let $\mathcal{F} = (\mathcal{F}_k)_{k \geq 1}$ be the filtration, where $\mathcal{F}_k$ is the σ-algebra generated by the random variables $(S_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_k)$. We have $M_n = \sum_{k=1}^{n} \Delta_k \left(S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - \mathbb{E}\left[S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) | \mathcal{F}_{k-1}\right]\right) \mathbb{1}_{W(s_{k-1}) \leq M}$ so this shows us that (M_n) is a $\mathcal{F}$-martingale. In addition to this we have:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}\left[\|M_k - M_{k-1}\|^2 \mid \mathcal{F}_{k-1}\right] &= \sum_{k=1}^{\infty} \mathbb{E}\left[\Delta_k^2 \|e_k\|^2 \mathbb{1}_{W(s_{k-1}) \leq M} \mid \mathcal{F}_{k-1}\right] \leq \sum_{k=1}^{\infty} \Delta_k^2 \mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_{k-1}\right] \\
&\leq \sum_{k=1}^{\infty} \Delta_k^2 \mathbb{E}\left[\|S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) - \mathbb{E}\left[S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k) | \mathcal{F}_{k-1}\right]\|^2 \mid \mathcal{F}_{k-1}\right] \\
&\leq \sum_{k=1}^{\infty} \Delta_k^2 \mathbb{E}\left[\|S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k)\|^2 \mid \mathcal{F}_{k-1}\right] .
\end{aligned}$$

We now evaluate this last integral term:

$$\begin{aligned}
\mathbb{E}\left[\|S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k)\|^2 \mid \mathcal{F}_{k-1}\right] &= \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} \int \|S(\boldsymbol{\beta}, \boldsymbol{\tau})\|^2 \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \prod_{i=1}^{n} p_{J_k, \eta_{k-1}}(\tau_i, \xi_{\tau_i, i}, y_i) Q(\xi_{\tau_i, i}) \mathrm{d}\xi_{\tau_i, i} \mathrm{d}\boldsymbol{\beta} \\
&\leq \left[\sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} \|S(\boldsymbol{\beta}, \boldsymbol{\tau})\|^2 \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}\right] \left[\int \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\xi_0, \xi) \mathrm{d}\xi\right] .
\end{aligned}$$

The last term equals one and again we only need to focus on the sufficient statistic which is not bounded itself. Indeed $S_{3,t}(\boldsymbol{\beta}, \boldsymbol{\tau})$ for all $1 \leq t \leq \tau_m$ so using the fact that the function V dominates this sufficient statistic, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|S_{3,t}(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k)\|^2 \mid \mathcal{F}_{k-1}\right] &\leq \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} \|S_{3,t}(\boldsymbol{\beta}, \boldsymbol{\tau})\|^2 \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \\
&\leq C \sum_{\boldsymbol{\tau}} \int_{\mathbb{R}^N} V(\boldsymbol{\beta})^2 \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \leq C \sum_{\boldsymbol{\tau}} \Pi^{J_k}_{\eta_{k-1}, \boldsymbol{\tau}} V(\boldsymbol{\beta}_0)^2 .
\end{aligned}$$

Applying Lemma 6.3 for $p = 2$, we get:

$$\mathbb{E}\left[\|S(\boldsymbol{\beta}_k, \boldsymbol{\tau}_k)\|^2 \mid \mathcal{F}_{k-1}\right] \leq C \sum_{\boldsymbol{\tau}} (\rho V(\boldsymbol{\beta}_0)^2 + C) \leq C \tau_m^n (\rho V(\boldsymbol{\beta}_0)^2 + C).$$

Finally it remains: $\sum_{k=1}^{\infty} \mathbb{E}\left[\|M_k - M_{k-1}\|^2 \mid \mathcal{F}_{k-1}\right] \leq C \sum_{k=1}^{\infty} \Delta_k^2$, which ensures that the previous series converges. This involves that $(M_k)_{k \in \mathbb{N}}$ is a martingale bounded in L^2 so that $\lim_{k \to \infty} M_k$ exists (see [13]). This proves the first part of (STAB2).

To conclude this proof we apply Theorem 4.1 and get that $\lim_{k \to \infty} d(s_k, \mathcal{L}') = 0$.

7. Conclusion and discussion

We consider the setting of Bayesian non-rigid deformable models building in the context of [1] and the associated MAP estimator. We approximate this estimator of this generative model parameters thanks to a stochastic algorithm which derives from an EM algorithm. We also prove its theoretical convergence toward a critical point of the observed likelihood. This is, to our best knowledge, the first convergent estimation algorithm of the templates and geometrical variabilities in the framework of mixture model for deformable templates.

The algorithm is based on a stochastic approximation of the EM algorithm using a MCMC approximation of the posterior distribution and truncation on random boundaries. We present experiments on the US-postal database as well as on some 2D medical data. This shows that the stochastic approach can be easily implemented with the algorithm detailed here and is robust to noisy situations, giving better result than the previous deterministic schemes.

Many interesting questions remain open.

The first goal of these model and algorithm is the estimation of some atlases in a given population and the acceptable deformations of these atlases that can explain the variability in the population. However, this model, as soon as the parameters are estimated, can be used to create a classifier. Given a new image, one can compute the most likely component that this image belongs to. This computation requires to evaluate the integral of the complete likelihood with respect to the posterior distribution as well as in the estimation process. A first proposition to overcome this difficulty has been given in [1] while approximating the posterior distribution by a Dirac on its mode. This gave very interesting results which are presented in that paper. However, in the case of noisy images, the same problem occurs and leads to bad classification ratios. Another way has been proposed in [2] using the same methods as in this paper, that is to say, using Monte Carlo Markov Chain methods. The results are impressive and the improvement noticeable.

We have presented here a set of experiments on 2D images. The generative model as well as the algorithm and the proof of its convergence do not depend on the dimension of the images. The implementation for 3D images is only a numerical issue. We are currently working on the 3D codes to test this algorithm on real medical databases.

An interesting extension would be to consider diffeomorphic mapping and not only displacement fields for the hidden deformation. This appears to be particularly interesting in the context of Computational Anatomy where a one to one correspondence between the template and the observation is usually needed and cannot be guaranteed with linear spline interpolation schemes. This extension could be done in principle using tangent models based on geodesic shooting in the spirit of [20]. Many numerical as well as theoretical work need to be done in this area.

References

[1] S. Allassonnière, Y. Amit and A. Trouvé, Toward a coherent statistical framework for dense deformable template estimation. *J. Roy. Stat. Soc.* **69** (2007) 3–29.

[2] S. Allassonnière, E. Kuhn and A. Trouvé, Map estimation of statistical deformable templates *via* nonlinear mixed effects models: Deterministic and stochastic approaches. In *Proc. Int. Workshop on the Mathematical Foundations of Computational Anatomy (MFCA-2008)*, edited by X. Pennec and S. Joshi (2008).

[3] S. Allassonnière, E. Kuhn and A. Trouvé, Construction of Bayesian deformable models *via* a stochastic approximation algorithm: A convergence study. *Bernoulli* **16** (2010) 641–678.

[4] Y. Amit, U. Grenander and M. Piccioni, Structural image restoration through deformable templates. *J. Am. Statist. Assoc.* **86** (1989) 376–387.

[5] C. Andrieu, R. Moulines and P. Priouret, Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** (2005) 283–312 (electronic).

[6] T.F. Cootes, G.J. Edwards and C.J. Taylor, Actives appearance models. In *5th Eur. Conf. on Computer Vision, Berlin*, Vol. 2, edited by H. Burkhards and B. Neumann. Springer (1998) 484–498.

[7] B. Delyon, M. Lavielle and E. Moulines, Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** (1999) 94–128.

[8] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data *via* the EM algorithm. *J. Roy. Statist. Soc.* **1** (1977) 1–22.

[9] C. Dorea and L. Zhao, Nonparametric density estimation in hidden Markov models. *Statist. Inf. Stoch. Process.* **5** (2002) 55–64.

[10] C.A. Glasbey and K.V. Mardia, A penalised likelihood approach to image warping. *J. Roy. Statist. Soc., Ser. B* **63** (2001) 465–492.

[11] J. Glaunès and S. Joshi, Template estimation form unlabeled point set data and surfaces for computational anatomy. In *Proc. Int. Workshop on the Mathematical Foundations of Computational Anatomy (MFCA-2006)*, edited by X. Pennec and S. Joshi (2006) 29–39.

[12] U. Grenander, *General Pattern Theory*. Oxford Science Publications (1993).

[13] P. Hall and C.C. Heyde, Martingale limit theory and its application. *Probab. Math. Statist.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York (1980).
[14] E. Kuhn and M. Lavielle, Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: PS* **8** (2004) 115–131 (electronic).
[15] H.J. Kushner and D.S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*, volume 26 of *Appl. Math. Sci.* Springer-Verlag, New York (1978).
[16] S. Marsland, C. Twining and C. Taylor, A minimum description length objective function for groupwise non rigid image registration. *Image and Vision Computing* (2007).
[17] S.P. Meyn and R.L. Tweedie, *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag, London Ltd. (1993).
[18] M.I. Miller, T.A. and L. Younes, On the metrics and Euler-Lagrange equations of computational anatomy. *Ann. Rev. Biomed. Eng.* **4** (2002) 375–405.
[19] C. Robert, *Méthodes de Monte Carlo par chaînes de Markov*. Statistique Mathématique et Probabilité. [Mathematical Statistics and Probability]. Éditions Économica, Paris (1996).
[20] M. Vaillant, I. Miller, M.A. Trouvé and L. Younes, Statistics on diffeomorphisms *via* tangent space representations. *Neuroimage* **23** (2004) S161–S169.