

Bayesian consistent estimation in deformable models using stochastic algorithms: Applications to medical images

Stéphanie Allassonnière, Estelle Kuhn, Alain Trouvé

▶ **To cite this version:**

Stéphanie Allassonnière, Estelle Kuhn, Alain Trouvé. Bayesian consistent estimation in deformable models using stochastic algorithms: Applications to medical images. Journal de la Societe Française de Statistique, 2010, 151 (1), pp.1-16. hal-02653610

HAL Id: hal-02653610
https://hal.inrae.fr/hal-02653610

Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal de la Société Française de Statistique
Vol. 151 No. 1 (2010)

Bayesian Consistent Estimation in Deformable Models using Stochastic Algorithms: Applications to Medical Images

Stéphanie Allassonnière [1], Estelle Kuhn [2] and Alain Trouvé [3]

Titre: Estimation Bayésienne Consistante de Modèles Déformables via des Algorithmes Stochastiques : Applications à l'Imagerie Médicale

Abstract: This paper aims at summarising and validating a methodology proposed in [2, 3, 4] for estimating a Bayesian Mixed Effect (BME) atlas, i.e. coupled templates and geometrical metrics for estimated clusters, in a statistically consistent way given a sample of images. We recall the generative statistical model applied to the observations which enables the simultaneous estimation of the clusters, the templates and geometrical variabilities (related to the metric) in the population. Following [2, 3, 4], we work in a Bayesian framework, use a Maximum A Posteriori estimator and approach its value using a stochastic variant of the Expectation Maximisation (EM) algorithm. The method is validated with two data set consisting of medical images of part of the human cortex and dendrite spines from a mouse model of Parkinson's disease. We present the performances of the method on the estimation of the template, the geometrical variability and the clustering.

Résumé : Cet article vise à résumer et valider sur données réelles la méthode proposée dans (2,3,4) pour l'estimation d'atlas appelé Bayesian Mixed Effect (BME) atlas. Un tel atlas est composé d'une image de référence et d'une métrique pour chaque sous-groupe d'une population ainsi que du poids de ce sous-groupe. L'estimation est consistante sur un échantillon d'images données non labellisées. Nous rappelons ici le modèle statistique génératif qui permet l'estimation simultanée des sous-groupes, de leurs poids, des images de référence et des variabilités géométriques (liées aux métriques). Comme proposé en (2,3,4), nous travaillons dans un cadre bayésien, utilisons l'estimateur de Maximum A Posteriori et approchons sa valeur par une variante stochastique de l'algorithme EM (Expectation Maximisation). Cette méthode est validée sur deux ensembles de données d'images médicales : une partie du cortex humain et des excroissances de dendrites de souris liées à la maladies de Parkinson. Nous présentons les performances de cette méthode sur l'estimation de l'image de référence, la variabilité géométrique et le label.

Keywords: generative statistical model, stochastic EM algorithm, MAP estimator, MCMC methods, medical imaging

Mots-clés : modèle statistique génératif, algorithme EM stochastique, estimateur MAP, méthodes MCMC, imagerie médicale

AMS 2000 subject classifications: 60J22, 62F10, 62F15, 62M40

[1] CMAP Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, France.
E-mail: Stephanie.Allassonniere@polytechnique.edu
[2] INRA, Domaine de Vilvert, 78352 Jouy en Josas, France.
E-mail: estelle.kuhn@jouy.inra.fr
[3] CMLA ENS Cachan, 61 av du Pres. Wilson, 94230 Cachan, France.
E-mail: Alain.Trouve@cmla.ens-cachan.fr

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

1. Introduction

In the field of Computational Anatomy, one aims at segmenting images, detecting pathologies and analysing the normal versus abnormal variability of segmented organs. The most widely used techniques are based on the comparisons of images from subjects to a prototype image (usually called template in the literature). Such a prototype is an image whose biological properties are known and which - in a sense to be defined - characterises the population being studied. This template contains common features of the population which would not be revealed by multiple inter-subject comparisons.

Regarding the large variability of anatomical structures, one template only may be not enough to summarise the diversity of a whole population. For example, two populations can have the same template but can be distributed quite differently around (very like points clouds in a manifold can be concentrated or spread in many different way around their means). Therefore, in addition to the template, a parametrisation of the shape variability around a given template is of importance in producing relevant statistical summary of a population. These two parameters will together be considered as an atlas in the following.

One way to estimate an atlas in a population is to do statistical inference on statistical models. Among all of them, generative statistical models make assumptions on how the observed images are derived from the atlas. These models do not only explain data but enable also to randomly generate new images. When simulating a large number of likely images (according to the model), one can better interpret and even exhibit unexpected behaviours that would not be easily detectable by a visual inspection of a small population (typical case in medical image analysis). Moreover, these models provide a characterisation of the population through the probabilistic distributions, in particular, they highlight the correlations between variables. Lastly, this setting provides a good mathematical and computational framework to work with.

The large heterogeneity of the control subjects for example leads us to consider that the population is composed of several sub-groups. We introduce a mixture of the previous models to take this point into account. To summarise the information about the population, one need now the weight of each cluster and an atlas for each of them. Since the clustering may not be known, the corresponding model enables an estimation of both the distribution of the sub-groups in the population and the cluster atlases at the same time.

Our special interest is the **construction** of an **atlas**, called Bayesian Mixed Effect (BME) atlas, as the estimation of the templates and their global geometric variabilities in estimated cluster for a given population in a statistically consistent way.

The usual way to measure the geometrical heterogeneity is to map the template to all the observations (or the other way around) and do some statistics on these deformations (typically PCA). Many registration methods have been developed for this purpose, for example in [21, 16, 7]. Based on this, several different approaches have been proposed recently to estimate templates. Some are based on a minimisation of a penalised energy function describing the cost to match the template to the observations [14, 19, 10]. Another view, closer to ours, is to propose a statistical model whose parameters are the template and the **mappings** between this template and the observations [5, 13, 18, 17] and the optimisation is done via maximum likelihood. Even if these methods lead to interesting results and effective computation schemes, they suffer from different limitations. First, in most cases, the deformation is applied to the observations instead of to the

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

template. However, these images are only noisy observations known on a discrete fixed grid of voxels. Applying the deformation to these discretly supported images requires interpolating between voxels and therefore creates errors which are difficult to control. Moreover, the modelling implies inexact matching. One way to model this is to consider that the difference between the deformed image and the template is an independent additive noise. This noise accounts for both the acquisition noise and the fact that the model does not describe the reality (but is only an approximation of the true distribution, providing that it exists). Assuming the deformation is invertible, applying the mapping to the observations is equivalent to apply its inverse to both the template and the noise. However, there is no suitable interpretation of this fact; there is no reason for the noise to be affected by the mapping which is only a mathematical tool we introduce. The last but not least drawback is that the deformations are considered as nuisance parameters which have to be optimised. Knowledge of these elements only gives information subject by subject and nothing about the global trend of the population. Moreover, the convergence of such procedures has not been proved and one of them has even been shown to fail for a toy example [2].

For these reasons, we consider the model proposed in [2]. Indeed, the authors consider the usual modelling called the Deformable Template model. This assumes that each observation is a random deformation of the template which is then corrupted by an additive Gaussian noise. This avoids the interpolation problem since the template is estimated on the whole domain as well as the lack of meaning of the deformed noise mentioned below. The deformations are *unobserved random* variables whose probabilistic distribution has to be estimated. This generative statistical model defines a global information of the geometrical variability inside the population. This distribution also characterises the metric on the deformation space. Thanks to this model, the estimation of the template is correlated to this estimated metric and vice versa.

To take into account the heterogeneity of the whole population, we use the extended model based on a mixture of the previous modelling [2, 4]: each observation belongs to one component of the mixture governed by its parameters (template, noise and metric). The observation memberships are specified through *hidden random* labels whose weights are estimated as well.

We summarise here this methodology, called Bayesian Mixed Effect (BME) template, to construct a BME atlas, i.e. clusters distribution, templates and geometrical metrics, via a consistent estimation, given a sample of images. We focus on its validation in the context of medical images of the splenium and of dendrite spines which have a large geometrical variability (various shapes) in order to show its performance in terms of estimation and generation of new plausible shapes.

In this paper, the model and the estimator are detailed in Section 2. We then present two algorithms and their properties in Section 3. The method is then illustrated in Section 4. We end this paper with some conclusions and a discussion in Section 5.

2. BME Template Model and MAP Estimation

2.1. BME Template Model

We consider a population of n gray level images which we aim to automaticaly cluster in a linearised number of groups called components later. We assume that each observation y belongs to an *unknown* component t picked among m possible ones. We work within the small deformation framework [5]: the deformation moves each point of the domain in the direction of its own vector.

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

Therefore, conditional on the image membership to component t, there exists an *unobserved* deformation field $z : \mathbb{R}^d \to \mathbb{R}^d$, $d \in \{2,3\}$, of a continuously defined template $I_t : \mathbb{R}^d \to \mathbb{R}$ and a Gaussian centred white noise [1] ε of variance σ_t^2 such that

$$y(s) = I_t(x_s - z(x_s)) + \varepsilon(s) = zI_t(s) + \varepsilon(s)\,, \tag{1}$$

where Λ is a discrete grid of pixels/voxels and the pixels/voxel location in $\mathbb{R}^d$ is denoted by $(x_s)_{s\in\Lambda}$. To apply the inverse deformation to the template, we have made the usual approximation $(Id+z)^{-1} \simeq Id - z$.

The inference will concern the templates and some characteristics of the deformation fields. Thus we choose parametrical models for these two quantities. Given $(p_k)_{1\le k\le k_p}$ a fixed set of uniformly distributed landmarks covering the image domain, the template functions I_t are parameterised by coefficients $\alpha_t \in \mathbb{R}^{k_p}$ through: $I_t(x) = \mathbf{K_p}\alpha_t(x) \triangleq \sum_{k=1}^{k_p} K_p(x,p_k)\alpha_t(k)$, where K_p is the kernel of the Reproducing Kernel Hilbert Space (RKHS) in which we search the template. The kernel controls the smoothness of the interpolation between landmarks. It is also nicely described as the covariance operator of a Gaussian random field globally defined on the image domain and defining a natural prior for the template. The restriction of these Gaussian fields on the p_k's is an easily tractable finite dimensional zero mean Gaussian vector with explicit covariance matrix. This has the advantage of giving a prior that is essentially independent of the number of landmarks k_p, and that only depends on the global choice made for the RKHS. In this context, the number of landmarks used determines a trade-off between accuracy of the approximations of functions in the respective spaces and the amount of required computation.

The same kind of decomposition with a second set of landmarks $(g_k)_{1\le k\le k_g}$ and kernel K_g is used to parametrize the deformation field z by the *unobserved random* vector β such that $z = \mathbf{K_g}\beta$. This random vector is assumed to follow a Gaussian distribution with zero mean and covariance matrix Γ_g^t depending on the component t (which could be the natural prior associated with K_g as a first guess but will be learnt from the data during the estimation process).

The model parameters of each component $t \in \{1,\dots,m\}$ are denoted by $\theta_t = (\alpha_t, \sigma_t^2, \Gamma_g^t)$. We assume that θ belongs to the open parameter space $\Theta \triangleq \{\ \theta = (\alpha_t,\sigma_t^2,\Gamma_g^t)_{1\le t\le m} |\ \forall t \in \{1,\dots,m\}\ ,\ \alpha_t \in \mathbb{R}^{k_p},\ \sigma_t^2 > 0,\ \Gamma_g^t \in \Sigma^+_{dk_g,*}(\mathbb{R})\ \}$. Here $\Sigma^+_{dk_g,*}(\mathbb{R})$ is the set of strictly positive symmetric matrices. The weights of the different mixtures are given by $\rho = (\rho_t)_{1\le t\le m}$ which belongs to the open simplex $\mathscr{R} = \{(\rho_t)_{1\le t\le m} \in\ ]\,0,1[^m\ |\ \sum_{t=1}^m \rho_t = 1\}$. Let $\eta = (\theta,\rho)$, the hierarchical Bayesian structure of our model is :

$$\begin{cases} \rho \sim \nu_\rho, \quad \theta = (\alpha_t,\sigma_t^2,\Gamma_g^t)_{1\le t\le m} \sim \otimes_{t=1}^m (\nu_p \otimes \nu_g) \\ \tau_1^n \sim \otimes_{i=1}^n \sum_{t=1}^m \rho_t \delta_t \mid \rho\,, \\ \beta_1^n \sim \otimes_{i=1}^n \mathscr{N}(0,\Gamma_g^{\tau_i}) \mid \tau_1^n,\ \eta \\ \\ y_1^n \sim \otimes_{i=1}^n \mathscr{N}(z_{\beta_i} I_{\alpha_i}, \sigma^2_{\tau_i} Id_\Lambda) \mid \beta_1^n,\ \tau_1^n,\ \eta \end{cases} \tag{2}$$

[1] This model is relevant for grey level images. One could slightly modify it in order to better interpret binary images. Instead of a Gaussian noise (usually used for image matching with a L^2 penalty term), one can use a Bernoulli distribution whose parameter would be a continuous map $r_t(x)$, analogous to our template $I_t(x)$. However, this model does not belong to the exponential family which make the coding more complicated. The convergence of the algorithm has not been proved in this case either.

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

$$\text{with} \quad \begin{cases} \nu_\rho(\rho) \propto \left(\prod_{t=1}^{m} \rho_t\right)^{a_\rho}, \quad \nu_g(d\Gamma_g) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g \rangle / 2) \frac{1}{\sqrt{|\Sigma_g|}}\right)^{a_g} d\Gamma_g, \\ \nu_p(d\sigma^2, d\alpha) \propto \left(\exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \frac{1}{\sqrt{\sigma^2}}\right)^{a_p} \cdot \exp\left(-\frac{1}{2}\alpha^t (\Sigma_p)^{-1} \alpha\right) d\sigma^2 d\alpha, \end{cases}$$

where the hyper-parameters are fixed (their effects has been discussed in [2]). All priors are the natural conjugate priors and are assumed independent to get easy calculations. This choice appears relevant while considering the equations involved in the maximisation step [2].

The Gaussian distribution set on the observations whose mean is the deformed template is the usual Deformable Template model used in image analysis and in particular image matching. This model is quite natural saying that the observation is, up to an independent noise, close to the deformed template. The Gaussian distribution used to model the deformation vector β is assumed to have zero mean. This assumption corresponds to the intuitive fact that once we are moving around the template -the "mean shape" of the population- the mean of all these movements should be close to zero. Therefore, we only estimate its covariance matrix. The last probabilistic distribution for τ is a common distribution on random variables on finite space, namely a finite sum of weighted Dirac measures.

The system of equations (2) can be interpreted top to bottom, which corresponds to the generation of some images. The generation process consists in first drawing the parameters from their prior distributions. Given these parameters, pick a membership according to the weighted distribution. This label points towards a component. For this particular component, draw a deformation with respect to this Gaussian law and apply it to the pointed template. Adding a random Gaussian noise whose variance is given by the membership to each voxel independently gives you a new image. The estimation process takes the images as observed elements and attempts to recover the parameters (giving that they follow some constrains given by the priors). This scheme can be summarised in Figure 1.

For each component t :
- ρ_t : probability of the component
- α_t : template parameter
- Γ_g^t : geometrical covariance matrix
- σ_t^2 : additive noise variance

For each observation y_i :
- τ_i : component label
- β_i : deformation parameters
- ε_i : additive noise

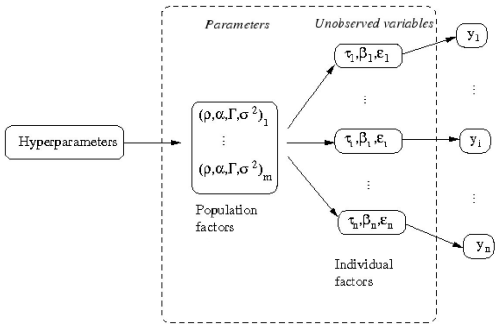


FIGURE 1. *Latent structure of BME-Template model.*

2.2. *MAP estimator and its theoretical properties*

In this context, in order to estimate the model parameters, we use a Maximum A Posteriori estimator, i.e. a value of the parameters which maximises the posterior density on η conditional

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

on y_1^n:

$$\hat{\eta}_n = \underset{\eta}{\operatorname{argmax}}\, q(\eta|y_1^n)\,. \tag{3}$$

It has been proved in [2] that the MAP estimator corresponding to the model reduced to only one component ($m = 1$) exists given a sample set. Moreover, this estimator is consistent i.e. as the number of images increases in the training set, the sequence of estimated parameters converges almost surely towards one maximiser of the expectation of the observed log-likelihood.

3. Algorithmic methods for estimation

Thanks to the Bayes' rule, the focus is on the observed likelihood. Its computation involves an integral over the hidden variables making the direct maximisation a difficult task. The natural approach in this context is to use iterative algorithms such as EM (Expectation-Maximisation) [12, 20] to maximise the penalised likelihood given the observations y_1^n. However, the classical EM algorithm cannot be directly applied here. Indeed, the E-step requires the computation of the conditional expectation of the complete log-likelihood which does not have a closed form. As a consequence, many "EM-like" procedures have been proposed.

We present in the next subsections two approaches to solve this issue: the first one is deterministic whereas the second one is stochastic.

3.1. Fast approximation with modes (FAM)

The expression in the E step requires the computation of the expectation with respect to the posterior distribution of $\beta_1^n, \tau_1^n|y_1^n$, known up to the re-normalisation constant, the density of y_1^n which is not computable. To overcome this obstacle, given an observation y_i and a label t, the posterior distribution of the random deformation field is approximated at iteration l by a Dirac law on its mode $\beta^*_{l,i,t}$. This yields the following computation :

$$\begin{aligned}
\beta^*_{l,i,t} &= \arg\max_{\beta} \log q(\beta|\alpha_{t,l}, \sigma^2_{t,l}, \Gamma^t_{g,l}, y_i) \\
&= \arg\min_{\beta} \left\{ \frac{1}{2}\beta^t (\Gamma^t_{g,l})^{-1}\beta + \frac{1}{2\sigma^2_{l,t}}|y_i - K_p^{\beta}\alpha_{t,l}|^2 \right\},
\end{aligned}$$

which is a standard template matching problem with the current parameters. We then approximate the joint posterior on (β_i, τ_i) as a discrete distribution concentrated at the m points $(\beta^*_{l,i,t})_{1\le t\le m}$ with weights given by: $w_{l,i}(t) \propto q(y_i|\beta^*_{l,i,t}, \alpha_{t,l}) q(\beta^*_{l,i,t}|\Gamma^t_{g,l})\rho_{t,l}$. The label $\tau_{l,i}$ is then sampled from the distribution $\sum_{t=1}^m w_{l,i}(t)\delta_t$ and the deformation is the mode of the drawn label $\beta_{l,i} = \beta^*_{l,i,\tau_i}$.

The maximisation is then done on this approximation of the likelihood.

3.2. Using a stochastic version of the EM algorithm : SAEM-MCMC

Another alternative to the computation of the E-step in a complex nonlinear context is to use the stochastic approximation EM algorithm (SAEM) [11] coupled with an MCMC procedure [15]

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

and a truncation on random boundaries. The combination of these different methods enables to overcome all the bottlenecks that are arising from the usual algorithms. In particular, the SAEM algorithm requires a simulation of the hidden variables with respect to their posterior distribution which in this context is not doable. Our model belongs to the exponential density family which means that:

$$q(y,\beta,\tau,\eta) = \exp\left[-\psi(\eta) + \langle S(\beta,\tau),\phi(\eta)\rangle\right],$$

where the sufficient statistic S is a Borel function on $\mathbb{R}^{dk_g} \times \{1,\dots,m\}$ taking its values in an open subset $\mathscr{S}$ of $\mathbb{R}^m$ and ψ, ϕ two Borel functions on $\Theta \times \rho$ (the dependence on y is omitted for sake of simplicity).

We introduce the following function: $L : \mathscr{S} \times \Theta \times \rho \to \mathbb{R}$ as $L(s;\eta) = -\psi(\eta) + \langle s, \phi(\eta)\rangle$. Direct generalisation of the proof in [2] to the multicomponent model gives the existence of a critical function $\hat{\eta} : \mathscr{S} \to \Theta \times \rho$ which satisfies: $\forall \eta \in \Theta \times \rho, \forall s \in \mathscr{S}, L(s;\hat{\eta}(s)) \geq L(s;\eta)$. Then, iteration l of this algorithm consists of the following four steps.

Simulation step: The missing data are drawn using a transition probability of a convergent Markov chain having the posterior distribution as stationary distribution:

$$(\beta_{l+1},\tau_{l+1}) \sim \Pi_{\eta_l}((\beta_l,\tau_l),\cdot).$$

Stochastic approximation step: Since the model is exponential, the stochastic approximation is done on the sufficient statistics using the simulated values of the missing data:

$$s_{l+1} = s_l + \Delta_{l+1}(S(\beta_{l+1},\tau_{l+1}) - s_l),$$

where $(\Delta_l)_l$ is a decreasing sequence of positive step-sizes.

Truncation step: A truncation is done on the stochastic approximation. Let $(\mathscr{K}_q)_{q\geq 0}$ be an increasing sequence of compact subsets of $\mathscr{S}$ such as $\cup_{q\geq 0}\mathscr{K}_q = \mathscr{S}$ and $\mathscr{K}_q \subset \mathrm{int}(\mathscr{K}_{q+1}), \forall q \geq 0$. If $\bar{s}_{l+1}$ wanders out of $\mathscr{K}_{l+1}$ then the algorithm is reinitialized in a given compact set. Otherwise, set $s_{l+1} = \bar{s}_{l+1}$.

Maximization step: The parameters are updated:

$$\eta_{l+1} = \hat{\eta}(s_{l+1}).$$

We now explain the choice of the transition kernel of the Markov chain Π_η used in the simulation step. As we aim to simulate (β_i,τ_i) through a transition kernel whose stationary distribution is $q(\beta,\tau|y_i,\eta)$, we first simulate τ_i with a kernel whose stationary distribution is $q(\tau|y_i,\eta)$ and then β_i through a transition kernel that has $q(\beta|\tau,y_i,\eta)$ as stationary distribution. Given any initial deformation field $\xi_0 \in \mathbb{R}^{dk_g}$, we run, for each component t, J_l iterations of a hybrid Gibbs sampler (for each coordinate of the vector, a Hasting-Metropolis sampling is done given the other coordinates) $\Pi_{\eta,t}$ using the conditional prior distribution $\beta^j|\beta^{-j}$ as the proposal for the j^{th} coordinate, β^{-j} referring to β without its j^{th} coordinate. So that we get J_l elements $\xi_{t,i} = (\xi_{t,i}^{(k)})_{1\leq k\leq J_l}$ of an ergodic homogeneous Markov chain whose stationary distribution is

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

$q(\cdot|y_i,t,\eta)$. Denoting $\xi_i = (\xi_{t,i})_{1\leq t\leq m}$, we simulate τ_i through the discrete density with weights given by:

$$\hat{q}_{\xi_i}(t|y_i,\eta) \propto \left(\frac{1}{J_l}\sum_{k=1}^{J_l}\left[\frac{f_t(\xi_{t,i}^{(k)})}{q(y_i,\xi_{t,i}^{(k)},t|\eta)}\right]\right)^{-1},$$

where f_t is the density of the Gaussian distribution $\mathcal{N}(0,\Gamma_{g,t})$. Then, we update β_i by re-running J_l times the hybrid Gibbs sampler Π_{η,τ_i} starting from a random initial point β_0.

Remark 1. *Note that when only considering one component in the population ($\tau_m = 1$), a simpler algorithm can be implemented as presented in [3]. Indeed, because of the mixture model, we face the trapping state problem which numerically freezes the labels. To overcome this issue, we introduced the previous more complex algorithm involving parallel Markov Chains to allow more ways out from the trapping states. The one component model only requires a single step of the Markov chain ($J_l = 1$) at each iteration l. The same hybrid Gibbs Sampler is used as transition kernel.*

3.3. Theoretical properties of the stochastic algorithms

It has been proved in [4], that the sequence $(\eta_l)_l$ generated through this algorithm converges a.s. towards a critical point of the penalised likelihood of the observations.

The combination of three different statistical tools -EM algorithm, stochastic approximation and MCMC methods- in a single algorithm led us to assume three usual types of conditions. The convergence assumptions due to the EM algorithm require some regularity of the model. The hypothesis concerning the stochastic approximation focuses on the step-size sequence and on the control of the random perturbation and the residual term. To ensure the convergence of the MCMC method, assumptions similar to the usual Drift conditions are sufficient. We refer to [4] for further details.

Remark 2. *The proof of convergence of the one component algorithm mentioned in Remark 1 has been addressed in [3]. The authors propose a general convergence theorem for stochastic approximations generalising Andrieu et al.'s theorem [6] which is then applied to the BME-Template model.*

3.4. Optimization on the representation, model and algorithms

Despite the fact that many parameters (e.g. the noise variance) are self-calibrated during the estimation process, the algorithm depends on some tuning parameters (coming from the modelling and the stochastic algorithm) and the prior hyper-parameters we would like to discuss briefly.

Data representation issues. The first point to be explained is the effect of the representation of the data, in particular the spline representation of both the template and the deformations. We have chosen Gaussian kernels $K_p(x,x') = \exp(-\frac{1}{2\sigma_p^2}\|x-x'\|^2)$ with $(x,x') \in [-1.5,1.5]^d$ and as well as for K_g with respective scale σ_g and $(x,x') \in [-1,1]^d$. The influence of their two scales can be seen on the template estimation. Indeed, choosing a too small geometric scale leads to very localised deformations around control points and the resulting template is more blurry. On the

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

opposite side, a very large scale induces very smooth deformations which would no longer be relevant for the kind of deformations required to explain the database. Concerning the photometric scale which corresponds to the spreading of the control points through the kernel interpolation, it is straightforward that a large scale will drive to blurry template. In addition, the effects of increasing scale can also be noticed on the learnt covariance matrix. Given a fatty template (larger shapes than expected), the deformations required to fit the database will be forced to contract the template. This phenomena is thus important in the learnt covariance matrix. When we generate new data thanks to the estimated parameters, we can see, that the template is contracted, which is relevant, but also enlarged since the distribution on β is symmetric (this particular point is detailed in the next paragraph). Those large images are not typical from the training set. We refer to [3] for some illustration of this phenomenon on hand-written digit images.

Model distribution issues. One question is the relevance of the Gaussian distribution chosen for the deformation field. It is natural to think that the mean of the deformations around an atlas is close to zero whereas the symmetry of the distribution (the probability of a deformation field + β equals its opposite one $-\beta$) is much more arguable.

Another issue about the model is the choice of the prior hyper-parameters. In particular, the effect of the inverse Wishart prior a_g on the geometric covariance matrix is important. Indeed, if we want to satisfy the theoretical requirements to the algorithms, we have to chose $a_g \geq d^2 k_g + 1$. However, the update formula is a barycenter between the expectation of the empirical covariance matrix and the prior with weights n and a_g respectively (cf: [2]). Since we are working with small sample sizes, this condition makes the update of Γ_g very constrained close to the prior Σ_g. This does not enable the geometry to be well estimated and the effects can be seen directly on the template but also on the classification rate [2]. The value of a_g used in those particular experiments is fixed to 0.5. Concerning the other weights $(a_p,\ a_\rho)$, their effects are less significant on the results and we fixed them to 200 and 2 respectively.

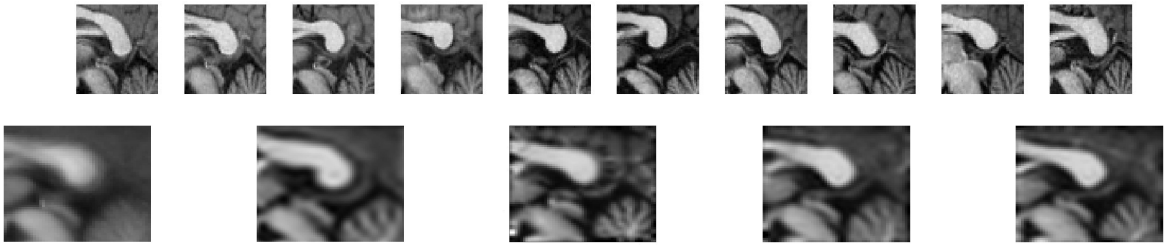
Stochastic algorithm issues. The FAM algorithm is deterministic and does not depend on any choice. Unfortunately, the stochastic algorithm requires several choices to optimize.

To optimize the choice of the transition kernel Π_η, we run the algorithm with different kernels and compare the evolution of the simulated hidden variables as well as the results on the estimated parameters. Some kernels, as an ordinary Hastings Metropolis algorithm using as proposal the prior or a standard random walk added to the current value, do not allow to visit well the entire support of the unobserved variable. From this point of view the hybrid Gibbs sampler we used has better properties and gives nice estimation results.

To prove the convergence of the stochastic algorithms, we have to suppose that as soon as the stochastic approximation wanders outside an increasing compact set, the unobserved variable needs to be projected inside a given compact set (this is the truncation on random boundaries). In practice however, this step is never required, the results presented were obtained without this control.

Finally, the initialization of the parameters can lead to undesirable effects. For example, if the first value of the photometric parameter α is set to 0, at the first iteration of the Gibbs sampler, the proposal will be accepted with probability one. Since the candidate coordinates are simulated according to the conditional a priori, the resulting vector β leads to a variation which does not

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238



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

FIGURE 2. *First row : Ten images of the training set representing the splenium and a part of the cerebellum. Second row : Results from the template estimation. (a) : gray level mean image of the 47 images. Templates estimated (b) : with the FAM (c) : with the stochastic algorithms on the simple model (d,e) : on the two component model.*

correspond to a relevant digit deformation. This implies some oscillations on the updated template. The next simulated deformation variable will try to take these oscillations into account to get closer and closer to the oscillating template, staying in its orbit.

4. Experiments

4.1. Corpus callosum

We have tested the algorithms on some medical images. The database we consider in this paragraph has 47 2D images, each of them representing the splenium (back of the corpus callosum) and a part of the cerebellum. Some of the training images are shown in Figure 2 first row.

The results of the estimation are presented in Figure 2 ((a) to (e)) where we can see the improvement from the gray level mean (a) to our estimations. Image (b), corresponding to the deterministic algorithm result, shows a well contrasted splenium whereas the cerebellum remains a little bit blurry (note that it is still much better that the simple mean). Image (c), corresponding to the stochastic EM algorithm result, presents some real improvement again. Indeed, the splenium is still very contrasted, the background is not blurry and overall, the cerebellum is well reconstructed with several branches. The two anatomical shapes are relevant representants of the ones observed in the training set.

The estimation has been done while enabling the decomposition of the database into two components. The two estimated templates (using the MCMC-SAEM algorithm) are presented in Figure 2 (d) and (e). The differences can be seen in particular on the shape of the splenium, where the fornix is more or less close to the boundary of the image and the thickness and convexity of the splenium varies. The number of branches in the two cerebella also tends to be different from one template to the other (4 in the first component and 5 in the second one). The estimation suffers from the small number of images we have. To be able to explain the huge variability of the two anatomical shapes, more components would be interesting but at the same time more images required so that the components will not end up empty.

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

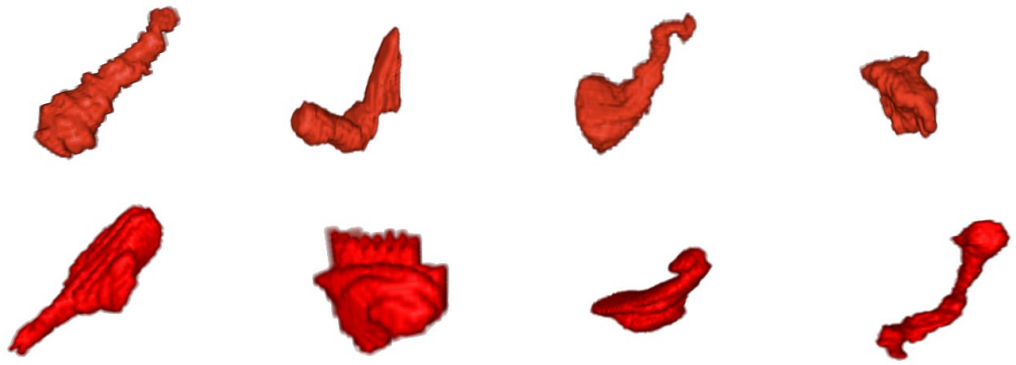


FIGURE 3. *3D view of eight samples of the data set of dendrite spines. Each image is a volume leading to a binary image.*

4.2. Murine dendrite spines

We run the stochastic algorithm on a set of murine dendrite spines [1, 8, 9]. The data set consists of 50 binary images of microscopic structures, tiny protuberances found on many types of neurons termed dendrite spines. The images are from control mice and knockout mice which have been genetically modified to mimic human neurological pathologies like Parkinson's disease. The acquisition process consisted of electron microscopy after injection of Lucifer yellow and subsequent photo-oxidation. The shapes were then manually segmented on the tomographic reconstruction of the neurons. The images are labelled by experts as belonging to six different categories: double, filopodia, long mushroom, mushroom, stubby and thin. Some of these images are presented in Figure 3. This figure shows a 3D view of some examples among the training set. Each image is a binary (background $= 0$, object $= 2$) cubic volume of size 56^3. We can notice here the large geometrical variability of this population of images.

The study in [1] showed a correlation between the spine type and its shape. This study is based on a template shape and a given metric to compare the spines through the computation of deformations. The estimation here aims at proposing one or more templates with their correlated metric in order to exhibit the common features of the population.

The computation of the Stochastic Approximation EM algorithm coupled with the MCMC procedure is performed in Matlab. Experiments were performed on 64bit system with 16GB of shared memory. Each run takes about a day with the whole data set. The main difficulty concerns the resolution of the linear system in α involved in the maximisation step at each iteration l of the algorithm. The matrix involved in this linear system is very ill-conditioned. The effects are edge effects on the template, i.e. some non-zero values of the voxel grey level on the sides of the template image. Therefore, incomplete LU factorisation as a pre-conditioner is performed to stabilise the numerical inversion. If this is insufficient (in extreme cases), one solution would be to use full or partial pivoting strategies as in Gaussian elimination. This leads to slightly longer algorithm but without numerical issues.

One step further in the optimisation of the processing time is to parallelise the loop on the observations. Indeed, given the current parameters, each observation is independent from the

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

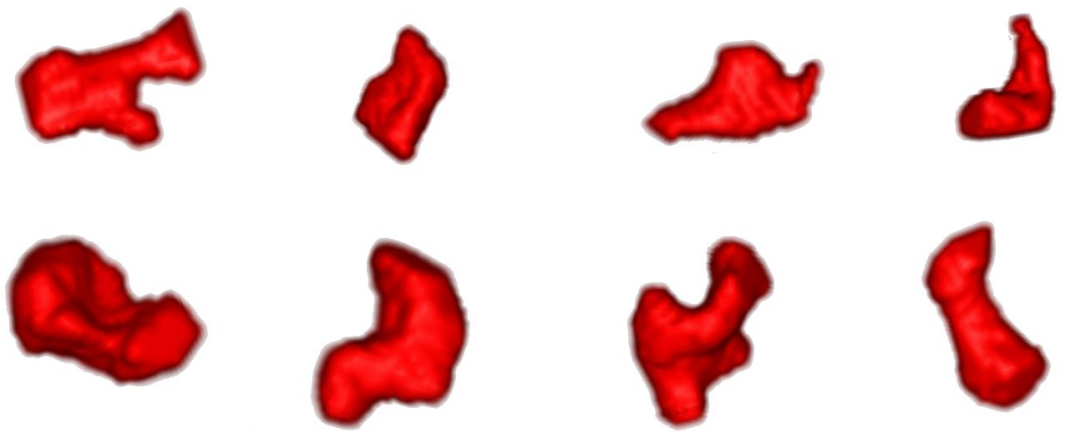


FIGURE 4. *3D view of eight synthetic data. The estimated template shown in Figure 5 is randomly deformation with respect to the estimated covariance matrix. The results are then thresholded in order to get a binary volume.*

others. The simulation step can therefore be run on separate processors. This divides the time of processing by the number of images.

4.2.1. One Component Model

In this section we present the result of the estimation using the single component model. Since the training set shows very different shapes for the six categories, a single template model might not be able to capture this large variability. In order to have a little bit smaller variability, we focused on 30 images of only three spine categoris to estimate our atlas with a single component model. We choose thin, long mushroom and stubby.

The estimated template is presented in the left column of Figure 5. The estimated image is real valued, in particular here in the segment $[0,2]$. We do not specify any criteria in order to impose a binary template. This is why the estimated volume looks blurred. For 3D visualisation, one can threshold the estimated image and binarise the values (most of the values are very close to the extrema and it only creates really sharp boundaries). The resulting shape is presented in the right column of Figure 5. As expected, the shape of this estimated spine is a relevant representation of the data set. It is smoother than the observations (as expected for an “average") but it could be one of them. The deformations that are allowed have a regularity which is given by the scale σ_g of the spline kernel K_g. Below this scale, the deformation is considered as noise. This leads to quite smooth simulated deformations which do not capture high frequency local deformations. Since the training set has very different images, the resulting estimated template is a tradeoff between this large variability and the fixed smoothness of the deformations.

One crucial improvement coming from our method is that we also get an estimation of the geometrical variability through the covariance matrix Γ_g. In order to visualise the accuracy of this coupled estimation and thanks to the generative model, we simulate new synthetic data using the estimated values of the parameters. Figure 4 shows eight images obtained by applying random deformations (sampled from $\mathcal{N}(0,\Gamma_g)$) to the estimated template. The resulting shapes look like potential dendrite spines. Indeed, we can see some similarities between these synthetic images

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

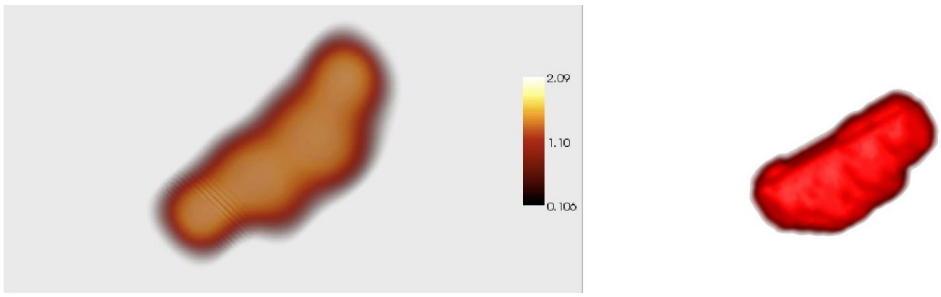


FIGURE 5. *Estimated template with the one component model: Left: 3D representation of the grey level volume. Right: 3D representation of the thresholded volume.*



FIGURE 6. *Estimated templates of the two components with the* 50 *image training set: 3D representation after thresholding.*

and some images of the data set presented in Figure 3. For example, the estimated geometrical variability has taken into account shrinking the template to get a long and thin appearance. It has also learnt to inflate one extremity and contract the other to get what is labelled as long mushroom and to make the shape more or less curved. Considering the huge dimensionality of the deformation space, this estimation is pretty good. In this model, the deformation is not constrained to be a diffeomorphism. This can affect the estimation in a way that the estimated geometrical variability could create holes or overlaps in the template. In these experiments, this problem did not occur. One way however to correct this would either be to tune the hyper-parameters which controls the deformation regularity or to use diffeomorphisms.

The last parameter which is estimated is the variance of the additive Gaussian noise. This parameter is quite interesting since it helps to see how close the model managed to fit the data. In our experiments, the estimated standard deviation of the noise in the one component case is 0.1387. Since the data set is very heterogeneous, it is very low. Indeed, as a comparison, one can look at the 2D experiments on hand written digits in [3]. The standard deviations of the digits were between 0.1 and 0.3. This suggests that the estimation in this 3D case of dendrite spine is relevant.

4.2.2. Two Component Model

The large geometrical variability of the spine shapes leads to consider several different sub-populations in the data set. However, since the data set is of small size (at most 50 images), the estimated parameters would not be accurate in a mixture model involving more than two sub-groups called components. Indeed, we have to estimate one template and one covariance

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

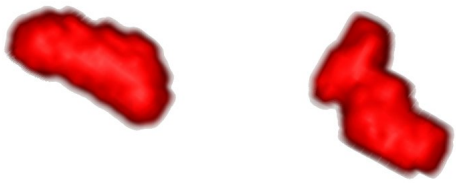


FIGURE 7. *Estimated templates of the two components with the* 30 *image training set: 3D representation after thresholding.*

matrix for each component. This leads to parameters of large dimension. The small number of images in each component would not give enough information to perform the computation of the corresponding atlas accurately. For this reason, we restrict the estimation to two components.

We ran the algorithm on the previous data set of 30 images of the three categories used for the single component estimation. We also use the whole data set of 50 images from the six dendrite spines. The estimated templates are shown in Figure 6 for the three categories and in Figure 7 for the whole training set. We only show the thresholded shapes for illustrating the differences between the two component templates.

The two estimated components show very different shapes. Indeed, we can see that the second template has a curved shape with a thin extremity and a larger one on the other side. The other template size is more isotropic. The curvature of the two shapes is also distinct. These two shapes are quite relevant representatives of the spine population. The first component looks to contain the stubby group which corresponds to plumper shapes. Whereas the second component gathers the thin and long mushroom groups. The estimated weights of the components in the population are respectively 0.32 and 0.68 which actually match the number of such shapes in the data set.

To see the impact of the different spine categories on the estimation, we ran the same algorithm with the whole data base of the 50 images with the six different categories. This training set has a larger geometrical variability than the previous one since we increase the number of spine categories considered. But the estimation may be sharper since more images are available for each component parameters to be estimated.

The two sub-groups are expected to be quite different from the previous ones and so their respective templates. These templates are shown in Figure 7. The estimated shapes are again good representations of the whole population. The subdivision is made between more isotropic shapes (similar to the previous stubby type) and longer ones, curved and with irregular boundaries. This summarises the differences which appear in the training set.

Concerning the experiment with a data set of 30 images, the estimated standard deviations are 0.1780 and 0.1659 respectively. One would expect a lower value compared to the one component. However, the small number of images leads to less precise parameters and therefore a slightly higher value of the standard deviation. For the last experiment with the whole data set, the values are 0.1521 and 0.1800. These values are quite good again compared to the 2D example of hand written digits. The slightly larger values (compared to the single component) may come from the fact that even if the training set is bigger, the variability increases as well.

It would be interesting to run the algorithm with a larger data base of only these six categories

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

and six possible components. It would also be interesting to either repeat the kind of study presented in [1] or to use the model as a classifier. Concerning this application of the model to classify new observations, this model may reach good performances in particular looking at the classification results obtained in [2] on some hand written digits where the huge geometrical variability is even higher due to some change of topology.

5. Discussion and Conclusion

We considered a generative statistical model and a stochastic algorithm to estimate mixtures of deformable templates to construct a BME Atlas. The theoretical statistical properties of the estimator and of the algorithm were proved. We validated them by numerical results. Indeed, we ran this estimation on highly variable 2D images of the splenium and 3D shapes of murine dendrite spines. The results in the one component model are relevant on both the estimation of the template image and of the geometrical variability around its template. Using the two component model, we capture more precisely the variability. This leads to two different templates representing characteristic shapes of the data set. This method can be used to estimate different population atlases such as healthy controls and Parkinson's disease populations and then compute likelihood ratios in order to classify new un-labelled images. Another possibility is to compute atlases at different stages of the disease in order to characterise its evolution. These applications may increase the knowledge and understanding of diseases. However, this modelling suffers from the fact that the deformations are not constrained to keep the topology of the shapes. Including a diffeomorphism constrain may help for clustering highlighting the topological differences within the populations.

Acknowledgements

Murine dendrite spines were initially obtained from Dr. M. Martone of NCMIR at UCSD. They were processed for image analysis at CIS under the support of NSF DMS-0101329.

References

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

[2] Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Toward a coherent statistical framework for dense deformable template estimation. *JRSS*, 69:3–29, 2007.

[3] Stéphanie Allassonnière, Estelle Kuhn, and Alain Trouvé. Bayesian deformable models bulding via stochastic approximation algorithm: A convergence study. *In Press in Bernoulli J.*

[4] Stéphanie Allassonnière and Estelle Kuhn. Stochastic algorithm for bayesian mixture effect template estimation. *In press in ESAIM Probab.Stat.*

[5] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86:376–387, 1989.

[6] Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312 (electronic), 2005.

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238

[7] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38:95–113, 2007.

[8] E. Ceyhan, L. Fong, T.N. Tasky, M.K. Hurdal, M.E. Beg, M.F.and Martone, and J.T. Ratnanather. Type-specific analysis of morphometry of dendrite spines of mice. *5th Int. Symp. Image Signal Proc. Analysis, ISPA*, pages 7–12, 2007.

[9] E. Ceyhan, R.Ç. Ölken, L. Fong, T.N. Tasky, M.K. Hurdal, M.F. Beg, M.E. Martone, and J.T. Ratnanather. Modeling metric distances of dendrite spines of mice based on morphometric measures. *Int. Symp on Health Informatics and Bioinformatics*, 2007.

[10] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[11] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1:1–22, 1977.

[13] C. A. Glasbey and K. V. Mardia. A penalised likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B*, 63:465–492, 2001.

[14] Joan Glaunès and Sarang Joshi. Template estimation form unlabeled point set data and surfaces for computational anatomy. In X. Pennec and S. Joshi, editors, *Proc. of the International Workshop on the Mathematical Foundations of Computational Anatomy (MFCA)*, pages 29–39, 2006.

[15] Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.*, 8:115–131 (electronic), 2004.

[16] 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.

[17] Frédéric. Richard, Adeline. Samson, and Charles A. Cuenod. A saem algorithm for the estimation of template and deformation parameters in medical image sequences. *Statistics and Computing*, 19:465–478, 2009.

[18] Mert Sabuncu, Serdar K. Balci, and Polina Golland. Discovering modes of an image population through mixture modeling. *Proceeding of the MICCAI conference*, LNCS(5242):381–389, 2008.

[19] Carole Twining, Tim Cootes, Stephen Marsland, Vladimir Petrovic, Roy Schestowitz, and Chris Taylor. Information-theoretic unification of groupwise non-rigid registration and model building. In *Proceedings of Medical Image Understanding and Analysis (MIUA)*, volume 2, pages 226–230, 2006.

[20] Florin Vaida. Parameter convergence for EM and MM algorithms. *Statist. Sinica*, 15(3):831–840, 2005.

[21] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *Neuroimage*, 45:61–72, 2009.

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 2102-6238