Bayesian mixed effect atlas estimation with a diffeomorphic deformation model

Stéphanie Allassonnière, Stanley Durrleman, Estelle Kuhn

Unité MIA Jouy-en-Josas
Mathématiques et Informatique Appliquées

INRA

Bayesian Mixed Effect Atlas Estimation with a Diffeomorphic Deformation Model

Stéphanie ALLASSONNIERE, Stanley DURRLEMAN, Estelle KUHN

Rapport technique 2014-2, juin 2014

UR341 Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France
http://www.jouy.inra.fr/mia

Stéphanie ALLASSONNIERE
Ecole Polytechnique, CMAP, Route de Saclay, 91128 Palaiseau, FRANCE

Stanley DURRLEMAN
INRIA, ARAMIS Team, 23 avenue d'Italie, 75013 Paris, FRANCE et Institut du Cerveau et de la Moelle épinière, Hôpital de la Pitié Salpétrière, 47 boulevard de l'hôpital, 75013 Paris, FRANCE

Estelle KUHN
INRA, Unité MIA, Domaine de Vilvert, 78352 Jouy-en-Josas, FRANCE
estelle.kuhn@jouy.inra.fr

Abstract

In this paper we introduce a diffeomorphic constraint on the deformations considered in the deformable Bayesian Mixed Effect (BME) Template model. Our approach is built on a generic group of diffeomorphisms, which is parametrized by an arbitrary set of control point positions and momentum vectors. This enables to estimate the optimal positions of control points together with a template image and parameters of the deformation distribution which compose the atlas. We propose to use a stochastic version of the Expectation-Maximization (EM) algorithm where the simulation is performed using the Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA). We propose also an extension of the model including a sparsity constraint to select an optimal number of control points with relevant positions. Experiments are carried out on the USPS database and on mandibles of mice.

Keywords: Deformable template model; atlas estimation; diffeomorphic deformations; stochastic algorithm; Anisotropic MALA; control point optimization; sparsity.

1 Introduction

In this paper, we are interested in summarizing the variability observed in a collection of images by a small number of characteristics. Each image is considered as a different instance of the same "shape" or "object", like the same digit written by different people or scans of the same bone observed in different individuals. This problem can be addressed in the framework of Computational Anatomy as introduced in (20). The goal is to find a representative image of the object, called template, and to quantify the observed variability in shape of this object by template-to-subject deformations. This geometric characteristic together with the template form the atlas. The corresponding model assumes that each observed image is a smooth deformation of the template plus an additive noise. The deformation is defined in the underlying space included in $\mathbb{R}^d$, where d equals 2 or 3 in applications. The template-to-subject deformations are used to quantify the geometric variability of the population via a metric on the set of characteristic mappings.

The study of the mathematical foundations of this deformable template model have been initiated in (19). The characterization of the geometric variability has been addressed in different ways for example in (33) or (32). This model could be considered from either a deterministic or stochastic point of view based on the idea that observations come from deformations of the template image. Such approaches were developed among others in (28; 25; 24; 27; 11) for instance, and have demonstrated great impact in the field of image analysis. Models of deformations usually differ in the smoothness constraint of the mappings, which has to be adapted to the observations. To deal with the large geometric variability observed in real data, one could not restrict deformations to be rigid-body and should consider non-rigid deformations that may have up to an infinite number of degrees of freedom. A simple model of deformations may be the so-called "linearized deformations". A linearized deformation ϕ is defined by the displacement field v of each point in the domain $D \subset \mathbb{R}^d$: $\forall r \in D, \quad \phi(r) = r + v(r)$. The main advantage of this class of deformations is its numerical simplicity as it parametrizes the deformation by a single vector field v. Nevertheless, even with regularity conditions on v, there is no guarantee that the deformation is invertible, meaning that the deformation may create holes or overlapping regions in the domain. To avoid such unrealistic behaviors, one wants to consider diffeomorphic maps which preserve the topology of the shapes in the image set. This amounts to assume that every sample has the same topology or equivalently that differences within sample shapes do not rely on changes of topology.

Diffeomorphic deformations can be built on linearized deformations in the framework of the Large Diffeomorphic Deformation Metric Mapping (LDDMM), which has been introduced in (33; 12) and further developed among others in (22; 30; 10; 18; 21; 5). In this framework, the above linearized deformations are considered as infinitesimal deformations, and the vector field v is seen as an instantaneous velocity field. The composition of such deformations creates a flow of diffeomorphisms, which can be written at the limit as the solution of a differential equation. The set of such diffeomorphisms can be equipped with a group structure and a right-invariant

metric, providing regularity on the driving velocity fields. It follows that the set of images is given the structure of an infinite-dimensional manifold, on which distances are computed as the geodesic length in the deformation group between the identity map and the diffeomorphism that maps one image to another one.

It has been shown in (15) that this infinite dimensional deformation set can be efficiently approximated by a finite control point parametrization carrying momentum vectors. This finite dimension reduction is a key aspect for statistical analysis. Durrleman et al. have enforced the velocity fields that are defined everywhere in the domain to be parameterized by a finite set of control points (see (16)). Positions of control points are not given as a prior but optimized as parameters of the statistical model. As a consequence, control points tend to move to the regions showing the largest variability among samples while optimizing a least-square criterion. This optimization allows at the same time to reduce the number of control points for the same matching accuracy, compared to the case where control points are fixed as the nodes of a regular lattice.

Once the deformation model has been fixed, one needs to estimate the parameters of the associated statistical model including in particular the template image. Different algorithms have been proposed to solve the template estimation. Most of them are based on a deterministic gradient descent. In particular, Durrleman et al. manage simultaneously the optimization in control point positions and momentum vectors thanks to a joint gradient descent (see (16)). Although providing visually interesting results in several practical cases, the nature of the limit is not identified. Moreover, this type of methods fails in specific cases, in particular using noisy training data. Another point of view is to consider stochastic algorithms. For example, Zhang et al. used an Hamiltonian Monte Carlo sampler into a Monte Carlo Expectation Maximization algorithm in the dense LDDMM setting, although there is no theoretical convergence property proved for this algorithm (see (35)). In the linearized deformation setting a convergent algorithm has been proposed in (4) to solve the atlas estimation issue. It is based on the Stochastic Approximation Expectation Maximization (SAEM) introduced in (13), and further extended using Monte Carlo Markov Chain (MCMC) methods in (23; 4), thus allowing for wider scope of applications.

In this paper, we aim at estimating an atlas of a population of images which is composed of a template image and a quantification of the geometric variability using the deformable template framework. We consider the LDDMM setting where the deformations are parametrized by a finite number of initial control point positions and momenta such as in (16). To this purpose, we extend the generative statistical model given in (1). In that model, the deformations are assumed to be linearized and are modeled as random variables which are not observed. This enables to estimate the representative parameters of their distribution which will characterize the geometric variability. On the one hand, we extend this approach to the LDDMM framework. On the other hand, we introduce the control point positions as population parameters into the model so that they can be optimized in the estimation process. This enables to better fit the deformation model leading to a more accurate

estimation of the geometric parameters.

From an algorithmic point of view, we propose to use the Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA) within SAEM algorithm introduced in (2). This algorithm has shown very interesting theoretical and numerical properties. Indeed, the AMALA sampler enables to better explore the target distribution support in very high dimension compared to other samplers. It also increases the speed of convergence of the estimation algorithm. Moreover, we take advantage in our sampler of the efficient computation used in the joint gradient descent in (16) so that the optimization of control point positions is of no additional cost at each iteration.

Another interesting question is how to optimize the number of control points required to parametrize the deformations. Indeed, the number of control points essentially depends on the variability in the data: it should be estimated rather than fixed by the user. In the geometric approach given in (16), control points were automatically selected using a L^1 type penalty that tends to zero out momentum vectors of small magnitude. Numerically it is solved by an adapted gradient descent called FISTA (see (9)). However, this penalty acts on each observation separately, meaning that a control point that is needed to match only a single observation will be kept in the final set of control points. From a statistical point of view, one may think about this control point as an outlier and would like to remove it from the basis. The L^1 penalty is also not suitable for statistical purposes, since its associated distribution, namely the Laplace prior, does not generate sparse variables. In other words, the criterion with L^1 penalty that is minimized in (16) could not be interpreted as the log likelihood of a statistical model generating sparse solutions.

In this paper, we propose to include a sparsity constraint in the parameter space of our statistical model through a thresholding step, borrowing ideas from the Group LASSO literature initiated in (8). This has the advantage to select control points based on their importance for the description of the variability of the whole population, and not only of one single sample. The thresholding step is then included in the Maximization step, so that the same AMALA-SAEM algorithm can be used for the estimation process. We also exhibit a criterion to select an optimal threshold leading to an optimal number of control points.

This paper is organized as follows. We first recall the LDDMM setting in the case of control point parametrization in Section 2. The generative statistical model derived for the atlas estimation issue is presented in Section 3. The AMALA-SAEM algorithm is detailed in Section 4. The extension toward sparsity is presented in Section 5. Section 6 is devoted to experiments on hand written digit and mouse mandible images. Section 7 gives conclusions and perspectives for this work.

2 Model of diffeomorphic deformations

2.1 Large Deformation Diffeomorphic Metric Mapping

The model of diffeomorphic deformations we choose is derived from the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework (see (33; 14; 28)), which generalizes the linearized deformation setting that has been used for the statistical estimation of atlases in (1). In the linearized deformation setting, the deformation ϕ is given by:

$$\phi(r) = r + v(r), \quad \forall\, r \in D\,, \tag{1}$$

with $d = 2$ or 3, and v a vector field on $\mathbb{R}^d$.

It is convenient to assume that v belongs to a Reproducible Kernel Hilbert Space (RKHS) in order to control its spatial smoothness. Essentially, the RKHS V is the set of the square integrable functions regularized by the convolution with a known kernel K_g : $V = L^2(D) * K_g$ (see (7) for more theoretical details.)

One may further assume that v is given as a finite linear combination of the RKHS basis vector fields, yielding:

$$v(r) = \sum_{k=1}^{k_g} K_g(r, c_k)\alpha_k, \tag{2}$$

where $(c_k)_k$ is a set of k_g control points and $(\alpha_k)_k$ the vector weights attached to the control points (called momenta in this context). The value of the vector field at any position r is obtained by interpolating the momenta located at the control points.

The advantage of this formulation is to restrict the vector v to belong to a finite-dimensional sub-space of the RKHS, which allows the straightforward definition and estimation of statistics such as means and covariance matrices. Another advantage is that v depends linearly on the momentum vectors, which greatly eases the derivation of the statistical model equations.

However, if the magnitude of the vector $v(r)$ or if the Jacobian of the vector field v becomes too large, then the linearized deformations are not invertible, thus leading to foldings or holes that may be undesirable for applications.

The LDDMM framework offers a generic way to define diffeomorphic maps, which guarantees their smoothness and invertibility. The approach introduced in (16) and (17) is a direct extension of the linearized deformation setting. It is still built on parameterization of the diffeomorphic maps lying in a finite dimensional subspace of a given RKHS. However, the dependence of the deformations in their parameters will not be linear anymore.

2.2 Parametric LDDMM construction

To build a diffeomorphic map, we use the linearized deformations given in Equation (1) as infinitesimal steps, and consider the corresponding vector field as an instantaneous velocity field. More precisely, we consider time-dependent velocity fields $(v_t)_t$ for a time-parameter t varying in $[0, 1]$. The motion of a point r_0 in the

domain of interest D describes a curve $t \to r(t)$ which is the integral curve of the following Ordinary Differential Equation (ODE) called flow equation:

$$\begin{cases} \dfrac{dr(t)}{dt} = v_t(r(t)) \\ r(0) = r_0 \,. \end{cases} \tag{3}$$

The deformation ϕ_1 is defined as follows:

$$\forall r_0 \in D, \quad \phi_1(r_0) = r(1) \,.$$

Conditions under which this map ϕ_1 is diffeomorphic can be found in (10). In particular, the existence, uniqueness and diffeomorphic property of the solution are satisfied if the velocity v_t belongs to a RKHS at all time t and is square integrable in time.

Under these conditions, the model builds a flow of diffeomorphic deformations $\phi_t : r_0 \longrightarrow r(t)$ for all $t \in [0, 1]$. The flow describes a curve in a sub-group of diffeomorphic deformations starting at the identity map. The RKHS V plays the role of the tangent space of such an infinite-dimensional Riemannian manifold at the identity map Id. We can provide this group of diffeomorphisms with a right-invariant metric, where the square distance between the identity map $Id = \phi_0$ and the final deformation ϕ_1 is given as the total kinetic energy used along the path: $d(Id, \phi_1)^2 = \int_0^1 \|v_t\|_V^2 dt$, where $\| \cdot \|_V$ is the norm in the RKHS. The existence and uniqueness of minimizing paths have been shown in (28).

According to mechanical principles, one can show that the kinetic energy is preserved along the geodesic paths, namely for all $t \in [0, 1]$ $\|v_t\|_V = \|v_0\|_V$. Moreover, the velocity fields (v_t) along such paths satisfy Hamiltonian equations, meaning that the geodesic is fully parametrized by the initial velocity field v_0. This velocity field plays the role of the Riemannian logarithm of the final diffeomorphism ϕ_1. Therefore, it belongs to a vector space and allows the definition of tangent-space statistics in the spirit of (34) and (31).

Following (17) and (16), we further assume that v_0 is the interpolation of momentum vectors $(\alpha_{k,0})_k$ at control point positions $(c_{k,0})_k$:

$$v_0(r) = \sum_{k=1}^{k_g} K_g(r, c_{k,0}) \alpha_{k,0} \,, \tag{4}$$

where K_g is the kernel associated to the RKHS V. In this context, it has been shown in (29) that the velocity fields $(v_t)_t$ along the geodesic path starting at the identity map in the direction of v_0 keep the same form:

$$v_t(r) = \sum_{k=1}^{k_g} K_g(r, c_k(t)) \alpha_k(t) \,, \tag{5}$$

where the control point positions $(c_k(t))_k$ and the momentum vectors $(\alpha_k(t))_k$ satisfy

the Hamiltonian equations:

$$\begin{cases} \dfrac{dc_k(t)}{dt} = \displaystyle\sum_{l=1}^{k_g} K_g(c_k(t), c_l(t))\alpha_l(t) \\ \dfrac{d\alpha_k(t)}{dt} = -\left(\displaystyle\sum_{l=1}^{k_g} d_{c_k(t)}(K_g(c_k(t), c_l(t))\alpha_l(t))\right)^t \alpha_k(t) \end{cases} \tag{6}$$

with initial conditions $c_k(0) = c_{0,k}$ and $\alpha_k(0) = \alpha_{0,k}$ for all $1 \leq k \leq k_g$. This is similar to the equations of motion of a set of k_g self-interacting particles, with K_g modeling the interactions. One can easily verify that the Hamiltonian defined as $H_t = \|v_t\|_V^2 = \sum_{k=1}^{k_g}\sum_{l=1}^{k_g} \alpha_l(t)^t K_g(c_l(t), c_k(t))\alpha_k(t)$ is constant in time when control point positions and momentum vectors satisfy the system (6).

This model defines a finite dimensional subspace of the group of diffeomorphisms. For a given set of initial control points, the diffeomorphisms are parametrized by the momentum vectors attached to them. For one instance of the initial momentum vectors, one builds the motion of the control points and of the momentum vectors by integrating the Hamiltonian system (6). Then, they define a dense velocity field at each time t according to Equation (5). Finally, one can find the motion $\phi_t(r_0)$ of any point r_0 in the domain D by integrating the flow equation (3). In this framework, the tangent-space representation of the diffeomorphic deformation ϕ_1 is given by the initial velocity field v_0 parametrized by $z = ((c_{0,k}, \alpha_{0,k}))_k$, called the initial state of the particle system. The position $\phi_1(r)$ depends on the parameters $((c_{0,k}, \alpha_{0,k}))_k$ via the integration of two non-linear differential equations in Equation (6) and Equation (3).

Remark 1 *The LDDMM framework formulation involves a coupling on the control point and the momentum evolutions along the geodesic path which is not the case in the linearized deformation setting. This joint equation introduces more constraints reducing the dimension of the solution space. Therefore, the identifiability of the control point positions may be expected in our LDDMM framework. This property would most probably fail in the linearized deformation setting where the momenta and the control points are not coupled.*

In Section 3, we will define the stochastic model of deformations based on parametric distributions of the initial state of particles.

3 Statistical model and parameter estimation

As pointed out in (1), the gradient descent optimization with respect to the template together with the momenta does not necessarily converge if the training set is noisy. To solve this problem, we introduce here a statistical model where we consider the deformations as well as the control point positions as non-observed random variables, in the spirit of the Bayesian Mixed Effect (BME) Template model given in (4).

3.1 Statistical generative model

We choose to model our data by a generative hierarchical model. This allows to generate images from the atlas. The ability to generate synthetic images is important to interpret the features captured by the model. It may highlight variability patterns that could not be perceived by simple visual inspection of the training images. In this model, the distribution of the deformations in the diffeomorphism group is parametrized. In a statistical approach, these parameters are estimated from the data, thus providing a metric in the shape space which is adapted to the data and takes into account the deformation constraints. This is in contrast to geometric approaches that estimate the template using a fixed metric.

More precisely, let I_0 be a template image: $I_0 : \mathbb{R}^d \to \mathbb{R}$. We consider an observation, namely an image y, as a noisy discretization on a regular grid Λ of a diffeomorphic deformation of the template image. Let ϕ_1^z be the solution of both the flow equation (3) and the Hamiltonian system (6) with initial condition $z = ((c_{0,k}, \alpha_{0,k}))_k$. Then, for all $s \in \Lambda$,

$$y(s) = I_0((\phi_1^z)^{-1}(r_s)) + \sigma\epsilon(s), \tag{7}$$

where $\sigma\epsilon$ denotes an additive centered Gaussian random noise on the grid Λ with variance σ^2, and r_s is the coordinate of the voxel s in the continuous domain D.

We are provided with n images $\mathbf{y} = (y_i)_{1\leq i\leq n}$ in a training set. We assume that each of them follows the probabilistic model (7) and that they are independent.

We consider the initial state of particles, namely the control point positions and the momentum vectors, as random variables and estimate their probabilistic distributions, restricting ourselves to the case of parametric distributions. We assume that control points live in the template domain D and that they are the same for all observations. By contrast, the momentum vectors attached to them are specific to each observation, as they parametrize the matching of the template with each sample image.

Therefore, we propose the following probabilistic model: we assume that the initial control point positions $\mathbf{c}_0 = (c_{0,k})_{1\leq k\leq k_g}$ are drawn through a Gaussian distribution with mean $\bar{\boldsymbol{c}}_{\mathbf{0}}$ and covariance $a_c Id$ where Id is the identity matrix of dimension dk_g. We define the initial momenta $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_0^i)_{1\leq i\leq n}$ with $\boldsymbol{\alpha}_0^i = (\alpha_{0,k}^i)_{1\leq k\leq k_g}$. We assume that the variables $(\boldsymbol{\alpha}_0^i)_{1\leq i\leq n}$ are independent identically distributed and follow a Gaussian distribution with mean 0 and covariance matrix Γ_g. Note that this covariance matrix depends on the initial control point positions as the momenta are attached to them. Moreover the momenta $\boldsymbol{\alpha}_0$ are assumed to be independent of the control point positions $\mathbf{c}_0$ given Γ_g.

Following the same lines as in (1), we parametrize the template function I_0 as a linear combination of gray level values of fixed voxels $(b_k)_{1\leq k\leq k_p}$ equidistributed on the domain D. The interpolation kernel is denoted by K_p and the combination weights are denoted by w. Thus we have for all $r \in D$,

$$I_0(r) = \sum_{k=1}^{k_p} K_p(r, b_k)\mathrm{w}_k\,. \tag{8}$$

The action of a diffeomorphism on this template is the linear combination of the deformed kernel with the same weights: $\forall r \in D$,

$$\mathbf{K}_{\mathbf{p}}^{\boldsymbol{z}}\mathrm{w}(r) = I_0 \circ (\phi_1^z)^{-1}(r) = \sum_{k=1}^{k_p} K_p\left((\phi_1^z)^{-1}(r), b_k\right)\mathrm{w}_k\,. \tag{9}$$

The parameters of the model are $\theta = (\mathrm{w}, \sigma^2, \Gamma_g, \bar{\boldsymbol{c}}_{\mathbf{0}})$ and the random variables $(\boldsymbol{\alpha}_0, \mathbf{c}_0)$ are considered as hidden random variables. As we often deal with small sample size in practice, we restrict our inference to a Bayesian setting. Some of the priors can be informative as the one of Γ_g. Other priors may be non-informative as for the expectation of the control point positions for which no additional information is available. The complete model writes therefore:

$$\begin{cases} \theta = (\mathrm{w}, \sigma^2, \Gamma_g, \bar{\boldsymbol{c}}_{\mathbf{0}}) \sim \nu_p \otimes \nu_g \\ \\ \mathbf{c}_0 \sim \mathcal{N}_{dk_g}(\bar{\boldsymbol{c}}_{\mathbf{0}}, a_c Id)|\ \theta, \\ \\ \boldsymbol{\alpha}_0^i \sim \mathcal{N}_{dk_g}(0, \Gamma_g)|\ \theta,\ \forall 1 \le i \le n\,, \\ \\ y_i \sim \mathcal{N}_{|\Lambda|}(\mathbf{K}_{\mathbf{p}}^{(\mathbf{c_0}, \boldsymbol{\alpha}_{\mathbf{0}}^{\mathbf{i}})}\mathrm{w}, \sigma^2 Id)\ |\ (\mathbf{c}_0, \boldsymbol{\alpha}_0^i),\ \ \theta,\ \forall 1 \le i \le n\,. \end{cases} \tag{10}$$

We define the prior distributions as follows:

$$\begin{cases} \nu_g(d\Gamma_g, d\bar{\boldsymbol{c}}_{\mathbf{0}}) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g\rangle_F/2)\frac{1}{\sqrt{|\det(\Gamma_g)|}}\right)^{a_g} \cdot \exp\big(-\tfrac{1}{2}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c)^t\Sigma_c^{-1}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c)\big)d\Gamma_g d\bar{\boldsymbol{c}}_{\mathbf{0}}, \\ \nu_p(d\mathrm{w}, d\sigma^2) \propto \exp\big(-\tfrac{1}{2}\mathrm{w}^t\Sigma_p^{-1}\mathrm{w}\big) \cdot \Big(\exp\Big(-\frac{\sigma_0^2}{2\sigma^2}\Big)\frac{1}{\sqrt{\sigma^2}}\Big)^{a_p} d\mathrm{w}d\sigma^2\,, \end{cases}$$

where $\langle ., . \rangle_F$ designs the Frobenius scalar product and the hyper-parameters satisfy $a_g \ge 4k_g + 1$, $\Sigma_g = Id$, $\sigma_0^2 > 0$, $a_p \ge 3$ and Σ_p is derived from the interpolation kernel K_p and the photometric grid $(b_k)_{1\le k\le k_p}$ (see (1) for more details). Concerning the hyper-parameters of the control point prior (μ_c, Σ_c), we choose μ_c to be the vector of the equidistributed grid coordinates. The covariance matrix Σ_c is assumed non-informative. All priors are the natural conjugate priors and are assumed independent to ease derivations.

Remark 2 *From a modeling point of view, the positions of the control points could have been consider as parameters of our model since they are fixed effects of the whole population as well as the template. However considering control points as parameters does not lead to a model belonging to the exponential family. Thus, we could not benefit from the convergence properties and efficient implementation of the SAEM algorithm for this class of models. Therefore, we model the control point positions as random variables following a Gaussian distribution.*

3.2 Parameter estimation

Let us define $\mathbf{y} = (y_1, ..., y_n)$. We consider the Maximum A Posteriori (MAP) estimator denoted by $\hat{\theta}_n$ obtained by maximizing the posterior density of θ conditional

to $\mathbf{y}$ as follows:

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta} q(\theta|\mathbf{y}). \tag{11}$$

For sake of simplicity all likelihoods will be denoted by q. We will state the existence and the consistency of this MAP estimator in the next paragraphs.

3.2.1 Existence of the MAP estimator

We first show that for any finite sample the maximum a posteriori will lie in the parameter set Θ; this is non-trivial due to the highly non-linear relationship between parameters and observations in the model.

Theorem 1 *For any sample* $\mathbf{y}$*, there exists* $\hat{\theta}_n \in \Theta$ *such that* $q(\hat{\theta}_n|\mathbf{y}) = \sup_{\theta\in\Theta} q(\theta|\mathbf{y})$.

Proof 1 *From Equation* (10) *we have that for any* $\theta = (\mathrm{w}, \sigma^2, \Gamma_g, \bar{\boldsymbol{c}}_{\mathbf{0}}) \in \Theta$

$$q(\mathbf{y}|\boldsymbol{\alpha}_0, \mathbf{c}_0, \mathrm{w}, \sigma^2) q(\boldsymbol{\alpha}_0|\Gamma_g) q(\mathbf{c}_0|\bar{\boldsymbol{c}}_{\mathbf{0}}) \leq (2\pi\sigma^2)^{-|\Lambda|/2} (2\pi)^{-k_g} |\det(\Gamma_g)|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\alpha}_0^t \Gamma_g^{-1} \boldsymbol{\alpha}_0\right)$$
$$\times (2\pi a_c)^{-dk_g/2} \exp\left(-\frac{1}{2a_c}\|\mathbf{c}_0 - \bar{\boldsymbol{c}}_{\mathbf{0}}\|^2\right) \tag{12}$$

so that integrating over $\boldsymbol{\alpha}_0$ *and* $\mathbf{c}_0$ *and adding the priors on each parameters, we get:*

$$\log(q(\theta|\mathbf{y})) \leq -\tfrac{a_g}{2}\langle R_g, \Sigma_g\rangle_F + \tfrac{a_g}{2}\log(|\det(R_g)|) - \tfrac{a_p\sigma_0^2}{2\sigma^2} - \tfrac{n|\Lambda|+a_p}{2}\log(\sigma^2)$$
$$- \frac{1}{2}\mathrm{w}^t\Sigma_p^{-1}\mathrm{w} - \frac{1}{2}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c)^t\Sigma_c^{-1}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c) + \ C$$

where $R_g = \Gamma_g^{-1}$*, and* C *does not depend on the parameters. If we denote by* η_g^0 *the smallest eigenvalue of* Σ_g *and by* $\|\cdot\|$ *the operator norm, we get*

$$\langle R_g, \Sigma_g\rangle_F \geq \eta_g^0\|R_g\| \ \textit{and} \ \log(|\det(R_g)|) \leq (2k_g - 1)\log(|\det(R_g)|) - \log(\|\Gamma_g\|)$$

so that

$$\lim_{\|R_g\|+\|\Gamma_g\|\to\infty} -\frac{a_g}{2}\langle R_g, \Sigma_g\rangle_F + \frac{a_g}{2}\log(|\det(R_g)|) = -\infty\,.$$

Similarly, we can show

$$\lim_{\sigma^2+\sigma^{-2}\to\infty} -\frac{a_p\sigma_0^2}{2\sigma^2} - \frac{n|\Lambda| + a_p}{2}\log(\sigma^2) = -\infty\,,$$

$$\lim_{\|\mathrm{w}\|\to\infty} -\frac{1}{2}\mathrm{w}^t\Sigma_p^{-1}\mathrm{w} = -\infty$$

and

$$\lim_{\|\bar{\boldsymbol{c}}_{\mathbf{0}}\|\to\infty} -\frac{1}{2}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c)^t\Gamma_c^{-1}(\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c) = -\infty\,.$$

Now considering the Alexandrov one-point compactification $\Theta \cup \{\infty\}$ *of* Θ*, we have*

$$\lim_{\theta\to\infty} \log q(\theta|\mathbf{y}) \to -\infty\,.$$

Since $\theta \to \log q(\theta|\mathbf{y})$ *is smooth on* Θ*, we get the result.*

3.2.2 Consistency of the MAP estimator

We are interested in the consistency properties of the MAP estimator without making strong assumptions on the distribution of the observations $\mathbf{y}$ denoted by P. We seek to prove the convergence of the MAP estimator to the set Θ_* defined by:

$$\Theta_* = \{\ \theta_* \in \Theta \mid E_P(\log q(y|\theta_*)) = \sup_{\theta\in\Theta} E_P(\log q(y|\theta))\}.$$

Theorem 2 *Assume that Θ_* is non empty. Then, for any compact set $K \subset \Theta$, for all $\varepsilon > 0$,*

$$\lim_{n\to+\infty} P(\ \delta(\hat{\theta}_n, \Theta_*) \geq \varepsilon\ \wedge \hat{\theta}_n \in K\ ) = 0\,,$$

where δ is any metric compatible with the usual topology on Θ.

The proof follows the lines of (1). Indeed, the observed likelihood of our diffeomorphic BME Template model has the same regularity properties and asymptotic behaviors in the parameters as the linearized one.

Remark 3 *In (1), the authors have proved that under a weak additional condition, Θ_* is not empty. This makes the use of an important property of the linearized deformations: the amplitude of the deformation increases as the amplitude of its coefficients increases. This enables to prove that large amplitude deformations would not be suitable to optimize the observed likelihood. In the LDDMM setting, this property cannot be guaranteed anymore. The relation between the range of the deformation and its momenta depends on the curvature of the diffeomorphisms space (which is flat in the linearized deformation framework). Therefore, proving that Θ_* is not empty will require to know the curvature of the deformation space. This is unfortunately not known except in very simple cases (see (26)) preventing from a direct generalization of the previous proof.*

4 Algorithmic method

In this section, we detail the estimation algorithm chosen to maximize the posterior distribution $q(\theta|\mathbf{y})$ in the parameter θ. We use the Stochastic Approximation Expectation Maximization (SAEM) algorithm introduced in (13) coupled with Monte Carlo Markov Chain (MCMC) method as suggested in (23) and (4). Let us first recall the principles of the SAEM-MCMC algorithm in the general case of a model belonging to the curved exponential family. This algorithm is iterative, each iteration consisting in four steps:

Simulation step The missing data (here positions of initial control points and momentum vectors) are drawn using a transition probability of a convergent Markov chain Π_θ having the conditional distribution $\pi_\theta(.) = q(.|\mathbf{y}, \theta)$ as stationary distribution.

Stochastic approximation step A stochastic approximation is done on the sufficient statistics of the model using the simulated value of the missing data and a decreasing sequence of positive step-sizes.

Projection on random boundaries If the result of the stochastic approximation falls outside a compact set of increasing size, it is projected back to a fixed compact set.

Maximization step The parameters θ are updated by maximizing the complete log-likelihood evaluated in the projected sufficient statistics.

Due to the high dimension of the hidden variables (i.e. the initial momenta and control points), we need to pay attention to the sampling step, as detailed in Subsection 4.1. The other steps of the algorithm are presented in Subsection 4.2. The theoretical properties of this estimation algorithm are discussed in Subsection 4.3.

4.1 Simulation step of the stochastic EM

4.1.1 AMALA sampler

In our applications, the missing variables composed of the initial momenta and positions of control points $\mathbf{z} = (\mathbf{c}_0, \boldsymbol{\alpha}_0)$ are of very high dimension. In this case, the AMALA sampler proposed in (2) seems better suited for our stochastic EM algorithm than more standard samplers. For example, the Gibbs sampler solves the problems of low numerical acceptation rate and trapping states by looping over each coordinate to better stride the target density support. However, this involves a huge number of loops and heavy computations in the acceptation ratio preventing from any use in very high dimension. By contrast, the AMALA sampler is more performant in terms of computational time while exploring the target support as well as the Gibbs sampler.

To be more precise, the AMALA sampler is an anisotropic version of the well-known Metropolis Adjusted Langevin Algorithm (MALA), where the covariance matrix of the proposal is optimized to take into account the anisotropy and the coordinate correlations of the target distribution. Using our previous notation, the drift vector denoted by $D_\theta(\mathbf{z})$ is equal to:

$$D_\theta(\mathbf{z}) = \frac{b}{\max(b, |\nabla \log \pi_\theta(\mathbf{z})|)} \nabla \log \pi_\theta(\mathbf{z}) \,, \tag{13}$$

with $b > 0$ a truncation boundary. This vector $D_\theta(\mathbf{z})$ is the concatenation of the truncated gradients with respect to $\mathbf{c}_0$ and $(\boldsymbol{\alpha}_0^i)_{1 \leq i \leq n}$ denoted respectively by $D_\theta^0(\mathbf{z}), D_\theta^1(\mathbf{z}), ..., D_\theta^n(\mathbf{z})$.

Starting from the current value $\mathbf{z}_k$ of the Markov chain, the candidate $\mathbf{z}_c$ is sampled from the Gaussian distribution with expectation $\mathbf{z}_k + \delta D_\theta(\mathbf{z}_k)$ and covariance matrix $\delta \Sigma_\theta(\mathbf{z}_k)$ where $\Sigma_\theta(\mathbf{z})$ is given as:

$$\Sigma_\theta(\mathbf{z}) = \varepsilon Id_{(n+1)dk_g} + diag\left(D_\theta^0(\mathbf{z})D_\theta^0(\mathbf{z})^t, D_\theta^1(\mathbf{z})D_\theta^1(\mathbf{z})^t, ..., D_\theta^n(\mathbf{z})D_\theta^n(\mathbf{z})^t\right) \,, \tag{14}$$

with $\varepsilon > 0$ a small regularization parameter and $\delta > 0$.

We denote by p_θ the probability density function (pdf) of this proposal distribution and by $\rho_\theta(\mathbf{z}_k, \mathbf{z}_c)$ the acceptance rate defined as:

$$\rho_\theta(\mathbf{z}_k, \mathbf{z}_c) = \min\left(1, \frac{\pi_\theta(\mathbf{z}_c) p_\theta(\mathbf{z}_c, \mathbf{z}_k)}{p_\theta(\mathbf{z}_k, \mathbf{z}_c) \pi_\theta(\mathbf{z}_k)}\right) \,. \tag{15}$$

Then, the new value $\mathbf{z}_{k+1}$ of the Markov chain equals $\mathbf{z}_c$ with probability $\rho_\theta(\mathbf{z}_k, \mathbf{z}_c)$ and $\mathbf{z}_k$ otherwise.

Remark 4 *For numerical efficiency, we do not take into account correlations between the momenta and the control point positions in the proposal. Moreover, the observations being independent, the covariance matrix of the momenta is block-diagonal since the momenta are independent conditionally to the control point positions.*

We now move to the computation of the gradient of the conditional distribution logarithm which appears in the drift D_θ. It happens that the conditional distribution logarithm is actually equaled to minus the usual energy used to compute the best match between images in the LDDMM framework. Therefore, we pay attention to the computation of the gradient of this quantity in the following paragraph.

4.1.2 Gradient computation in the LDDMM deformation model

We recall here the result established in (16). For clarity purposes, we adopt compact matrix notation. The initial state of the system, which consists of the initial positions of control points $\mathbf{c}_0$ and their associated momentum vectors $\boldsymbol{\alpha}_0$ is denoted by $\mathbf{z} = (\mathbf{c}_0, \boldsymbol{\alpha}_0)$. The position of this set of particles at later time t is denoted by $\mathbf{z}(t) = (\mathbf{c}_0(t), \boldsymbol{\alpha}(t))$, and satisfies the set of coupled ODEs (6). This system of ODE can be re-written in short as:

$$\begin{cases} \dot{\mathbf{z}}(t) = F(\mathbf{z}(t)) \\ \mathbf{z}(0) = \mathbf{z}\,. \end{cases} \tag{16}$$

Let $X^z(t,.)$ denotes the mapping: $r \in D \to X^z(t,r) = \phi_t^z((\phi_1^z)^{-1}(r))$. For $t = 1$, $X^z(1,.) = \mathrm{Id}_{L^2(D)}$ is the identity map. For $t = 0$, $X^z(0,.) = (\phi_1^z)^{-1}(.)$ is the inverse mapping of the domain D that is needed to deform the images. The interest of using the flow $\phi_t^z \circ (\phi_1^z)^{-1}$ (and not $(\phi_t^z)^{-1}$ for instance) is that the trajectory of any pixel under this flow in D is exactly the same as for the direct flow, but in the reverse direction. More precisely, $X^z(t,.)$ is solution of the following ODE integrated backward from $t = 1$ to $t = 0$:

$$\begin{cases} \dfrac{\partial X^z(t,.)}{\partial t} = -v_t(X^z(t,.)) = -\displaystyle\sum_{k=1}^{k_g} K_g(X^z(t,.), c_k(t))\alpha_k(t) \\ X^z(1,.) = \mathrm{Id}_{L^2(D)}\,, \end{cases} \tag{17}$$

which can be re-written in short as:

$$\begin{cases} \dfrac{dX^z(t,r)}{dt} = G(X^z(t,r), z(t)) \\ X^z(1,r) = r\,, \end{cases} \tag{18}$$

for all $r \in D$.

The solution at $t = 0$, $X^z(0,.)$, is used to deform the template image I_0:

$$I_0((\phi_1^z)^{-1}(r)) = I_0(X^z(0,r)). \tag{19}$$

From a numerical point of view, we discretize the image domain D into an array of $|\Lambda|$ pixels. The map $X^z(t,.)$ is therefore a vector of dimension $d|\Lambda|$: $\{X^z(t,r_s)\}_{s=1,\dots,|\Lambda|}$, which gives the trajectory of any pixel r_s under the flow equation (17) (where G is a map from $\mathbb{R}^{d|\Lambda|} \times \mathbb{R}^{dn_g}$ to $\mathbb{R}^{d|\Lambda|}$). The grey value of the deformed image at pixel r_s is computed by interpolating the grey values in I_0 located at pixels around position $X^z(0,r_s)$ using Equation (8).

Proposition 1 *Let us denote by* $\mathbf{z} = (\mathbf{c}_0, \boldsymbol{\alpha}_0)$ *the* $(n+1)dk_g$ *parameters of a generic criterion* E_θ *of the form:*

$$E_\theta(\mathbf{z}) = A(X^{\mathbf{z}}(0,.)) + L(\mathbf{z}),$$

where:

$$\begin{aligned} A(X^{\mathbf{z}}(0,.)) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} \|y_i - I_0((\phi_1^{z_i})^{-1})\|^2 \\ L(\mathbf{z}) &= \sum_{i=1}^{n} (\boldsymbol{\alpha}_0^i)^t \Gamma_g^{-1} \boldsymbol{\alpha}_0^i \end{aligned} \quad . \tag{20}$$

$$\begin{aligned} &\dot{\mathbf{z}}(t) = F(\mathbf{z}(t)) && \mathbf{z}(0) = \mathbf{z} \\ &\dot{X}^{\mathbf{z}}(t,.) = G(X^{\mathbf{z}}(t,.), \mathbf{z}(t)) && X^{\mathbf{z}}(1,.) = Id_{L^2(D)} \end{aligned}$$

and $X^{\mathbf{z}}(t,.) \in L^2(D,\mathbb{R}^d)$ *for all* t *and* A *is a differentiable map from* $L^2(D,\mathbb{R}^d)$ *to* $\mathbb{R}$.

Then, the gradient of E_θ *is given by:*

$$\nabla_{\mathbf{z}} E_\theta = \xi(0) + \nabla_{\mathbf{z}} L, \tag{21}$$

where two auxiliary variables $\xi(t)$ *(in* $\mathbb{R}^{(n+1)dk_g}$*) and* $\eta(t,.)$ *(in* $L^2(D,\mathbb{R}^d)$*) satisfy the following linear ODEs:*

$$\begin{cases} \dot{\eta}(t,.) = -(\partial_1 G(X^{\mathbf{z}}(t,.), \mathbf{z}(t)))^* \eta(t,.) \\ \eta(0,.) = -\nabla_{X^{\mathbf{z}}(0,.)} A\,, \end{cases} \tag{22}$$

$$\begin{cases} \dot{\xi}(t) = -\partial_2 G(X^{\mathbf{z}}(t,.), \mathbf{z}(t))^* \eta(t,.) - d_{\mathbf{z}(t)} F^t \xi(t) \\ \xi(1) = 0\,, \end{cases} \tag{23}$$

where * *denotes the adjoint operator in* $L^2(D,\mathbb{R}^d)$.

This proposition states that the gradient is computed by integrating two linear ODEs that couple the information in the initial momenta and in the initial control points. Computing the gradient only with respect to the initial momenta does not decrease the computation time. The coupling implies that the gradient with respect to each coordinate of the hidden variables are computed simultaneously. The expression in coordinates of the terms in Proposition 1 as well as its proof can be found in (16).

Remark 5 *The gradient of E_θ is indeed equal to $-\nabla_{\mathbf{z}} \log \pi_\theta$. This vector belongs to the tangent space of the Riemannian manifold of the data at the template point. Thus it provides crucial information about the local shape of this manifold. Therefore, it is of great interest to include this quantity into the estimation algorithm. It has been done from a deterministic point of view in (16) through a gradient descent on E_θ. Nevertheless, this algorithm may be stuck into local minima of this energy E_θ. To avoid such behaviors, stochastic algorithms are well-known powerful tools. In particular, the AMALA-SAEM combines both advantages: it is a stochastic algorithm whose samples are based on this gradient direction.*

Remark 6 *The AMALA-SAEM requires the computation of this gradient at each iteration of the algorithm. Its numerical cost has to be compared with the cost when using other samplers. The hybrid Gibbs sampler that was chosen in (4) for the linearized deformation may also be used here. Although it does not require to compute the gradient, it needs to loop over each coordinate of the hidden variables. This loop in our generalized LDDMM model would involve an integration of the Hamiltonian system* (6) *and the flow equation* (3) *for each coordinate. This scheme would be particularly inefficient due to the coupling between all control points and initial momenta in these equations: one would need to compute the full set of coupled equations each time one updates a coordinate. For k_g control points in dimension d and n observations, the hidden variable is of dimension $(n+1)dk_g$. The Gibbs sampler needs then to integrate $(n+1)dk_g$ times the Hamiltonian system* (6) *and the flow equation* (3) *which are differential systems in dimension $(n+1)dk_g$ and $d|\Lambda|$. The AMALA only requires one single of this step and the gradient computation which involves two differential equations* (22) *and* (23) *both in dimension $(n+1)dk_g$. Although the differential equations in the gradient are more complex than those in the Hamiltonian system (as they require the Hessian of the kernel and not only its gradient), the AMALA sampler is still much more efficient in very high dimension than the Gibbs sampler.*

4.2 Stochastic approximation, projection and maximization steps of the algorithm

We detail here the other steps of the algorithm. The generalized large deformation BME Template model belongs also to the curved exponential family. Indeed, the log-likelihood writes:

$$\begin{aligned}
\log q(\mathbf{y}, \mathbf{c}_0, \boldsymbol{\alpha}_0, \theta) &= \sum_{i=1}^{n} \left(-\tfrac{|\Lambda|}{2} \log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} \|y_i - \mathbf{K}_{\mathbf{p}}^{(\mathbf{c_0}, \boldsymbol{\alpha}_\mathbf{0}^\mathbf{i})} \mathrm{w}\|^2 \right) \\
&+ \sum_{i=1}^{n} \left(-\tfrac{dk_g}{2} \log(2\pi) - \tfrac{1}{2} \log(|\det(\Gamma_g)|) - \tfrac{1}{2} (\boldsymbol{\alpha}_0^i)^t \Gamma_g^{-1} \boldsymbol{\alpha}_0^i \right) - \tfrac{1}{2a_c} \|\mathbf{c}_0 - \bar{\boldsymbol{c}}_\mathbf{0}\|^2 \\
&+ a_g \left(-\tfrac{1}{2} \langle \Gamma_g^{-1}, \Sigma_g \rangle_F - \tfrac{1}{2} \log(|\det(\Gamma_g)|) \right) + a_p \left(-\tfrac{\sigma_0^2}{2\sigma^2} - \tfrac{1}{2} \log(\sigma^2) \right) \\
&- \tfrac{1}{2} \mathrm{w}^t \Sigma_p^{-1} \mathrm{w} - \tfrac{1}{2} (\bar{\boldsymbol{c}}_\mathbf{0} - \mu_c)^t \Sigma_c^{-1} (\bar{\boldsymbol{c}}_\mathbf{0} - \mu_c) + \ \mathrm{C}\,,
\end{aligned} \tag{24}$$

where $\forall u \in \Lambda, \quad \mathbf{K}_{\mathbf{p}}^{(\mathbf{c_0}, \boldsymbol{\alpha}_\mathbf{0}^\mathbf{i})} \mathrm{w}(r_u) = \sum_{k=1}^{k_p} K_p((\phi_1^{\mathbf{c}_0, \boldsymbol{\alpha}_0^i})^{-1}(r_u), b_k) \mathrm{w}_k.$

This enables to exhibit the following sufficient statistics:

$$\left\{\begin{array}{rcl} S_0(\mathbf{z}) & = & \mathbf{c_0} \\ S_1(\mathbf{z}) & = & \sum\limits_{1\leq i\leq n} \left(\mathbf{K_p^{(c_0,\alpha_0^i)}}\right)^t y_i \\ S_2(\mathbf{z}) & = & \sum\limits_{1\leq i\leq n} \left(\mathbf{K_p^{(c_0,\alpha_0^i)}}\right)^t \left(\mathbf{K_p^{(c_0,\alpha_0^i)}}\right) \\ S_3(\mathbf{z}) & = & \sum\limits_{1\leq i\leq n} (\boldsymbol{\alpha}_0^i)^t \boldsymbol{\alpha}_0^i \quad . \end{array}\right. \tag{25}$$

For simplicity, we denote $S(\mathbf{z}) = (S_0(\mathbf{z}), S_1(\mathbf{z}), S_2(\mathbf{z}), S_3(\mathbf{z}))$ for any $\mathbf{z} = (\mathbf{c}_0, \boldsymbol{\alpha}_0) \in \mathbb{R}^{(n+1)dk_g}$. We define the sufficient statistic space

$$\mathcal{S} = \left\{(s_0, s_1, s_2, s_3) | s_0 \in \mathbb{R}^{dk_g},\ s_1 \in \mathbb{R}^{k_p},\ s_2 + \sigma_0^2\Sigma_p^{-1} \in \Sigma^+_{k_p,*}(\mathbb{R}),\quad s_3 + a_g\Sigma_g \in \Sigma^+_{2k_g,*}(\mathbb{R})\right\} .$$

Identifying in the sequel s_2 and s_3 with their lower triangular part, the set S can be viewed as an open convex set of $\mathbb{R}^{n_s}$ with $n_s = dk_g + k_p + \frac{k_p(k_p+1)}{2} + k_g(2k_g+1)$.

Moreover, there exists a maximizing function $\hat{\theta}$ satisfying

$$\forall \theta \in \Theta, \forall s \in \mathcal{S}, L(s; \hat{\theta}(s)) \geq L(s; \theta)\,, \tag{26}$$

which yields:

$$\left\{\begin{array}{rcl} \hat{\mathrm{w}}(s) & = & \left(s_2 + \hat{\sigma}(s)^2(\Sigma_p)^{-1}\right)^{-1} s_1 \\ \hat{\sigma}^2(s) & = & \frac{1}{n|\Lambda|+a_p}\left(\|\mathbf{y}\|^2 + \hat{\mathrm{w}}(s)^t s_2 \hat{\mathrm{w}}(s) - 2\hat{\mathrm{w}}(s)^t s_1 + a_p\sigma_0^2\right)\,, \\ \hat{\Gamma}_g(s) & = & \frac{1}{n+a_g}(s_3 + a_g\Sigma_g)\,, \\ \hat{\tilde{\boldsymbol{c}}}_{\mathbf{0}}(s) & = & s_0\,. \end{array}\right. \tag{27}$$

Since we are not dealing with hidden variables with compact support, we introduce a usual projection of the sufficient statistics on random boundaries. 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 $\mathcal{K}_q \subset \mathrm{int}(\mathcal{K}_{q+1}), \forall q \geq 0$. Let $(\varepsilon_k)_{k\geq 0}$ be a monotone non-increasing sequence of positive numbers and $\mathbb{R}^{n_z}$ where $n_z = (n+1)dk_g$ a subset of $\mathbb{R}^{n_s}$. We construct the sequence $((s_k, \mathbf{z}_k))_{k\geq 0}$ as follows. As long as the stochastic approximation does not fall out the current compact set and is not too far from its previous value, we run the AMALA-SAEM algorithm. If one of the two previous conditions is no more satisfied, we reinitialize the sequences of s and $\mathbf{z}$ using a projection $(\tilde{s}, \tilde{\mathbf{z}}) \in \mathcal{K}_0 \times \mathrm{K}$ (for more details see (6)).

We are now able to summarize the complete estimation algorithm (see Algorithm 1).

4.3 Discussion on theoretical properties

The AMALA-SAEM algorithm has already been applied to the BME Template model in the context of linearized deformations (see (2)). In that paper, the almost

Algorithm 1 AMALA-SAEM with truncation on random boundaries

Set $\kappa_0 = 0$, $s_0 \in \mathcal{K}_0$ and $\mathbf{z}_0 \in \mathrm{K}$.
for all $k \geq 1$ **do**
Sample $\bar{\mathbf{z}}$ from the AMALA transition kernel :
Sample $\mathbf{z}_c \sim \mathcal{N}(\mathbf{z}_k + \delta D_{\theta_k}(\mathbf{z}_k), \delta\Sigma_{\theta_k}(\mathbf{z}_k))$ where

$$D_\theta(\mathbf{z}) = \frac{b}{\max(b, |\nabla \log \pi_\theta(\mathbf{z})|)} \nabla \log \pi_\theta(\mathbf{z}) \,,$$

and

$$\Sigma_\theta(\mathbf{z}) = \varepsilon Id_{(n+1)dk_g} + diag\left(D^0_\theta(\mathbf{z})D^0_\theta(\mathbf{z})^t, D^1_\theta(\mathbf{z})D^1_\theta(\mathbf{z})^t, ..., D^n_\theta(\mathbf{z})D^n_\theta(\mathbf{z})^t\right) \,,$$

with $\varepsilon > 0$ a small regularization parameter and $\delta > 0$.
Then set $\bar{\mathbf{z}} = \mathbf{z}_c$ with probability $\rho_{\theta_k}(\mathbf{z}_k, \mathbf{z}_c)$ and $\bar{\mathbf{z}} = \mathbf{z}_k$ otherwise, where ρ_θ is given in Equation (15).

Compute $\bar{s} = s_{k-1} + \Delta_{k-1}(S(\bar{\mathbf{z}}) - s_{k-1})$ where $(\Delta_k)_{k\in\mathbb{N}}$ is a decreasing positive step size sequence.
if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ and $|\bar{s} - s_{k-1}| \leq \varepsilon_{k-1}$ **then**
set $(s_k, \mathbf{z}_k) = (\bar{s}, \bar{\mathbf{z}})$ and $\kappa_k = \kappa_{k-1}$,
else
set $(s_k, \mathbf{z}_k) = (\tilde{s}, \tilde{\mathbf{z}}) \in \mathcal{K}_0 \times \mathrm{K}$ and $\kappa_k = \kappa_{k-1} + 1$,
where $(\tilde{s}, \tilde{\mathbf{z}}) \in \mathcal{K}_0 \times \mathrm{K}$.
end if
$\theta_k = \hat{\theta}(s_k)$
end for

sure convergence of the parameter sequence as well as its asymptotic normality (Theorems 1 and 2 in (2)) have been proven under usual assumptions on both the model and the step size sequences. Thus, we can wonder whether our LDDMM BME Template model fits into these assumptions. First of all, we notice that our model belongs to the curved exponential family. Moreover, it satisfies the regularity and integrability conditions required in assumptions (M1-M6) and (B2) of Theorem 1 in (2). However, due to the very complex dependencies of the LDDMM model, the super-exponential property (B1) of the conditional density and, related to it, its polynomial upper bound (M8) cannot be guaranteed. Nevertheless, both assumptions sound reasonable in the applications that we are targeting. In the following experiments, the convergence of the algorithm is demonstrated, thus corroborating our hypothesis.

5 Extension toward sparse representation of the geometric variability

Obviously, the number of degrees of freedom needed to describe the variability of a given shape should be adapted to this shape. Therefore, the number of control points in our model should be estimated as a parameter of the model and not fixed by the user. This leads to automatically optimize the dimension of the deformation model. We propose here to simultaneously optimize the positions of the control points and select a subset of the most relevant ones for the description of the variability.

In (16), the control point selection is done adding an L^1 penalty on the momenta to the energy E_θ and performing an adapted gradient descent called FISTA (see (9)). The effect of this penalty is to zero out momenta of small magnitude and to slightly decrease the magnitude of the other ones. A control point which does not contribute to *at least* one of the template-to-observation deformations at the convergence of the algorithm is called inactive. Note that since control points move in the domain, inactive control points may become active during the optimization process, and vice-versa.

This method suffers from three main limitations. First, the Laplace prior associated to the L^1 penalty does not generate sparse observations. Second, the method keeps active control points that may contribute to only few template-to-observation deformations. Lastly, L^1 penalty implies a soft thresholding step on the momentum vectors, thus reducing the norm of these vectors keeping the direction and therefore the local curvature. As a consequence, important momenta for the description of the variability will also be penalized. In the following, we propose to select control points given their importance to describe the variability of the *whole* population, and not of outliers. The idea is to inactivate a control point if the distribution of the momenta attached to it is not strongly correlated with the momentum distribution of other control points. Therefore our procedure selects control point positions and their number, relevant with regards to the whole population.

This constraint on the momenta is taken into account in the model by assuming that the geometric covariance matrix Γ_g is of the form $\Gamma_g = A_g + \varepsilon_g Id$, where ε_g

is a small positive real number and A_g is a sparse symmetric positive matrix. To construct A_g, we modify the third update in Equation (27). Let c_k^g be one of the control points. We compute the sum of the Frobenius norms of the sub-matrices of the sufficient statistic s_3 given by the stochastic approximation of the empirical covariance of this control point with all others:

$$t_k = \sum_{j=1}^{k_g} \|s_3(c_k^g, c_j^g)\|_F \,, \tag{28}$$

where $s_3(c_k^g, c_j^g)$ is the $d \times d$ sub-matrix of s_3 corresponding to the control points c_k^g and c_j^g. Let us fix a positive threshold λ. The control point c_k^g is said active if:

$$t_k \geq \lambda \,. \tag{29}$$

Let us denote $\mathcal{A}$ the set of all active points. Then, we define the sparse matrix A_g as follows:

$$\forall (k,j) \in \{1, ...k_g\} \quad A_g(c_k^g, c_j^g) = s_3(c_k^g, c_j^g) \mathbb{1}_{c_k \in \mathcal{A}} \mathbb{1}_{c_j \in \mathcal{A}} \,. \tag{30}$$

By analogy with Equation (27), the matrix Γ_g is updated as follows:

$$\Gamma_g = \frac{1}{n + a_g} (A_g + a_g Id_{dk_g}) \,, \tag{31}$$

which also corresponds to introduce a specific prior on Γ_g.

This update is performed at each iteration of the estimation algorithm in the M-step.

The threshold λ in our approach plays an equivalent role as the weight of the L^1 penalty in the criterion optimized in (16). The larger, the sparser the solution.

In order to be self-adapted to the data, it could be a benefit to fix the threshold λ as a ratio of the maximum correlations between control points instead of setting a fixed value as in Equation (29). Thus, a control point c_k^g is now active if

$$t_k \geq \lambda \max_{1 \leq j \leq k_g} t_j \,. \tag{32}$$

Moreover, starting to threshold before the Markov chain reaches its stationarity can lead to poor covering of the target distribution support. Therefore, in practice, we start the threshold process after the burn-in period of the estimation algorithm.

To go one step further, we propose to automatically select an optimal threshold λ. We consider a criterion based on two relevant quantities namely the data attachment residual over the training images (denoted by A in Section 4.1.2) and the number of active control points. Indeed, the larger the threshold, the larger the residual and the lower the number of active control points. These quantities are computed for different values of the threshold. These sequences are then normalized to 1. The optimal threshold is chosen to be the point where the two normalized sequences intersect.

6 Experiments

6.1 Handwritten digit experiments

Our first experiments are run on the USPS handwritten digit database which is a traditional benchmark for quantitative performance evaluation of template estimation. Twenty images of each digit are used as the training sample which is presented in Figure 1. This sample shows a large geometric and photometric variability. We consider the model with random control points presented in Equation (10) as well as its simplified version where the control points are fixed. The number of control points is chosen equal to 4, 9 or 16 depending on the experiments. We infer the atlas of each digit independently using our stochastic estimation algorithm for the two models.

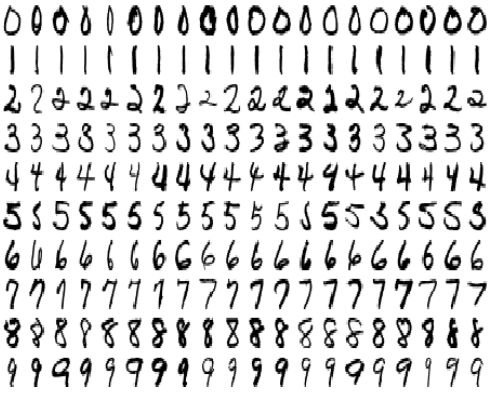

Figure 1: Training set of the USPS database (20 images per digit - inverse video)

We present the estimated templates obtained with both models and varying number of control points in Figure 2. The first row shows the template images estimated with control points fixed. The second one provides the estimated templates together with the estimated control point positions.

As expected, the contours in the template image become sharper in both cases as the number of control points is increased. Moreover, the number of control points being fixed, the sharpness of the estimated template is improved by allowing the control points to move toward optimized positions. We can also note that the estimated control points are informative as they tend to move toward the contours of the digits, and in particular toward those that correspond to the regions of highest variability among samples. It is particularly noticeable on digits 5 and 6 for example.

Note that we checked empirically the identifiability of the control point positions by running several times the same experiment with different random initializations.

We evaluate the relevance of the estimated covariance matrix via the generation of synthetic samples. In Figure 3, we compare the geometry captured with 9 control points using fixed (top) and estimated (bottom) control point models. Although the template of the digit 6 looks similar in both cases, this experiment shows that the geometric variability captured by the model is rather different. The model with equidistributed fixed control points generates unrealistic shape of the digit 6 and

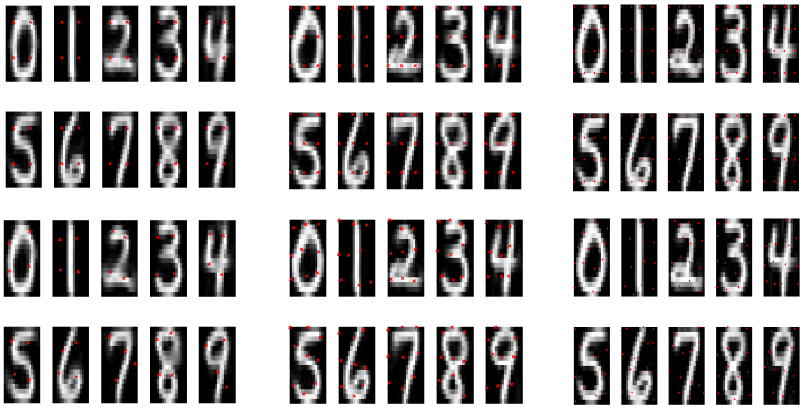

Figure 2: Estimated templates with varying numbers of control points: 4 (left), 9 (middle) and 16 (right), with either fixed (top) or estimated (bottom) control points positions.

therefore does not reflect well the geometric variability observed in the training set. Optimizing for control point positions enables to retrieve a much more natural geometric variability. This optimization increases the number of hidden variables to sample, although the dimension of the covariance matrix remains the same, namely $dk_g \times dk_g$. Updates in control point positions optimize the sub-group of diffeomorphisms of fixed dimension that is the most adapted to describe the variability of a given data set.

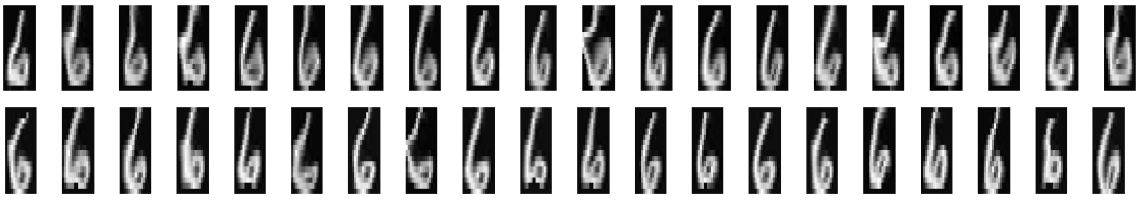

Figure 3: Synthetic samples from the generative model with either fixed (top) or estimated (bottom) control point positions.

In Figure 4, we compare the geometry captured with 4 (top) and 9 (bottom) estimated control points for digit 8. As shown in Figure 2, the contours of the template image with only 4 control points is less sharp particularly in its upper part as the one with 9 control points. For the 4 point model, there are only two close control points in the lower part of the shape, whereas there are three of them spread around the loop with 9 control points. These additional degrees of freedom make the deformation model more flexible as highlighted in Figure 4. Not only the template looks better with an increasing number of control points but the captured geometric variability is also improved.

Nonetheless, one can notice that beyond a certain number of control points, the improvement is less obvious (see Figure 2). This suggests that there may be an *intrinsic dimension* of the deformation space that is optimal (neither too small nor too redundant) for the description of the variability of a given data set. This is also highlighted by the following classification experiment. To perform the classification,

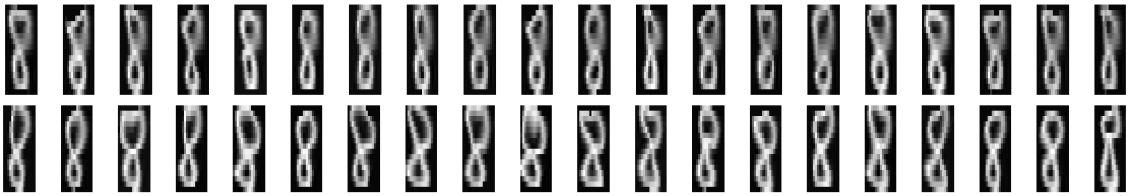

Figure 4: Synthetic samples from the generative model with estimated control point positions. Top: 4, bottom: 9 control points.

we use the test set available in the USPS data base. It contains 2007 digit images. The allocated class for a test image is calculated as follows: we approximate the posterior distribution $q(c|y;\theta_c)$ of the class c given this image y using

$$q(c|y;\theta_c) \simeq Cq(y|(\boldsymbol{\alpha}_0^*)_c, (\bar{\boldsymbol{c}}_{\mathbf{0}})_c, \theta_c)q((\boldsymbol{\alpha}_0^*)_c|\theta_c)q((\bar{\boldsymbol{c}}_{\mathbf{0}})_c|\theta_c)q(\theta_c)\,, \tag{33}$$

where $(\bar{\boldsymbol{c}}_{\mathbf{0}})_c$ is the estimated vector of control points for the class c and

$$(\boldsymbol{\alpha}_0^*)_c = \operatorname*{argmax}_{\boldsymbol{\alpha}_0} q(\boldsymbol{\alpha}_0|y, (\bar{\boldsymbol{c}}_{\mathbf{0}})_c, \theta_c)\,.$$

This approximation has already been used in (1).

Classification results are presented in Figure 5 for 4, 9, 16, 36 and 64 control points using both fixed (blue) and random (red) control point models. The scale of the Gaussian interpolation kernel K_g is fixed such that considering 36 control points leads to one point every kernel scale.

With no control point, the model classifies according to the L^2 similarity with the grey level average image. This mean image, though very fuzzy, is still informative and leads to a classification score of about 85%. If the number of control points is increased, the model incorporates deformations. The template images become less fuzzy (see Figure 2), deformations explain part of the shape variability in an interpretable way (see Figure 3) and the classification scores increase (see Figure 5). Near the maximal classification score, models with estimated control points perform better. As already noted, the slight increase in classification score goes with a much more realistic and interpretable representation of the variability (see Figure 3 and Figure 4). If the number of control points is drastically increased, overall classification scores drop down, as we fall typically in an overfitting situation. Allowing control point positions to be optimized further increases the dimension of parameters. In this case, the deformation model becomes so flexible that it can accommodate for any small differences in shapes, and does not generalize well.

The best performances are reached for in between numbers of control points. In this region, estimating the positions of these control points allows to reach higher classification scores. This confirms the idea of the existence of an intrinsic dimension of the deformation space. How to find such dimension is the purpose of the sparse extended model presented in Section 5 and experimented in Section 6.3.

6.2 Mouse mandible experiment

We consider a second training set composed of 36 X-ray scans of mouse mandibles. Five of them are presented in Figure 6. The estimated template images resulting

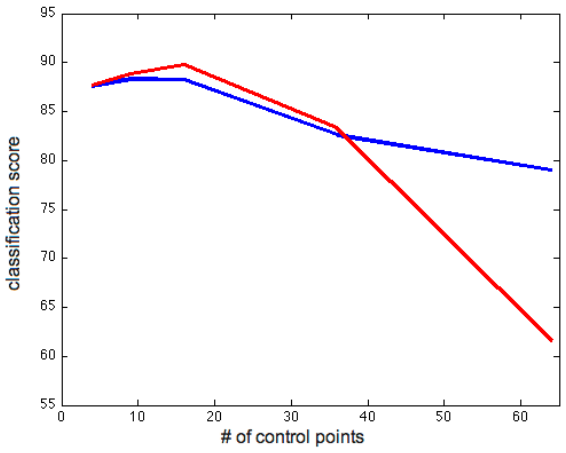

Figure 5: Evolution of the classification score for varying numbers of control points either fixed (blue) or estimated (red).

from three different experiments are shown in Figure 7. The image on the left shows the template estimated using 260 fixed equidistributed control points. The image on the middle (resp. right) shows the estimated template using 117 (resp. 70) estimated control points. These templates look similar, thus showing that the same photometric invariants have been captured in each experiment. These invariants include the main bones of the mandibles (i.e. the brightest areas in the image). The decrease in number of control points is balanced by the optimization of their optimal positions. Control points in the right image are noticeably located on the edges of the shape in order to drive the dilation, contraction and opening of the mandible. Depending on the desired precision of the atlas, we can reduce even more the number of control points. This enables a faster estimation task at the cost of providing less information about the data.

Figure 6: Five training images from the mouse mandibles.

6.3 Toward sparse representation

We test our estimation procedure for the sparse constrained model on a toy example in order to exhibit the stability of the estimated geometry with respect to outliers. We create a data set of 20 images which are composed of vertical translations and vertical dilations of a given rectangle. An outlier image is then introduced into this data base. This outlier has an excrescence on its left border (see Figure 8). We run

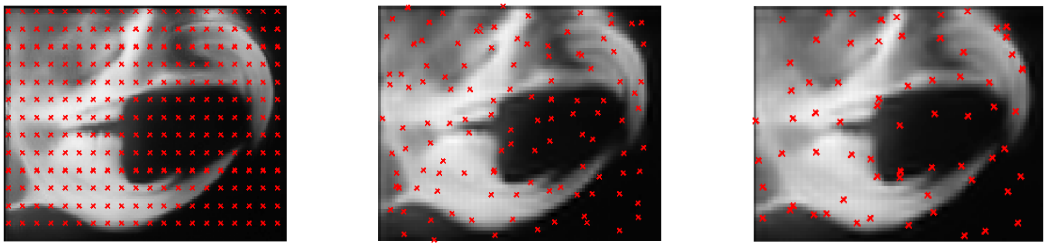

Figure 7: Estimated templates of the mouse mandible images obtained with 260 fixed control points (left), with 117 (middle) and 70 (right) estimated control points.

first our algorithm with the 20 regular training images. Then we run our algorithm and the one proposed in (16) with these 20 images together with the outlier. The three estimated templates with their respective optimized control point positions are shown in Figure 9. The grey level differences are negligible. Runing our estimation procedure for the training images with or without the outlier leads to very similar estimate of the control point positions, thus showing the robustness of our estimate with respect to outliers. Samples generated from both estimated models are presented in Figure 10. They show only vertical deformations up to the isotropy of the Gaussian interpolation kernel. This confirms the ability of the threshold process to limit the effect of the outlier in the data set. By contrast, the method proposed in (16) run on the same data set including the outlier exhibits a very different result. On the left border of the shape, two control points, with momentum magnitude of the same order as the other ones, play an important role in this model although only explaining the variability of the outlier.

Figure 8: Synthetic training sample: 10 examplars among the 20 regular base and the outlier on the right end.

We run the estimation algorithm presented above with the extension described in Section 5 and the threshold rule (32) on the USPS database presented above. We conduct different experiments with different thresholds λ between 0.3 and 0.8 in order to see the evolution of the sparsity with respect to this parameter and also to capture the most interesting one (depending on the training digit). The initial number of control points is set to 16. The results of these experiments are presented in Figure 11 and Figure 12.

As expected, increasing the threshold λ decreases the final number of selected control points, whose effects on template sharpness and description of variability have been presented in Figure 2, Figure 3 and Figure 4. Using the modified prior given in Equation (31) to enforce sparsity allows to automatically select a subset of

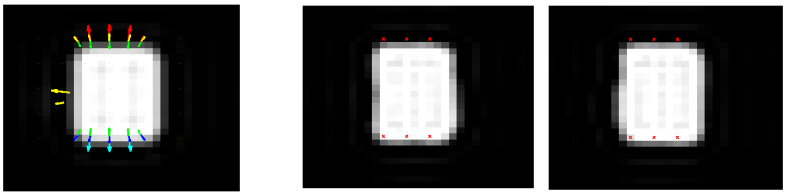

Figure 9: Estimated templates. Left: template estimated with the deterministic gradient descent of (16) for the data set including the outlier. The colored arrows represent the initial momenta for different subjects (in different colors) which are given as output of the algorithm. Middle and right: templates estimated with the stochastic algorithm for the 20 regular images (middle) and the 20 regular images plus the outlier (right).

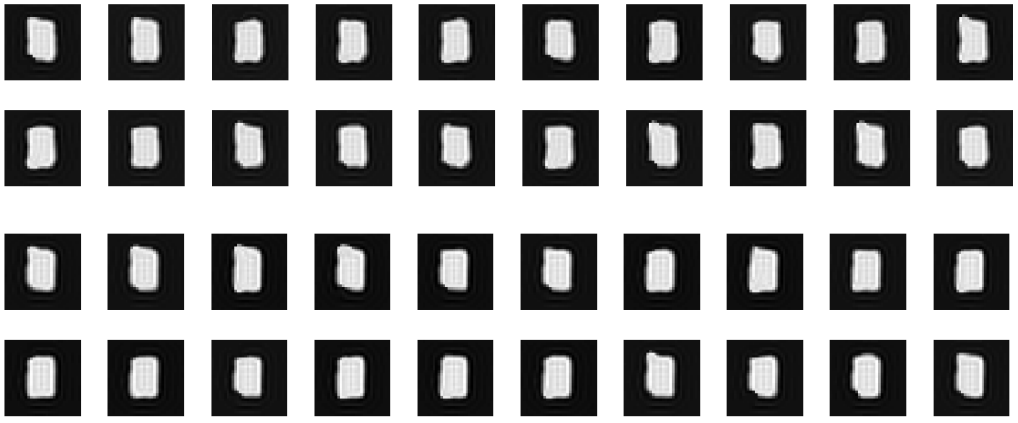

Figure 10: Estimated geometry: samples generated from the 20 regular image estimated model (top) and from the 20 regular images plus the outlier estimated model (bottom).

control points leading to estimation results of the same accuracy (see Figure 11 and Figure 12). Contrary to the L^1 prior used in (16), our sparsity prior selects a small number of control points without penalizing the magnitude of the momenta. Hence the variability of the model is not under-estimated. In this respect, our thresholding process has an effect which is closer to the expected L^0 norm than its surrogate L^1 norm.

Independently of the threshold λ, control points move in areas where the shape is the most variable. This can be noticed in the loop of the digit 2 which is highly variable, especially in contrast to the loop of the digit 6 which is much more stable in shape across observations. This can be seen by a fastest decrease in number of control points when the threshold λ is increased for the digit 6 compared to digit 2. It is also interesting to notice how our model deals with a mixture of 2 that could be written with or without a loop. Such variability violates the hypothesis of our model, which assumes that observations derive from a *diffeomorphic* deformations of the template image. In this situation, the model estimates a template image that is fuzzy in the region of the loop: the non-diffeomorphic variability has been interpreted as a photometric variation. To overcome this problem, one may investigate the use of several template images in the atlas along the lines of (3).

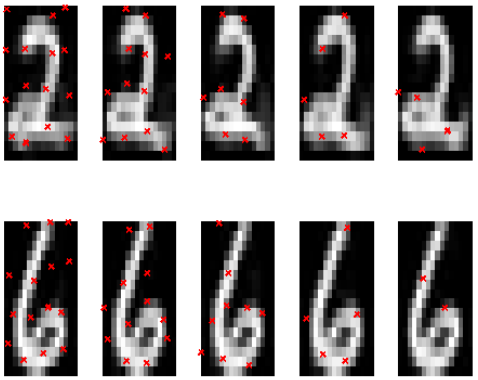

Figure 11: Evolution of the estimated templates and of their number of active control points with respect to the threshold parameter. From left to right: λ equals to $0.3, 0.45, 0.6, 0.75$ and 0.8.

The optimal threshold is chosen applying the criterion described in Section 5. Figure 13 shows the estimated templates with their control points corresponding to the optimal threshold. The number of control points reflects the variability of the digits. In particular, very constrained shapes (see digits 1 and 9) require fewer control points than very complex irregular forms (see digits 3 and 8). Note that in most of the case (nine digits among ten) the selection criterion enables to select thresholds between 50 and 60%.

Figure 12: Synthetic samples of digit 2 from the generative model using the estimated parameters for thresholds 0.3 (top), 0.6 (middle) and 0.8 (bottom).

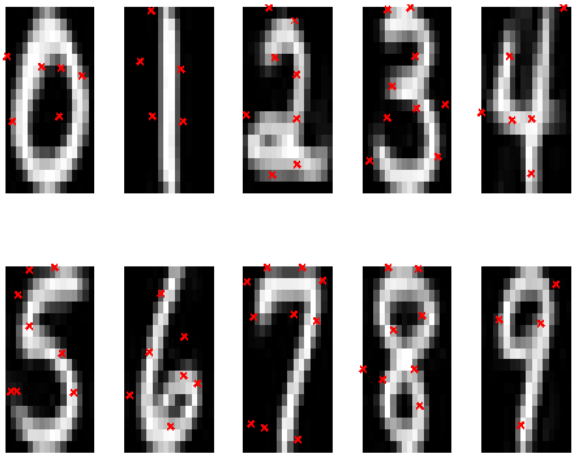

Figure 13: Estimated templates with their optimal numbers and positions of control points.

7 Conclusion and perspectives

In this paper, we presented a generalization of the BME Template model where a diffeomorphic constraint has been added on the deformations. Moreover, a finite dimensional parametrization of these deformations via control points has been used which enabled to include the positions of these points as parameters of the model. The AMALA-SAEM algorithm appears to be particularly well suited to estimate the parameters of such models, especially in comparison with the Gibbs-SAEM. This opened up the possibility to root the usual atlas estimation using large diffeomorphic deformations into a rigorous statistical framework and to propose tractable stochastic algorithm to estimate its parameters. The results on both handwritten digits and mouse mandibles show the interest in such model. Moreover, the issue of the optimal number of control points has been addressed including a selection step to only keep the most informative points. We proposed an empirical criterion to optimize the threshold leading to a model selection. This choice is usually done by cross-validation which requires a large training data base and is computationally costly.

A natural extension is to consider the mixture model introduced in (3) using the same LDDMM formulation with estimated control points. This model can easily be proposed, however, the difficulty stands in adapting the stochastic algorithm with the AMALA sampler so that it remains tractable.

Another remark concerns the choice of a Gaussian distribution for the momenta. In the linearized deformation model, this looks reasonable; indeed, in a flat manifold, the mean of the deformation (and thus the equivalent of the momentum α) starting from the ideal template to all the data should be close to 0. The global behaviors can be well approximated by some Gaussian behavior. But as soon as you consider large deformations, you are no longer in a flat manifold and the curvature has to be taken into account. In this manifold, matching one point (the template) to two close points (two different targets) will not necessarily imply that the two corresponding momenta are close to each other. This leads up to think that the Gaussian model should be changed to some other which will take the curvature of points (images) into account.

Finally, looking at the analytical expression of the observed log likelihood, we recognize the terms coming from the Gaussian distributions on the observations and on the initial momenta as the two terms of a LDDMM registration energy (resp. data attachment and L^2 penalty terms). For this reason, it would be coherent with this setting to use the metric Γ_g both into the penalty term and the definition of the velocity field using interpolation matrix K_g. One further interest of this generalization will be to include a correlation between these two matrices which is not straightforward.

References

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

[2] S. Allassonnière and E. Kuhn, *Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation*, submitted.

[3] S. Allassonnière and E. Kuhn, *Stochastic algorithm for Bayesian mixture effect template estimation*, ESAIM Probab. Stat., 14 (2010), pp. 382–408.

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

[5] S. Allassonnière, A. Trouvé, and L. Younes, *Geodesic shotting and diffeomorphic matching via textured meshes*, in Proc. of the Energy Minimization Methods for Computer Vision and Pattern Recognition (EMMCVPR), A. Y. Anand Rangarajan, Baba Vemuri, ed., November 9-11 2005, pp. 365–381.

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

[7] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[8] F. R. Bach, *Consistency of the group lasso and multiple kernel learning*, J. Mach. Learn. Res., 9 (2008), pp. 1179–1225.

[9] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[10] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, *Computing large deformation metric mappings via geodesic flows of diffeomorphisms*, Int J. Comp. Vis., 61 (2005), pp. 139–157.

[11] J. Bigot, S. Gadat, and J.-M. Loubes, *Statistical M-estimation and consistency in large deformable models for image warping*, J. Math. Imaging Vision, 34 (3) (2009), pp. 270–290.

[12] G. Christensen, R. Rabbitt, and M. I. Miller, *Deformable templates using large deformation kinematics*, IEEE trans. Image Proc., (1996).

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

[14] P. Dupuis, U. Grenander, and M. I. Miller, *Variational problems on flows of diffeomorphisms for image matching*, Quart. Appl. Math., 56 (1998), pp. 587–600.

[15] S. Durrleman, *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution.*, PhD thesis, Ecole Normale Supérieure de Cachan, France, 2010.

[16] S. Durrleman, S. Allassonnière, and S. Joshi, *Sparse adaptive parameterization of variability in image ensembles*, Int. J. Comput. Vis., 101 (2013), pp. 161–183.

[17] S. Durrleman, M. Prastawa, G. Gerig, and S. Joshi, *Optimal data-driven sparse parameterization of diffeomorphisms for population analysis*, in Information Processing in Medical Imaging (IPMI), G. Székely and H. Hahn, eds., vol. 6801 of LNCS, 2011, pp. 123–134.

[18] J. Glaunès, M. Vaillant, and M. I. Miller, *Landmark matching via large deformation diffeomorphisms on the sphere*, J. Math. Imaging Vision, 20 (2004), pp. 179–200. Special issue on mathematics and image analysis.

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

[20] U. Grenander and M. I. Miller, *Computational anatomy: an emerging discipline*, Quart. Appl. Math., 56 (1998), pp. 617–694. Current and future challenges in the applications of mathematics (Providence, RI, 1997).

[21] D. Holm, J. Ratnanather, A. Trouvé, and L. Younes, *Soliton dynamics in computational anatomy*, Neuroimage, 23 (2004), pp. 170–178.

[22] S. Joshi and M. I. Miller, *Landmark matching via large deformation diffeomorphisms*, IEEE transactions in image processing, 9 (2000), pp. 1357–1370.

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

[24] J. Ma, M. I. Miller, A. Trouvé, and L. Younes, *Bayesian template estimation in computational anatomy*, Neuroimage, 42 (1) (2008), pp. 252–261.

[25] S. Marsland and C. Twining, *Constructing diffeomorphic representations for the groupwise analysis of non-rigid registrations of medical images*, IEEE Transactions on Medical Imaging, 23 (2004), pp. 1006–1020.

[26] M. Micheli, P. W. Michor, and D. Mumford, *Sectional curvature in terms of the cometric, with applications to the Riemannian manifolds of landmarks*, SIAM J. Imaging Sci., 5 (2012), pp. 394–433.

[27] M. Miller, C. Priebe, A. Qiu, B. Fischl, A. Kolasny, T. Brown, Y. Park, J. Ratnanather, E. Busa, J. Jovicich, P. Yu, B. Dickerson, and R. Buckner, *Morphometry BIRN. collaborative computational anatomy: An MRI morphometry study of the human brain via diffeomorphic metric mapping*, Human Brain Mapping, 30 (2009), pp. 2132–2141.

[28] M. I. Miller, A. Trouvé, and L. Younes, *On the metrics and Euler-Lagrange equations of computational anatomy*, Annual Review of Biomedical Engineering, 4 (2002), pp. 375–405.

[29] M. I. Miller, A. Trouvé, and L. Younes, *Geodesic shooting for computational anatomy*, Journal of Mathematical Imaging and Vision, 24 (2006), pp. 209–228.

[30] M. I. Miller and L. Younes, *Group action, diffeomorphism and matching: a general framework*, Int. J. Comp. Vis, 41 (2001), pp. 61–84.

[31] X. Pennec, *Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements*, Journal of Mathematical Imaging and Vision, 25 (2006), pp. 127–154.

[32] J.-P. Thirion, *Image matching as a diffusion process: an analogy with maxwell's demons*, Medical Image Analysis, 2 (1998), pp. 243–260.

[33] A. Trouvé, *Diffeomorphism groups and pattern matching in image analysis*, International Journal of Computer Vision, 28 (1998), pp. 213–221.

[34] M. Vaillant, M. I. Miller, A. Trouvé, and L. Younes, *Statistics on diffeomorphisms via tangent space representations*, Neuroimage, 23 (2004), pp. 161–169.

[35] M. Zhang, N. Singh, and P. T. Fletcher, *Bayesian estimation of regularization and atlas building in diffeomorphic image registration*, IPMI, (2013).