# Stochastic Algorithm For Parameter Estimation For Dense Deformable Template Mixture Model

Stéphanie Allassonniere, Estelle Kuhn

# Stochastic Algorithm For Parameter Estimation For Dense Deformable Template Mixture Model

S. Allassonnière[1], E. Kuhn[2]

February 11, 2008

CIS - Johns Hopkins University[1]
3400 North Charles Street
MD, 21218 Baltimore, USA

LAGA- Université Paris 13[2]
99 Av. J-B Clément
93430 Villetaneuse, France

**Abstract**

Estimating probabilistic deformable template models is a new approach in the fields of computer vision and probabilistic atlases in computational anatomy. A first coherent statistical framework modelling the variability as a hidden random variable has been given by Allassonnière, Amit and Trouvé in [1] in simple and mixture of deformable template models. A consistent stochastic algorithm has been introduced in [2] to face the problem encountered in [1] 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 model. We also prove the convergence of this algorithm toward a critical point of the penalised likelihood of the observations and illustrate this with handwritten digit images.

**keywords** stochastic approximations, non rigid-deformable templates, shapes statistics, Bayesian modelling, MAP estimation.

# 1 Introduction

This paper deals with the representation and the analysis of geometrical structures upon which some deformations can act. 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" [10]. 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.

Many solutions have been proposed to face the problem of the template definition. They go from some generalised Procruste's means with a variational [9] or statistical [8] point of view to some statistical models like Active Appearance Model [4] or Minimum Description Length methods [13]. Unfortunately, all these methods are only focussing 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. Another issue addressed here is the clustering problem. Given a set of images, the statistical estimation of the component weights and 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 the K-means) but the geometry, and related to this the metric 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. Finally, all these iterative algorithms 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, Amit and Trouvé 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 and random deformation of a random template image picked 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 template, the parameters of the deformation laws and the components weight are the parameters of interest. This generative model automatically decomposes 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 penalised Maximum A Posteriori (MAP). This estimation problem is carried out 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 [2]. We propose to use a stochastic version of the EM algorithm to reach the posterior distribution of the hidden variables. We use the SAEM algorithm introduced by Delyon et al in [5] coupled with a Monte Carlo Markov Chain (MCMC) method. Contrary to the one component model where we can couple the iteration of the SAEM algorithm with the Markov chain evolution (introduced by Kuhn and Lavielle in

[12] and extended in [2]), we show here that it cannot be driven numerically. We need to consider another method. We propose to simulate the hidden variables using some subsidiary Markov chains, one per component, to approach the posterior distributions of the labels in particular. We prove the convergence of this particular algorithm for a non compact setting by adapting Delyon's theorem about general stochastic approximations and introducing truncation on random boundaries as in [3].

The paper is organised as follows: in Section 2 we first recall the observation mixture model proposed by Allassonnière, Amit and Trouvé in [1]. In Section 3, we describe the stochastic algorithm used in our particular setting. Section 4 is devoted to the experiments. Section 5 gathers the proof of the convergence of the algorithm.

# 2 The Observation Model

We are working with the 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 $\{r_s \in D \subset \mathbb{R}^2, s \in \Lambda\}$ where $D$ is a continuous domain and $\Lambda$ the pixel network, we are looking for some template images which will explain the panel. Each of them 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 $\Lambda$ of the deformation of a template 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 $s \in \Lambda$ : $y(s) = I_0(r_s - z(r_s)) + \sigma\epsilon(s)$ , where $\sigma\epsilon$ denotes an additive noise.

## 2.1 Models for Templates and Deformations

We use the same framework as chosen in [1] to describe both the template $I_0$ and the deformation field $z$. Our model takes into account two complementary sides: photometry -indexed by $p$, and geometry -indexed by $g$. The template $I_0$ and the deformation $z$ are assumed to belong to some finite dimensional subspace of two reproducing kernels

Hilbert spaces $V_p$ and $V_g$ (determined by their respective kernel $K_p$ 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 $D$: $(r_{p,k})_{1\leq k\leq k_p}$ respectively $(r_{g,k})_{1\leq k\leq k_g}$. They are therefore parametrised by the coefficients $\alpha \in \mathbb{R}^{k_p}$ and $\beta \in (\mathbb{R}^{k_g})^2$ which yield: $\forall r \in D$,

$$I_\alpha(r) = (\mathbf{K_p}\alpha)(r) = \sum_{k=1}^{k_p} K_p(r, r_{p,k})\alpha(k) \text{ and } z_\beta(r) = (\mathbf{K_g}\beta)(r) = \sum_{k=1}^{k_g} K_g(r, r_{g,k})\beta(k).$$

## 2.2 Parametrical Model

In this paper, we consider a mixture of the deformable template model which enables a fixed number $\tau_m$ of components in each training set. This means that the data will be separated in $\tau_m$ (at most) different components by the algorithm. The algorithm automatically decompose the data increasing the posterior likelihood while assigning a label to each image of the training set. Therefore, for each observation $y_i$, we consider the pair $(\beta_i, \tau_i)$ of unobserved variables which correspond respectively to the deformation field and the label of image $i$. We denote below $y_1^n \triangleq (y_i)_{1\leq i\leq n}$, $\beta_1^n \triangleq (\beta_i)_{1\leq i\leq n}$ and $\tau_1^n \triangleq (\tau_i)_{1\leq i\leq n}$. The generative model is:

$$\left\{ \begin{array}{l} \tau_1^n \sim \otimes_{i=1}^n \sum_{t=1}^{\tau_m} \rho_t \delta_t \mid (\rho_t)_{1\leq t\leq \tau_m}\,, \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_{g,\tau_i}) \mid \ \tau_1^n,\ (\Gamma_{g,t})_{1\leq t\leq \tau_m}\,, \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_{\alpha_{\tau_i}}, \sigma^2_{\tau_i} Id_{|\Lambda|}) \mid \beta_1^n,\ \tau_1^n,\ (\alpha_t, \sigma_t^2)_{1\leq t\leq \tau_m}\,, \end{array} \right. \tag{1}$$

where $z_\beta I_\alpha(s) = I_\alpha(r_s - z_\beta(r_s))$, for $s$ in $\Lambda$ and $\delta_t$ 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 noise, 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 the parameters $(\theta_t, \rho_t)_{1\leq t\leq \tau_m}$ so that $\theta_t$ corresponds to the parameters composed of the photometric part $(\alpha_t, \sigma_t^2)$ and the geometric part $\Gamma_{g,t}$

for all $1 \leq t \leq \tau_m$. 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 $\Theta$ defined as $\Theta = \{\ \theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p},\ |\alpha| < R,\ \sigma > 0,\ \Gamma_g \in \mathrm{Sym}^+_{2k_g,*}(\mathbb{R})\ \}$, 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 setx $\varrho = \{(\rho_t)_{1\leq t\leq \tau_m} \in ]0,1[^{\tau_m} \mid \sum_{t=1}^{\tau_m} \rho_t = 1\}$ .

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

## 2.3 The Bayesian Model

Even though the parameters are finite dimensional, their high dimensionality can lead to degenerated maximum-likelihood estimator when the training sample is small. Introducing prior distributions, estimation with small samples is still possible and their importance has been shown in the parameter update formula in [1]. We use a generative model which includes standard conjugate prior distributions with *fixed* parameters: a normal prior on $\alpha_t$ and inverse-Wishart priors on $\sigma_t^2$ and $\Gamma_{g,t}$ for all $1 \leq t \leq \tau_m$. All priors are assumed independent. Let $\forall 1 \leq t \leq \tau_m$, $\theta_t = (\alpha_t, \sigma_t^2, \Gamma_{g,t}) \sim \nu_p \otimes \nu_g$ and $\langle A, B\rangle_F \triangleq tr(A^t B)$ for two matrices $A$ and $B$ where

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

For the prior law $\nu_\rho$ we choose the Dirichlet distribution with density

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

# 3 Parameter Estimation with a Stochastic Approximation EM Algorithm

For the sake of simplicity, let us denote in the sequel $x \triangleq \beta_1^n \in \mathbb{R}^N$ with $N \triangleq 2nk_g$ the vector collecting all the missing deformation variables and $\lambda \triangleq \tau_1^n \in \mathcal{T}$ with $\mathcal{T} \triangleq \{1, \ldots, \tau_m\}^n$ the collection of missing labels. 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 estimation the parameters:

$$\hat{\eta}_n = \underset{\eta}{\operatorname{argmax}} \, q(\eta|y) \, , \tag{3}$$

where $q(\eta|y)$ denotes the a posteriori likelihood. (We will use $q$ to denote all the density functions below.)

To reach this estimator, we maximise the posterior likelihood using a Stochastic Approximation EM algorithm coupled with a MCMC method. Indeed, due to the intractable computation of the E step encountered in this complex non linear setting, we follow in a stochastic way the EM setting introduced by [6]. Unfortunately, the direct generalisation of the algorithm presented in [2] turns out to be of no use in practice. This suggests to go back to some extension of the SAEM procedure proposed in [5].

## 3.1 The SAEM Algorithm Using MCMC Methods

Let us first recall the SAEM algorithm. The $k^{th}$ iteration consists in three steps:

(i) the missing data, here the deformation parameters and the labels, $(x, \lambda) = (\beta_1^n, \tau_1^n)$, are drawn using the current parameter according to the posterior distribution denoted $\pi_\eta$,

- Simulation step : $(x_k, \lambda_k) \sim \pi_{\eta_{k-1}}$,

(ii) a stochastic approximation is done on the complete likelihood using the simulated value of the missing data,

- Stochastic approximation : let $(\Delta_k)_k$ be a decreasing sequence of positive step-sizes:

$$Q_k(\eta) = Q_{k-1}(\eta) + \Delta_k[\log q(y, x_k, \lambda_k, \eta) - Q_{k-1}(\eta)]\,, \tag{4}$$

(iii) the parameters are updated in the M-step,

- Maximisation step : $\eta_k = \underset{\eta}{\operatorname{argmax}}\, Q_k(\eta)$.

Initialised values of $Q$ and $\eta$ 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. The complete likelihood can be written as: $q(y, x, \lambda, \eta) = \exp\left[-\psi(\eta) + \langle S(x, \lambda), \phi(\eta)\rangle\right]$ , 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 $\psi$, $\phi$ two Borel functions on $\Theta \times \varrho$. (Note that $S$, $\phi$ and $\psi$ may depend also on $y$, but since $y$ will stay fixed in the sequel, we omit this dependency.) With such a likelihood, the stochastic approximation can be done on the complete log-likelihood as well as on sufficient statistics. This yields to the following stochastic approximation:

$$s_k = s_{k-1} + \Delta_k(S(x_k, \lambda_k) - s_{k-1})\ .$$

We 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 make

$\nabla L$ vanishe. 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 becomes: $\eta_k = \hat{\eta}(s_k)$.

Unfortunately, the first step is, in this particular model, untractable and requires the use of some MCMC methods to reach the simulation of the missing data. We will explain this procedure in the two next paragraphs.

Another tool needs to be introduced. Some of the convergence assumptions of such algorithms [5, 12] will not be satisfied since we are working with unbounded missing data (the deformation fields $\beta$ are assumed Gaussian). This leads to consider a truncation algorithm as suggested in [5] and extended in [2].

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 $\Pi_\eta$ be a transition kernel of an ergodic Markov chain on $\mathbb{R}^N$ having $\pi_\eta$ as stationary distribution. We construct the sequence $((x_k, \lambda_k, s_k, \kappa_k))_{k\geq 0}$ as explained in Algorithm 1. As long as the stochastic approximation does not wander out the current compact set and is not too far from its previous value, we run our "SAEM like" algorithm. As soon as one of the two previous conditions is not satisfied, we reinitialise the sequences of $s$ and $(x, \lambda)$ using a projection (for more details see [5]).

**Algorithm 1** Stochastic approximation with truncation on random boundaries

Set $\kappa_0 = 0$, $s_0 \in \mathcal{K}_0$, $x_0 \in \text{K}$ and $\lambda_0 \in \mathcal{T}$.
**for all** $k \geq 1$ **do**
  `compute` $\bar{s} = s_{k-1} + \Delta_k(S(\bar{x}, \bar{\lambda}) - s_{k-1})$
  `where` $(\bar{x}, \bar{\lambda})$ `are sampled from a transition kernel` $\Pi_{\eta_{k-1}}$.
  **if** $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ **then**
    `set` $(s_k, x_k, \lambda_k) = (\bar{s}, \bar{x}, \bar{\lambda})$ `and` $\kappa_k = \kappa_{k-1}$,
  **else**
    `set` $(s_k, x_k, \lambda_k) = (\tilde{s}, \tilde{x}, \bar{\lambda}) \in \mathcal{K}_0 \times \text{K} \times \mathcal{T}$ `and` $\kappa_k = \kappa_{k-1} + 1$,
    `where` $(\tilde{s}, \tilde{x})$ `can be chosen through different ways (cf [5]).`
  **end if**
  $\eta_k = \underset{\eta}{\text{argmax}}\, \hat{\eta}(s_k)$.
**end for**

### 3.2 The intuitive idea: the usual Gibbs Sampler

In this particular setting, we can try to simulate the unobserved variables $(x, \lambda)$ using a Markov chain which has $q(x, \lambda|y, \eta)$ as stationary distribution and couple the iteration of the SAEM algorithm with the Markov chain evolution as done in [2]. If we consider the full vector $(x, \lambda)$ as a single vector of missing data, we can try and use the hybrid Gibbs Sampler on $\mathbb{R}^{N+n}$ as detailed in Algorithm 2. For any $b \in \mathbb{R}$ and $1 \leq j \leq N$, let us denote $x_{j,b}$ the unique configuration which is equal to $x$ everywhere except the coordinate $j$ where $x^j_{j,b} = b$ and $x_{-j}$ the vector $x$ without the coordinate $j$. Each coordinate of the deformation field $x^j$ is updated using a Hastings Metropolis procedure where the proposal is given by the conditional distribution of $x^j|x_{-j}, \lambda$ coming from the current Gaussian distribution with the corresponding parameters (pointed by $\lambda$).

**Algorithm 2** Transition step $k \to k+1$ using an hybrid Gibbs Sampler on $(x, \lambda)$

**Require:** $x = x_k$, $\lambda = \lambda_k$; $\eta = \eta_k$

`Gibbs Sampler` $\Pi_{\eta,\lambda}$`:`

**for all** $j = 1 : N$ **do**

`Hasting-Metropolis procedure:`

$b \sim q(b|x_{-j}, \lambda, \eta)$

Compute $r_{\eta,\lambda,j}(x, b) = \left[\frac{q(y|x_{j,b},\lambda,\eta)}{q(y|x,\lambda,\eta)} \wedge 1\right]$

With probability $r_{\eta,\lambda,j}(x, b)$, update $x^j$: $x^j \leftarrow b$

**end for**

Update $x_{k+1} \leftarrow x$

Update $\lambda$ through the following distribution:

$$\lambda_{k+1} \sim \otimes_{i=1}^{n} \sum_{t=1}^{\tau_m} q(t|y_i, \beta_{i,k+1}, \eta)\delta_t$$

where $\delta$ is the Dirac function and the weights $(q(t|y_i, \beta_{i,k+1}, \eta))_{t,i}$ are proportional to $(q(y_i, \beta_{i,k+1}, t|\eta))_{t,i}$ and their sum equals to one.

Even if this procedure provided an estimated parameter sequence which theoretically converged toward the MAP estimator, in practice, as mentioned in [15], 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 one 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, thanks to its label, it simulates a deformation field which only takes into account the parameters of this given component. Indeed, the simulation of $x$ through the Gibbs Sampler involves a proposal whose corresponding Markov chain has $q(x|\lambda, y, \eta)$ as stationary distribution. Therefore, the deformation tries to match $y$ to the deformed template of the given components $\lambda$. 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.

This suggests that this algorithm should not be used in our case. To overcome the trapping state problem, we will simulate the optimal label, using as many Markov chains in $x$ 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 label.

**Remark 1** *This is a point that was done in [1] while computing the best matching for all components by minimising the corresponding energies.*

## 3.3 Using multicomponent Markov chains

Since we aim to simulate $(x, \lambda)$ through a transition kernel that has $q(x, \lambda|y, \eta)$ as stationary distribution, we simulate $\lambda$ with a kernel whose stationary distribution is $q(\lambda|y, \eta)$ and then $x$ through a transition kernel that has $q(x|\lambda, y, \eta)$ 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. So we will make an approximation. Indeed, for any density function $f$, for any image $y_i$ and for all $1 \leq t \leq \tau_m$, we have

$$q(t, y_i|\eta) = \left( \mathbb{E}_{q(\beta|y_i,t,\eta)} \left[ \frac{f(\beta)}{q(y_i, \beta, t|\eta)} \right] \right)^{-1} . \tag{5}$$

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

The solution we propose is the following: suppose we are at the $k^{th}$ iteration of the algorithm and let $\eta$ 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 detailed in Algorithm 3 whose stationary distribution is $q(\cdot|y_i, t, \eta)$. Let us denote $\xi_i = (\xi_{t,i})_{1 \leq t \leq \tau_m}$ 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:

$$p_J(t|\xi_i, y_i, \eta) \propto \left( \frac{1}{J} \sum_{l=1}^{J} \left[ \frac{f(\xi_{t,i}^{(l)})}{q(y_i, \xi_{t,i}^{(l)}, t|\eta)} \right] \right)^{-1} , \tag{6}$$

where the normalisation is done such that their sum over $t$ equals one, involving the dependence on all the auxiliary variables $\xi_i$. The ergodic theorem ensures the convergence of our approximation toward the expected value. We then simulate $\lambda$ 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 $x$ by re-running $J$ times the hybrid Gibbs

sampler $\Pi_{\eta,\lambda}$ on $\mathbb{R}^N$ starting from a random initial point $x_0$ 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 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 on: $f_t(\xi) = \frac{1}{\sqrt{2\pi}^{2k_g}\sqrt{|\Gamma_{g,t}|}} \exp\left(-\frac{1}{2}\xi^T\Gamma_{g,t}^{-1}\xi\right)$ . Algorithm 3 shows the detailed iteration.

**Algorithm 3** Transition step $k \to k+1$ using an hybrid Gibbs Sampler on $(x, \lambda)$

**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**
          `Hasting-Metropolis procedure:`
          $b \sim q(b|\xi_{-j}, t, \eta)$
          Compute $r_{\eta,t,j}(\xi, b) = \left[\frac{q(y_i|\xi_{j,b},t,\eta)}{q(y_i|\xi,t,\eta)} \wedge 1\right]$
          With probability $r_{\eta,t,j}(\xi, b)$, update $\xi^j$: $\xi^j \leftarrow b$
        **end for**
        $\xi_{t,i}^{(l)} = \xi$
      **end for**

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

    **end for**
  **end for**

$$\lambda_{k+1} \sim \otimes_{i=1}^n \sum_{t=1}^{\tau_m} p_{J_k}(t|\xi_i, y_i, \eta)\delta_t \quad \text{and} \quad x_{k+1} \sim \Pi_{\eta,\lambda_{k+1}}^{J_k}(x_0).$$

## 3.4 Convergence theorem of the multicomponent procedure

In this particular case, we are not working with the SAEM-MCMC algorithm which couples the iteration of the Markov Chain to the EM iterations. To prove the convergence of our parameter estimate toward the MAP, we need a convergence theorem which deals with general stochastic approximations.

We consider the following Robbins Monroe stochastic approximation procedure:

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

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 the stochastic excitation.

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

$$\begin{cases} \text{if } \bar{s}_k \in \mathcal{K}_{\kappa_{k-1}} \begin{cases} s_k = \bar{s}_k \,, \\ \kappa_k = \kappa_{k-1} \,, \end{cases} \\ \text{if } \bar{s}_k \notin \mathcal{K}_{\kappa_{k-1}} \begin{cases} s_k = \tilde{s}_k \,, \\ \kappa_k = \kappa_{k-1} + 1 \,. \end{cases} \end{cases} \tag{7}$$

As already done in [5], 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:

**Theorem 1** *(Delyon, Lavielle, Moulines) Assume that :*

**SA0** *w.p.1, for all $k \geq 0$, $s_k \in \mathcal{S}$.*

**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 $\mathcal{S}$ 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 closed convex set $\mathcal{S}_a \subset \mathcal{S}$ for which $s \to \rho(h(s) + e(x) + r(x)) \in \mathcal{S}_a$ for any $\rho \in [0, 1]$ and $(s, x) \in \mathcal{S}_a \times \mathbb{R}^N$ ($\mathcal{S}_a$ is absorbing), 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, considering $(s_k)_{k \geq 0}$ given by the truncated procedure, w.p.1, $\limsup_{k \to \infty} d(s_k, \mathcal{L}') = 0$.*

We want to apply this theorem to 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 Theorem 3 in [5]).

Let us consider the following stochastic approximation: $(x_k, \lambda_k)$ are simulated by the transition kernel described in the previous section and $s_k = s_{k-1} + \Delta_k (S(x_k, \lambda_k) - s_{k-1})$, which can be connected to the Robbins Monro procedure using the notations introduced in [5]: let $\mathcal{F} = (\mathcal{F}_k)_{k \geq 1}$ be the filtration where $\mathcal{F}_k$ is the $\sigma-$algebra generated by the random variables $(S_0, x_1, \ldots, x_k, \lambda_1, \ldots, \lambda_k)$ $\mathbb{E}_{\pi_\eta}$ is the expectation with respect to the posterior distribution $\pi_\eta$ and

$$
\begin{aligned}
h(s_{k-1}) &= \mathbb{E}_{\pi_{\hat{\eta}(s_{k-1})}}\left[S(x,\lambda)\right] - s_{k-1}\,, \\
e_k &= S(x_k,\lambda_k) - \mathbb{E}\left[S(x_k,\lambda_k)|\mathcal{F}_{k-1}\right]\,, \\
r_k &= \mathbb{E}\left[S(x_k,\lambda_k)|\mathcal{F}_{k-1}\right] - \mathbb{E}_{\pi_{\hat{\eta}(s_{k-1})}}\left[S(x,\lambda)\right]\,.
\end{aligned}
$$

**Theorem 2** *Let* $w(s) = -l\circ\hat{\eta}(s)$ *where* $l(\eta) = \log\sum_\lambda \int q(y,x,\lambda,\eta)dx$ *and* $h(s) = \sum_\lambda \int_x (S(x,\lambda) - s)\pi_{\hat{\eta}(s)}(x,\lambda)dx$ *for* $s \in \mathcal{S}$. *Assume that:*

*(A1) The sequence* $(\Delta_k)_{k\geq 1}$ *is non-increasing, positive and satisfy:*
$\sum_{k=1}^{\infty} \Delta_k = \infty$ *and* $\sum_{k=1}^{\infty} \Delta_k^2 < \infty$.

*(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 (7),* 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* $x_0 \in \mathrm{K}$, $\lambda_0 \in \mathcal{T}$ *and* $s_0 \in \mathcal{K}_0$, *we have*

$$
\lim_{k\to\infty} d(s_k, \mathcal{L}') = 0 \ \bar{\mathbb{P}}_{x_0,\lambda_0,s_0,0} \text{ -}a.s.\,,
$$

*where* $\bar{\mathbb{P}}_{x_0,\lambda_0,s_0,0}$ *is the probability measure associated with the chain* $Z_k = (x_k,\lambda_k,s_k,\kappa_k)$ *for* $k \geq 0$ *starting at* $(x_0,\lambda_0,s_0,0)$.

The proof of this theorem is given in Appendix. It will follow the scheme of the proof of Theorem 5 in [5]. The only difference between our algorithm and SAEM is the simulation of the missing data which is not done through the posterior law but through an approximation which can be arbitrarily close.

**Corollary 1** *Under the assumptions of Theorem 2 we have for all* $x_0 \in \mathrm{K}$, $\lambda_0 \in \mathcal{T}$ *and* $\eta_0 \in \Theta \times \varrho$, $\lim_{k\to\infty} d(\eta_k, \mathcal{L}) = 0 \ \ \bar{\mathbb{P}}_{x_0,\lambda_0,s_0,0}$-$a.s$, *where* $\bar{\mathbb{P}}_{x_0,\lambda_0,s_0,0}$ *is the probability measure associated with the chain* $Z_k = (x_k,\lambda_k,s_k,\kappa_k)$, $k \geq 0$ *starting at* $(x_0,\lambda_0,s_0,0)$ *and* $\mathcal{L} \triangleq \{\ \eta \in \hat{\eta}(\mathcal{S}),\ \frac{\partial l}{\partial \eta}(\eta) = 0\}$.

***proof* 1** *This is a direct consequence of the smoothness of the function* $s \mapsto \hat{\eta}(s)$ *on* $\mathcal{S}$ *and Lemma 2 of [5].*

# 4 Experiments

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 $(16 \times 16)$ gray level image with intensity in $[0, 2]$ where 0 corresponds to the black background.

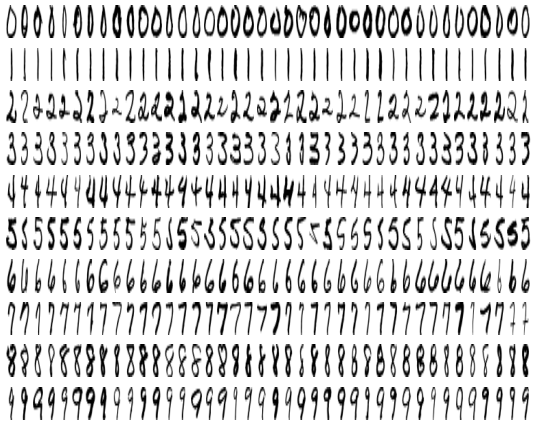

Figure 1: Some example of the training set: 40 images per class (inverse video).

In Figure (1) we show some of the training images used for the statistical estimation.

A natural choice for the prior laws on $\alpha$ and $\Gamma_g$ is to set 0 for the mean on $\alpha$ 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 $M_p(k, k') = K_p(r_{p,k}, r_{p,k'})\ \forall 1 \leq k, k' \leq k_p$ , and $M_g(k, k') = K_g(r_{g,k}, r_{g,k'})\ \forall 1 \leq k, k' \leq k_g$ . 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 $K_g$, where the standard deviations are fixed: $\sigma_p = 0.2$ and $\sigma_g = 0.3$ (for an estimation on $[-1.5, 1.5]^2$ and $[-1, 1]$ respectively).

For the stochastic approximation step-size, we allow a heating period 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 $q(\beta, \lambda|y)$ before exploring this particular region. In the experiments presented, the heating time 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 seems to be reached. The power of the decreasing sequence is chosen to equal $d = 0.6$.

The multicomponent case has to face the problem of its computational time. Indeed, as we have to approximate the posterior density by running $J$ elements of $\tau_m$ independent Markov chains, the computation time increases linearly with $J$. In our experiments, we have chosen a fixed $J$ for every EM iteration, $J = 50$.

## 4.1 The estimated templates

We are showing here the results of the statistical learning algorithm for our generative model. The initialisation of the parameters is an important choice. To avoid the problems shown in [2], we choose the same initialisation of the template parameter $\alpha$ as they did, that is to say, we set the initial value of $\alpha$ such that the corresponding $I_\alpha$ is the mean of the gray-level training images.

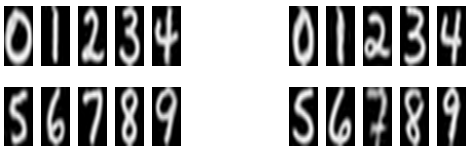

Figure 2: Estimated prototypes of the two components for each digit (40 images per class; 100 iterations; two components per class).

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, the two components reached are meaningful, such as the 2 with and

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

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 the 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 5, 8 or 9).

## 4.2 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: 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.

## 4.3 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 4.

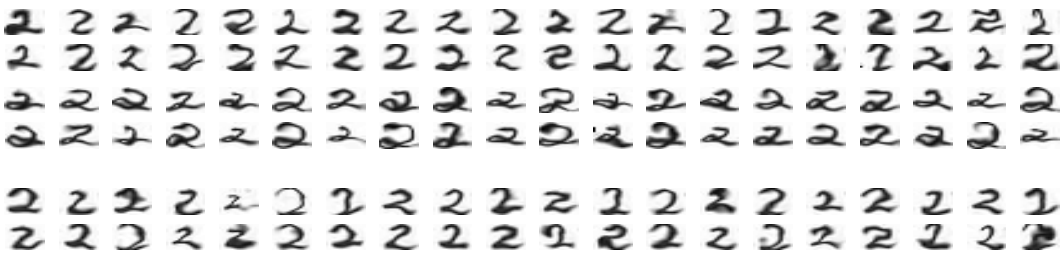

Figure 4: Some synthetic examples of the components of digit 2: First four rows: templates of the two components deformed through some deformation field $\beta$ and $-\beta$ drawn from their respective geometric covariance. Two last row: template of the first component from Figure 2 with deformations drawn with respect to the second component covariance matrix.

For each component, we have drawn the deformation given by the variable $\beta$ and its opposite $-\beta$ since, as soon as one is learnt, because of the symmetry of the 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.

The simulation is done using a common standard Gaussian distribution which is then multiplied by the 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 4, 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 are not as natural. This shows the variability of the models into classes.

# 5 Appendix

Here is the proof of Theorem 2. First let exhibit sufficient statistics for the model. The complete log-likelihood equals:

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

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

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

Thus we apply the stochastic approximation at iteration $k$ of the algorithm leading to: $s_{k,i,t} = s_{k-1,i,t} + \Delta_k(S_{i,t}(x_k,\lambda_k) - s_{k-1,i,t})$ for $0\leq i\leq 4$ and rewrite the maximisation step. The weights and the covariance matrix are updated as follows:

$$\rho_{\tau,k} = \frac{s_{k,0,\tau}+a_\rho}{n+\tau_m a_\rho} \text{ and } \Gamma_{g,\tau,k} = \frac{1}{s_{k,0,\tau}+a_g}(s_{k,0,\tau}s_{k,3,\tau}+a_g\Sigma_g)\,. \tag{8}$$

The photometric parameters are solution of the following system:

$$\begin{cases}\alpha_{\tau,k} &= \left(s_{k,0,\tau}s_{k,2,\tau}+\sigma_{\tau,k}^2(\Sigma_p)^{-1}\right)^{-1}\left(s_{k,0,\tau}s_{k,1,\tau}+\sigma_{\tau,k}^2(\Sigma_p)^{-1}\mu_p\right)\,,\\ \sigma_{\tau,k}^2 &= \frac{1}{s_{k,0,\tau}|\Lambda|+a_p}\left(s_{k,0,\tau}\left(s_{k,4,\tau}+(\alpha_{\tau,k})^t s_{k,2,\tau}\alpha_{\tau,k}-2(\alpha_{\tau,k})^t s_{k,1,\tau}\right)+a_p\sigma_0^2\right)\,,\end{cases} \tag{9}$$

which can be solved iteratively for each component $\tau$ starting with the previous values.

We will now apply Theorem 1 to prove Theorem 2. (**SA0**) is trivially satisfied as well as (**SA1**) since we can choose our step-size sequence $(\Delta_k)_k$.

(**SA2**) holds as already mentioned for the one component case with $w(s) = -l(\hat{\eta}(s))$ such as (**STAB1'(i)**) with the same function $W(s) = -l(\hat{\eta}(s))$ (see [2]).

We need to suppose, like in the one component case, that the critical points of our model are in a compact subset of $\mathcal{S}$ which stands for (**STAB1'(ii)**).

We will now focus on (**STAB2**) and show first 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 $\pi_k = \pi_{\hat{\eta}(s_k)}$ for any $k \geq 0$. We have $r_k = \mathbb{E}\left[S(x_k, \lambda_k)|\mathcal{F}_{k-1}\right] - \mathbb{E}_{\pi_{k-1}}\left[S(x, \lambda)\right]$ thus,

$$\begin{aligned} r_k &= \sum_\lambda \int_x S(x,\lambda)\Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x)\prod_{i=1}^{n}\int p_{J_k}(\tau_i|\xi_i,y_i,\eta_{k-1})\prod_{t=1}^{\tau_m}\prod_{l=1}^{J_k}\Pi_{\eta,t}(\xi_{t,i}^{(l-1)},\xi_{t,i}^{(l)})d\xi_{t,i}^{(l)}dx \\ &\quad -\sum_\lambda \int_x S(x,\lambda)\pi_{\eta_{k-1}}(x,\lambda)dx\,. \end{aligned}$$

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

$$\begin{aligned} |r_k| &\leq \left|\sum_\lambda \int_x S(x,\lambda)\left[\Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x)R_{J_k}(\lambda|y,\eta_{k-1})dx - \pi_{\eta_{k-1}}(x,\lambda)\right]dx\right| \\ &\leq \sum_\lambda \left|\int_x S(x,\lambda)\left[\Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x) - q(x|\lambda,y,\eta_{k-1})\right]dx\right| |R_{J_k}(\lambda|y,\eta_{k-1})| \\ &\quad + \sum_\lambda \left|\int_x S(x,\lambda)q(x|\lambda,y,\eta_{k-1})dx\right| |R_{J_k}(\lambda|y,\eta_{k-1}) - q(\lambda|y,\eta_{k-1})|\,. \end{aligned}$$

Denoting $\mathcal{M}_{\eta_{k-1}} = \max_\lambda \int_x |S(x,\lambda)|q(x|\lambda,y,\eta_{k-1})dx$, we obtain finally

$$\begin{aligned} |r_k|\mathbb{1}_{W(s_{k-1})\leq M} &\leq \sum_\lambda \left|\int_x S(x,\lambda)\left[\Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x) - q(x|\lambda,y,\eta_{k-1})\right]dx\right| \mathbb{1}_{W(s_{k-1})\leq M} \quad (10) \\ &\quad + \mathcal{M}_{\eta_{k-1}}\sum_\lambda |R_{J_k}(\lambda|y,\eta_{k-1}) - q(\lambda|y,\eta_{k-1})|\,\mathbb{1}_{W(s_{k-1})\leq M}. \quad (11) \end{aligned}$$

We will first show that the Gibbs sampler kernel $\Pi_{\eta,\lambda}$ satisfies a minoration condition and a Drift condition (MDRI) to get its geometric ergodicity (as it has been done in [2]).

**(MDRI)** For any $s \in \mathcal{S}$ and any $\lambda \in \mathcal{T}$, $\Pi_{\hat{\eta}(s),\lambda}$ is irreducible and aperiodic. In addition there exist a small set $\mathtt{C}$ ( defined below) and a function $V : \mathbb{R}^N \to [1, \infty[$ such that for any $p \geq 2$ and any compact subset $\mathcal{K} \subset \mathcal{S}$, there exist an integer $m$, constants $0 < \kappa < 1$, $B$ , $\delta > 0$ and a probability measure $\nu$ such that

$$\sup_{s\in\mathcal{K},\lambda\in\mathcal{T}} \Pi^m_{\hat{\eta}(s),\lambda}(x, A) \geq \delta\nu(A) \quad \forall x \in \mathtt{C}, \forall A \in \mathcal{B}(\mathbb{R}^N)\,, \tag{12}$$

$$\sup_{s\in\mathcal{K},\lambda\in\mathcal{T}} \Pi^m_{\hat{\eta}(s),\lambda}V^p(x) \leq \kappa V^p(x) + B\mathbb{1}_{\mathtt{C}}(x)\,. \tag{13}$$

**Notation 1** *Let $(e_j)_{1\leq j\leq N}$ be the canonical basis of the $x$-space and for any $1 \leq j \leq N$, let $E_{\eta,\lambda,j} \triangleq \{\, x \in \mathbb{R}^N \mid \langle x, e_j\rangle_{\eta,\lambda} = 0\}$ be the orthogonal of Span$\{e_j\}$ and $p_{\eta,\lambda,j}$ be the orthogonal projection on $E_{\eta,\lambda,j}$ i.e. $p_{\eta,\lambda,j}(x) \triangleq x - \frac{\langle x,e_j\rangle_{\eta,\lambda}}{|e_j|^2_{\eta,\lambda}} e_j$ , where $\langle x, x'\rangle_{\eta,\lambda} = \sum_{i=1}^n \beta_i^t \Gamma^{-1}_{g,\tau_i}\beta_i$ and $x = \beta^n_1$, $x' = \beta'^n_1$ (i.e. the natural dot product associated with the covariance matrices $(\Gamma_{g,t})_t$).*

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

We first recall the definition of a small set:

**Definition 1** *( [14]) A set $\mathcal{E} \in \mathcal{B}(\mathcal{X})$ is called a **small set** for the kernel $\Pi$ if there exist an integer $m > 0$ and a non trivial measure $\nu_m$ on $\mathcal{B}(\mathcal{X})$, such that for all $x \in \mathcal{E}$, $B \in \mathcal{B}(\mathcal{X})$, $\Pi^m(x, B) \geq \nu_m(B)$. When this holds, we say that $\mathcal{E}$ is $\nu_m$-small.*

We now prove the following lemma:

**Lemma 1** *Let $\mathcal{E}$ be a compact subset of $\mathbb{R}^N$ then $\mathcal{E}$ is a small set of $\mathbb{R}^N$ for $(\Pi_{\hat{\eta}(s),\lambda})_{s\in\mathcal{K},\lambda\in\mathcal{T}}$.*

***proof* 2** *First note that there exists $a_c > 0$ such that for any $\eta \in \Theta \times \varrho$, any $x \in \mathbb{R}^N$ and any $b \in \mathbb{R}$, the acceptance rate $r_{\eta,\lambda,j}(x,b)$ is uniformly lower bounded by $a_c$ so that for any $1 \leq j \leq N$ and any non-negative function $f$,*

$$\Pi_{\eta,\lambda,j} f(x) \geq a_c \int_{\mathbb{R}} f(x_{-j} + be_j) q(b|x_{-j},\lambda,\eta) db = a_c \int_{\mathbb{R}} f(p_{\eta,\lambda,j}(x) + ze_j/|e_j|_{\eta,\lambda}) g_{0,1}(z) dz \,,$$

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

$$\Pi_{\eta,\lambda} f(x) \geq a_c^N \int_{\mathbb{R}^N} f\left( p_{\eta,\lambda,1,N}(x) + \sum_{j=1}^N z_j p_{\eta,\lambda,j+1,N}(e_j)/|e_j|_{\eta,\lambda} \right) \prod_{j=1}^N g_{0,1}(z_j) dz_j \,, \tag{14}$$

*where $p_{\eta,\lambda,q,r} = p_{\eta,\lambda,r} \circ p_{\eta,\lambda,r-1} \circ \cdots \circ p_{\eta,\lambda,q}$ for any integer $q \leq r$ and $p_{\eta,\lambda,N+1,N} = Id_{\mathbb{R}^N}$. Let $A_{\eta,\lambda} \in \mathcal{L}(\mathbb{R}^N)$ be the linear mapping on $z_1^N = (z_1, \cdots, z_N)$ defined by $A_{\eta,\lambda} z_1^N = \sum_{j=1}^N z_j p_{\eta,\lambda,j+1,N}(e_j)/|e_j|_{\eta,\lambda}$. One easily checks that for any $1 \leq k \leq N$, Span$\{ p_{\eta,\lambda,j+1,N}(e_j),\ k \leq j \leq N\}$ = Span$\{e_j, k \leq j \leq N\}$ so that $A_{\eta,\lambda}$ is an invertible mapping. By a change of variable, we get*

$$\int_{\mathbb{R}^N} f\left( p_{\eta,\lambda,1,N}(x) + A_{\eta,\lambda} z_1^N \right) \prod_{j=1}^N g_{0,1}(z_j) dz_j = \int_{\mathbb{R}^N} f(u) g_{p_{\eta,\lambda,1,N}(x), A_{\eta,\lambda} A_{\eta,\lambda}^t}(u) du \,,$$

*where $g_{\mu,\Sigma}$ stands for the density of the normal law $\mathcal{N}(\mu,\Sigma)$. Since $(\eta,\lambda) \to A_{\eta,\lambda}$ is smooth on the set of invertible mappings in $(\eta,\lambda)$, we deduce that there exists $c > 0$ such that $cId \leq A_{\eta,\lambda} A_{\eta,\lambda}^t \leq Id/c$ and $g_{p_{\eta,\lambda,1,N}(x), A_{\eta,\lambda} A_{\eta,\lambda}^t}(u) \geq C g_{p_{\eta,\lambda,1,N}(x), Id/c}(u)$ uniformly for $\eta = \hat{\eta}(s)$ with $s \in \mathcal{K}$ and $\lambda \in \mathcal{T}$. Assuming that $x \in \mathcal{E}$, since $\eta \to p_{\eta,\lambda,1,N}$ is smooth and $\mathcal{E}$ is compact, we have $\sup_{x \in \mathcal{E}, \eta = \hat{\eta}(s),\ s \in \mathcal{K}, \lambda \in \mathcal{T}} |p_{\eta,\lambda,1,N}(x)| < \infty$ so that there exists $C' > 0$ and $c' > 0$ such that for any $(u,x) \in \mathbb{R}^N \times \mathcal{E}$ and any $\eta = \hat{\eta}(s),\ s \in \mathcal{K}, \lambda \in \mathcal{T}$*

$$g_{p_{\eta,\lambda,1,N}(x), A_{\eta,\lambda} A_{\eta,\lambda}^t}(u) \geq C' g_{0, Id/c'}(u). \tag{15}$$

*Using* (14) *and* (15)*, we deduce that for any* $A$ $\Pi_{\eta,\lambda}(x,A) \geq C' a_c^N \nu(A)$ *, with* $\nu$ *equal to the density of the normal law* $\mathcal{N}(0, Id/c')$*. This yields the existence of the small set as well as equation (12).*

This property also implies the $\phi$-irreducibility of the Markov Chain generated by $\Pi_{\eta,\lambda}$. Moreover, the existence of a $\nu_1$-small set implies the aperiodicity of the chain [14].

Now consider the Drift condition (13).

We set $V : \mathbb{R}^N \to [1, +\infty[$ as the following function $V(x) = 1 + |x|^2$,where $|\cdot|$ denotes the Euclidian norm. Define for any $g : \mathbb{R}^N \to \mathbb{R}^{n_s}$ the norm $\|g\|_V = \sup_{x \in \mathbb{R}^N} \frac{|g(x)|}{V(x)}$ 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 $\lambda \in \mathcal{T}$, we introduce a $(\eta, \lambda)$ dependent function $V_{\eta,\lambda}(x) \triangleq 1 + \langle x, x \rangle_{\eta,\lambda}$.

**Lemma 2** *Let* $K$ *be a compact subset of* $\Theta \times \varrho$*. For any integer* $p \geq 1$*, there exist* $0 \leq \rho < 1$ *and* $C > 0$ *such that for any* $\eta \in K$*, any* $\lambda \in \mathcal{T}$*, any* $x \in \mathbb{R}^N$ *we have* $\Pi_{\eta,\lambda} V_{\eta,\lambda}^p(x) \leq \rho V_{\eta,\lambda}^p(x) + C$ *.*

***proof* 3** *The proposal distribution for* $\Pi_{\eta,\lambda,j}$ *is given by* $q(x \mid x_{-j}, \lambda, y, \eta) \stackrel{law}{=} p_{\eta,\lambda,j}(x) + U_{\eta,\lambda} e_j$*, where* $U_{\eta,\lambda} \sim \mathcal{N}(0, |e_j|_{\eta,\lambda}^{-2})$*. Since the acceptance rate* $r_{\eta,\lambda,j}$ *is uniformly bounded from below by a positive number* $a_c > 0$*, we deduce that there exists* $C_K$ *such that for any* $x \in \mathbb{R}^N$ *and any measurable set* $A \in \mathcal{B}(\mathbb{R}^N)$

$$\Pi_{\eta,\lambda,j}(x,A) = (1 - r_{\eta,\lambda,j}(x)) \mathbb{1}_A(x) + r_{\eta,\lambda,j}(x) \int_{\mathbb{R}} \mathbb{1}_A(p_{\eta,\lambda,j}(x) + z e_j) \gamma_{\eta,\lambda}(dz) \,,$$

*where* $\gamma_{\eta,\lambda} \leq C_K \gamma_K$ *and* $\gamma_K$ *equals to the density of the normal law* $\mathcal{N}(0, \sup_{\eta \in K, \lambda \in \mathcal{T}} |e_j|_{\eta,\lambda}^{-2})$*.*

*Since* $\langle p_{\eta,\lambda,j}(x), e_j\rangle_{\eta,\lambda} = 0$, *we get* $V^p_{\eta,\lambda}(p_{\eta,\lambda,j}(x) + ze_j) = (V_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) + z^2|e_j|^2_{\eta,\lambda})^p$ *and*

$$\begin{aligned}
\Pi_{\eta,\lambda,j} V^p_{\eta,\lambda}(x) &= (1 - r_{\eta,\lambda,j}(x))V^p_{\eta,\lambda}(x) + r_{\eta,\lambda,j}(x) \int_{\mathbb{R}} \left(V_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) + z^2|e_j|^2_{\eta,\lambda}\right)^p \gamma_{\eta,\lambda}(dz) \\
&\leq (1 - r_{\eta,\lambda,j}(x))V^p_{\eta,\lambda}(x) + r_{\eta,\lambda,j}(x) \times \\
&\quad \left(V^p_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) + (2^p - 1)C_K V^{p-1}_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) \int_{\mathbb{R}} (1 + z^2|e_j|^2_{\eta,\lambda})^{p-1} \gamma_K(dz)\right) \\
&\leq (1 - r_{\eta,\lambda,j}(x))V^p_{\eta,\lambda}(x) + r_{\eta,\lambda,j}(x)V^p_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) + C'_K V^{p-1}_{\eta,\lambda}(p_{\eta,\lambda,j}(x))\,.
\end{aligned}$$

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

$$\Pi_{\eta,\lambda,j} V^p_{\eta,\lambda}(x) \leq (1 - a_c)V^p_{\eta,\lambda}(x) + (a_c + \ell)V^p_{\eta,\lambda}(p_{\eta,\lambda,j}(x)) + C_{K,\ell}\,.$$

*By induction, starting with j=1, we get*

$$\Pi_{\eta,\lambda} V^p_{\eta,\lambda}(x) \leq \sum_{u\in\{0,1\}^N} \prod_{j=1}^N (1 - a_c)^{1-u_j}(a_c + \ell)^{u_j} V^p_{\eta,\lambda}(p_{\eta,\lambda,u}(x)) + \frac{C_{K,\ell}}{\ell}((1 + \ell)^{N+1} - 1)\,,$$

*where* $p_{\eta,\lambda,u} = ((1 - u_N)Id + u_N p_{\eta,\lambda,N}) \circ \cdots \circ ((1 - u_1)Id + u_1 p_{\eta,\lambda,1})$.

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

$$\Pi_{\eta,\lambda} V^p_{\eta,\lambda}(x) \leq b_{c,\ell} V^p_{\eta,\lambda}(x) + (a_c + \ell)^N V^p_{\eta,\lambda}(p_{\eta,\lambda}(x)) + \frac{C_{K,\ell}}{\ell}((1 + \ell)^{N+1})$$

*for* $b_{c,\ell} = \left(\sum_{u\in\{0,1\}^N,\ u\neq\mathbf{1}} \prod_{j=1}^N (1 - a_c)^{1-u_j}(a_c + \ell)^{u_j}\right)$. *To end the proof, we need to check that* $p_{\eta,\lambda}$ *is strictly contracting uniformly on* $K$. *Indeed,* $|p_{\eta,\lambda}(x)|_{\eta,\lambda} = |x|_{\eta,\lambda}$ *implies that* $p_{\eta,\lambda,j}(x) = x$ *for any* $1 \leq j \leq N$ *so that* $\langle x, e_j\rangle_{\eta,\lambda} = 0$ *and* $x = 0$ *since* $(e_j)_{1\leq j\leq N}$ *is a basis. Using the continuity of the norm of* $p_{\eta,\lambda}$ *and the compactness of* $K$, *we deduce that there*

*exists* $0 < \rho_K < 1$ *such that* $|p_{\eta,\lambda}(x)|_{\eta,\lambda} \leq \rho_K |x|_{\eta,\lambda}$ *for any* $x$, *any* $\eta \in K$ *and any* $\lambda \in \mathcal{T}$. *Changing* $\rho_K$ *for* $1 > \rho'_K > \rho_K$ *we get* $(1 + \rho_K^2 |x|_{\eta,\lambda}^2)^q \leq \rho'^{2q}_K (1 + |x|_{\eta,\lambda}^2)^q + C''_K$ *for some uniform constant* $C''_K$ *so that*

$$\Pi_{\eta,\lambda} V^p_{\eta,\lambda}(x) \leq b_{c,\ell} V^p_{\eta,\lambda}(x) + \rho'^{2p}_K (a_c + \ell)^N V^p_{\eta,\lambda}(x) + C''_{K,\ell}.$$

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

**Lemma 3** *For any compact set* $K \subset \Theta \times \varrho$, *any integer* $p \geq 0$, *there exist* $0 < \rho < 1$, $C > 0$ *and* $m_0$ *such that* $\forall m \geq m_0$ , $\forall \eta \in K$, $\forall \lambda \in \mathcal{T}$ $\Pi^m_{\eta,\lambda} V^p(x) \leq \rho V^p(x) + C$ .

***proof* 4** *Indeed, there exist* $0 \leq c_1 \leq c_2$ *such that* $c_1 V(x) \leq V_{\eta,\lambda}(x) \leq c_2 V(x)$ *for any* $(x, \eta, \lambda) \in \mathbb{R}^N \times K \times \mathcal{T}$. *Then, using the previous lemma, we have* $\Pi^m_{\eta,\lambda} V^p(x) \leq c_1^{-p} \Pi^m_{\eta,\lambda} V^p_{\eta,\lambda}(x) \leq c_1^{-p} (\rho^m V^p_{\eta,\lambda}(x) + C/(1-\rho)) \leq (c_2/c_1)^p (\rho^m V^p(x) + C/(1-\rho))$. *Choosing* $m$ *large enough for* $(c_2/c_1)^p \rho^m < 1$ *gives the result.*

This finishes the proof of (13). Thanks to this property we can use the following proposition and lemma applied to every sequence $(\xi^{(l)}_{t,i})_l$ with stationary distribution $q(\cdot|y_i, t, \eta)$ for all $1 \leq t \leq \tau_m$ and all $1 \leq i \leq n$: (cf: [14], [3] Proposition B1 and Lemma B2).

**Proposition 1** *Suppose that* $\Pi$ *is irreducible and aperiodic and that* $\Pi^m(x_0, .) \geq \mathbb{1}_{\mathcal{C}}(x_0) \delta \nu(.)$ *for a set* $\mathcal{C} \in \mathcal{B}(X)$, *some integer* $m$ *and* $\delta > 0$ *and that there is a Drift condition to* $\mathcal{C}$ *in the sense that, for some* $\kappa < 1$, $B$ *and a function* $V : X \to [1, +\infty[$,

$$\Pi V(x_0) \leq \kappa V(x_0) \ \forall x_0 \notin \mathcal{C} \ and \ \sup_{x_0 \in \mathcal{C}} (V(x_0) + \Pi V(x_0)) \leq B \, .$$

*Then, there exist constants* $K$ *and* $0 < \rho < 1$, *depending only upon* $m, \delta, \kappa, B$, *such that, for all* $x_0 \in X$, *and all* $g \in \mathcal{L}_V$ $\|\Pi^n g(x_0) - \pi(g)\|_V \leq K \rho^n \|g\|_V$ .

**Lemma 4** *Assume that there exist an integer $m$ and constants $\kappa < 1$ and $\varsigma$ such that $\Pi^m V(x_0) \leq \kappa V(x_0)\ \forall x_0 \notin \mathcal{C}$ and $\Pi V(x_0) \leq \varsigma V(x_0)\ \forall x_0 \in X$ for some function $V : X \to [1, +\infty[$. Then there exists a function $\tilde{V}$ and constants $0 < \rho < 1, c$ and $C$, depending only upon $m, \kappa, \varsigma$, such that, $\Pi\tilde{V}(x_0) \leq \kappa\tilde{V}(x_0)\ \forall x_0 \notin \mathcal{C}$ and $cV \leq \tilde{V} \leq CV$ .*

So each Gibbs sampler kernel $\Pi_{\eta,\lambda}$ is geometrically ergodic.

We will now use the term $\mathbb{1}_{W(s_{k-1})\leq M}$ to show that the parameters $\eta_{k-1}$ 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 (1) implies that for any $\theta \in \Theta$ we have:

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

so that denoting $C$ as a constant:

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

It was shown in [1] that we have $\lim_{||\Gamma||+||\Gamma^{-1}||\to\infty} -\frac{a_g}{2}\langle\Gamma^{-1}, \Sigma_g\rangle + \frac{1+a_g}{2}\log|\Gamma^{-1}| = -\infty$, $\lim_{\sigma^2+\sigma^{-2}\to\infty} -\frac{a_p\sigma_0^2}{2\sigma^2} - \frac{|\Lambda|+a_p}{2}\log(\sigma^2) = -\infty$ 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 $\lim_{\eta\to\partial(\Theta\times\varrho)}\log q(y,\eta) = -\infty$. 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 $\sigma_t^2 + \sigma_t^{-2} \geq \ell$ 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 $|\alpha_t| \leq \ell$, $||\Gamma_t|| + ||\Gamma_t^{-1}|| \leq \ell$, $\sigma_t^2 + \sigma_t^{-2} \leq \ell$ and $\frac{1}{\ell} \leq \rho_t \leq 1 - \frac{1}{\ell}$ because $\sum_{t=1}^{\tau_m}\rho_t = 1$. Let us denote $\mathcal{V}_\ell = \Theta_\ell^{\tau_m} \times \left\{(\rho_t)_{1\leq t\leq\tau_m} \in \left[\frac{1}{\ell}, 1-\frac{1}{\ell}\right]^{\tau_m} \mid \sum_{t=1}^{\tau_m}\rho_t = 1\right\}$, where $\Theta_\ell = \left\{\theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p},\ \sigma > 0\,,\ \Gamma_g \in \mathrm{Sym}^+_{2k_g,*}(\mathbb{R}) \mid |\alpha| \leq \ell, \frac{1}{\ell} \leq \sigma^2 \leq \ell, \frac{1}{\ell} \leq ||\Gamma_g|| \leq \ell\right\}$. 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 (10) can be bounded as follows:

$$\sum_{\lambda} \left| \int_x S(x,\lambda) \left[ \Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x) - q(x|\lambda,y,\eta_{k-1}) \right] dx \right| \mathbb{1}_{W(s_{k-1})\leq M}$$
$$\leq \sum_{\lambda} \sup_{\eta\in\mathcal{V}_\ell} \left| \int_x S(x,\lambda) \left[ \Pi^{J_k}_{\eta,\lambda}(x_0,x) - q(x|\lambda,y,\eta) \right] dx \right| .$$

Since for each $\lambda$ the function $x \to S(x,\lambda)$ belongs to $\mathcal{L}_V$, since we have proved that each transition kernel $\Pi_{\eta,\lambda}$ is geometrically ergodic and since the set $\mathcal{V}_\ell$ is compact, we can deduce that the first term (10) converges to zero as $J_k$ tends to infinity.

We now consider the second term (11). 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 $(j,m) \in \{1,...,2k_g\}^2$:

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

where $C(\mathcal{V}_\ell)$ is a constant depending only on the set $\mathcal{V}_\ell$, $\hat{\Gamma}_{g,\lambda}$ is the diagonal block matrix with all the $\Gamma_{g,\tau_i}$ given by the label vector $\lambda$ and we have changed the variable in the last inequality and $Q$ is a quadratic form in $x$ whose coefficients are continuous functions of elements of the matrix $\Gamma_g$. 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 (11).

Let us denote $\mathcal{R}_{\lambda,y,k}$ for the term $|R_{J_k}(\lambda|y,\eta_{k-1}) - q(\lambda|y,\eta_{k-1})|$. Thus we have:

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

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:

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

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

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

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:

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

So we deduced that:

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

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 [7], 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 $(\eta_k)$ only moves in a compact set $\mathcal{V}_\ell$ thanks to the condition $\mathbb{1}_{W(s_{k-1}\leq M)}$. This yields:

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

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 $\xi$, we have for all $\eta \in \mathcal{V}_\ell$:

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

where $\sigma_\ell$ is the lower bound of the variance $\sigma$ on the compact set $\mathcal{V}_\ell$.

So choosing the sequences $(\zeta_k)_k$ and $(J_k)_k$ such that $\lim_{k\to\infty}\zeta_k = 0$ and $\lim_{k\to\infty} J_k\zeta_k^2 = +\infty$ we get the result. We can take for example $J_k = k$ and $\zeta_k = k^{-1/3}$ 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 $\sigma-$algebra generated by the random variables $(S_0, x_1, \ldots, x_k, \lambda_1, \ldots, \lambda_k)$. We have $M_n = \sum_{k=1}^{n}\Delta_k\left(S(x_k,\lambda_k) - \mathbb{E}\left[S(x_k,\lambda_k)|\mathcal{F}_{k-1}\right]\right)\mathbb{1}_{W(s_{k-1})\leq M}$ so this shows us that $(M_n)$ is a $\mathcal{F}$-martingale. In addition to this we have:

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

We now evaluate this last integral term:

$$\begin{aligned}\mathbb{E}\left[|S(x_k,\lambda_k)|^2 \mid \mathcal{F}_{k-1}\right] &= \sum_{\lambda}\int_x\int_\xi |S(x,\lambda)|^2\Pi_{\eta_{k-1},\lambda}^{J_k}(x_0,x)\prod_{i=1}^{n}p_{J_k,\eta_{k-1}}(\tau_i,\xi_{\tau_i,i},y_i)Q(\xi_{\tau_i,i})d\xi_{\tau_i,i}dx \\ &\leq \left[\sum_{\lambda}\int_x |S(x,\lambda)|^2\Pi_{\eta_{k-1},\lambda}^{J_k}(x_0,x)dx\right]\left[\int_\xi \Pi_{\eta_{k-1},\lambda}^{J_k}(\xi_0,\xi)d\xi\right].\end{aligned}$$

The last term equals one and again we only need to focus on the sufficient statistic which

is not bounded itself. Indeed $S_{3,t}(x,\lambda)$ for all $1 \leq t \leq \tau_m$ so using the fact that the function $V$ dominates this sufficient statistic, we obtain:

$$\begin{aligned}\mathbb{E}\left[|S_{3,t}(x_k,\lambda_k)|^2 \mid \mathcal{F}_{k-1}\right] &\leq \sum_{\lambda}\int_x |S_{3,t}(x,\lambda)|^2 \Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x)dx \\ &\leq C\sum_{\lambda}\int_x V(x)^2 \Pi^{J_k}_{\eta_{k-1},\lambda}(x_0,x)dx \leq C\sum_{\lambda}\Pi^{J_k}_{\eta_{k-1},\lambda}V(x_0)^2\,.\end{aligned}$$

Applying Lemma 3 for $p=2$, we get:

$$\mathbb{E}\left[|S(x_k,\lambda_k)|^2 \mid \mathcal{F}_{k-1}\right] \leq C\sum_{\lambda}(\rho V(x_0)^2 + C) \leq C\tau_m^n(\rho V(x_0)^2 + C)\,.$$

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 [11]). This proves the first part of (**STAB2**).

To conclude this proof we apply Theorem 1 and get that $\lim_{k\to\infty} d(s_k,\mathcal{L}') = 0$.

# References


[1] *Toward a coherent statistical framework for dense deformable template estimation* S. Allassonnière, Y. Amit, A. Trouvé, Journal of the Royal Statistical Society, 69(3-29), 2007.

[2] *Bayesian Deformable Models Bulding via Stochastic Approximation Algorithm: A convergence Study* S. Allassonnière, E. Kuhn, A. Trouvé submitted.

[3] *Stability of stochastic approximation under verifiable conditions* C. Andrieu, E. Moulines, P. Priouret SIAM J. Control Optim, 2005, 44(283-312).

[4] *Actives Appearance Models* T.F. Cootes,G.J. Edwards, C.J. Taylor 5th European Conference on Computer Vision, Berlin, 1998, 2(484-498).

[5] *Convergence of a stochastic approximation version of the EM algorithm* B. Delyon, M. Lavielle, E. Moulines The Annals of Statistics, 27(94-128), 1999

[6] *Maximum likelihood from incomplete data via the EM algorithm* A.P. Dempster, N.M. Laird, D.B. Rubin Journal of the Royal Statistical Society, 1977, (1-22).

[7] *Nonparametric Density Estimation in Hidden Markov Models* C.C.Y. Dorea, L.C. Zhao Statistical Inference for Stochastic Processes, 5(55-64), 2002.

[8] *A penalised likelihood approach to image warping* C.A. Glasbey, K.V. Mardia Journal of the Royal Statistical Society, Series B,63(465-492), 2001.

[9] *Template estimation form unlabeled point set data and surfaces for Computational Anatomy* J. Glaunès, S. JoshiProc. of the International Workshop on the Mathematical Foundations of Computational Anatomy, (29-39), 2006.

[10] *General Pattern Theory* U. Grenander Oxford Science Publications, 1993.

[11] *Martingale limit theory and its application* P. Hall, C.C. Heyde Probability and Mathematical Statistics, Academic Press Inc., 1980.

[12] *Coupling a stochastic approximation version of EM with an MCMC procedure* E. Kuhn, M. Lavielle ESAIM Probab. Stat. 8(115-131), 2004.

[13] *A minimum description length objective function for groupewise non-rigid image registration* S. Marsland, C.J. Twining, C.J. Taylor Image and Vision Computing, 2007.

[14] *Markov chains and stochastic stability* S.P. Meyn, R.L. Tweedie Springer-Verlag London Ltd. 1993

[15] *Méthodes de Monte Carlo par chaînes de Markov* C. RobertÉditions Économica, 1996.