Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories

Vianney Debavelaere, Stanley Durrleman, Stéphanie Allassonnière

Noname manuscript No.
(will be inserted by the editor)

Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories

Vianney Debavelaere · Stanley Durrleman · Stéphanie Allassonnière · for the Alzheimers Disease Neuroimaging Initiative*

Received: date / Accepted: date

Abstract Given repeated observations of several subjects over time, i.e. a longitudinal data set, this paper introduces a new model to learn a classification of the shapes progression in an unsupervised setting: we automatically cluster a longitudinal data set in different classes without labels. Our method learns for each cluster an average shape trajectory (or representative curve) and its variance in space and time. Representative trajectories are built as the combination of pieces of curves. This mixture model is flexible enough to handle independent trajectories for each cluster as well as fork and merge scenarios. The estimation of such non linear mixture models in high dimension is known to be difficult because of the trapping states effect that hampers the optimisation of cluster assignments during training. We address this issue by using a tempered version of the stochastic EM algorithm. Finally, we apply our algorithm on different data sets. First, synthetic data are used to show that a tempered scheme achieves better convergence. We then apply our method to different real data sets: 1D RECIST score used to monitor tumors growth, 3D facial expressions and meshes of the hippocampus. In particular, we show how the method can be used to test different scenarios of hippocampus atrophy in ageing by using an heteregenous population of normal ageing individuals and mild cognitive impaired subjects.

Keywords Longitudinal data analysis · Mixture model · Branching population · Stochastic Optimization · Statistical Model · Riemannian manifold.

Vianney Debavelaere
Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France
E-mail: vianney.debavelaere@polytechnique.edu

Stanley Durrleman
ARAMIS Lab, Institut du Cerveau et de la Moelle épinière , 47 Boulevard de l'Hopital, Paris, France

Stéphanie Allassonnière
Centre de Recherche des Cordeliers, Universite Paris Descartes, Paris, France

* Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: adni.loni.usc.edu.

Aknowledgment

This work has been partly funded by the European Research Council with grant 678304.

1 Introduction

The emergence of large longitudinal data sets (subjects observed repeatedly at different time points) has allowed the construction of different models improving the understanding of biological or natural phenomenon. Longitudinal studies have numerous applications: understating of the differences of progression in neurodegenerative disease such as Alzheimer's, chemotherapy monitoring, facial recognition, etc.. Such medical studies enable to retrieve the global progression of the disease while explaining the inter subject variability. In particular, it would be interesting to highlight the influence of a disease on a normal ageing process and to be able to differentiate those two processes. Clinicians are also interested in the possibility to detect the moment when a disease begins to manifest itself, i.e. the moment at which a subject branches from the normal

dynamic. For instance, in the case of the Alzheimer's disease, we still do not know if the disease has a very early genesis, leading to a specific aging pattern from an early age or if it is a sudden deviation from the normal ageing process. Another example is the monitoring of tumors along treatment. Indeed, it is well known that the whole population will not react the same way to a given drug. Therefore, clustering patients would enable a specific care. In both situations, the evolution may not be smooth in the sense that the disease can show variations in dynamics according to the stage of its development. To tackle those problems, we consider that populations can follow different dynamics over time. Moreover, in order to detect subgroups with specific patterns, we implement an unsupervised clustering of the dataset. Here, our populations are therefore heterogeneous but without prior knowledge on the sub-groups composing them, thus preventing from the use of supervised approaches.
We design our model such that it is able to detect a certain fixed number of different dynamics in the population and, for each of them, to estimate a representative trajectory of that population together with the inter subjects variability. The difficulty is in fact further increased in this spatiotemporal setting since clustering may take various forms: sub-groups may follow independent trajectories, or they may follow trajectories that fork or merge at specific time-points. The former case is relevant to discover pathological sub-types having different disease course. The latter is interesting for a disease that is seen as a progressive deviation from a normal aging scenario.

Usually, shape spaces are built by considering shape data as points on a Riemannian manifold (for instance, Kendall spaces (Kendall, 1984), currents (Vaillant and Glaunès, 2005) or varifolds (Charon and Trouvé, 2013)). In such shape spaces, descriptive (Donohue et al., 2014) or generative (Jedynak et al., 2012; Durrleman et al., 2013; Allassonnière et al., 2015) models have been constructed. To deform the shapes, different frameworks can be used, among others diffeomorphic demons (Vercauteren et al., 2009) or the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework. We will here use the last. It allows us to compute the deformation from one shape to the other by coding deformations as geodesics on a Riemannian manifold and using flows of deformations (Miller et al., 2006). Given a data set of shapes, it is then possible to construct an atlas. An atlas is composed of a shape that is representative of the population, as well as the spatial variability within this population (Fletcher, 2013; Allassonnière and Kuhn, 2010; Lorenzen et al., 2005; Su et al., 2014). The next logical step is to handle longitudinal data sets. Once again, the trajectory of a shape from one time point to the other will be constructed by using flows of diffeomorphisms (Bône et al., 2018; Lorenzi et al., 2011; Singh et al., 2016; Muralidharan and Fletcher, 2012; Kim et al., 2017; Chakraborty et al., 2017). In this framework, a longitudinal atlas consists of a representative trajectory, or template, and of the spatiotemporal variability of the population. The representative trajectory is a long-term scenario of changes informed by sequences of short-term individual data. It can be seen as a geodesic (Bône et al., 2018; Schiratti et al., 2017) or a piecewise geodesic (Allassonniere et al., 2017) curve on the manifold. For instance in the case of a sphere, a geodesic on the manifold is just a great circle. Spatial and temporal deformations are then considered to generate subjects from this representative trajectory. In particular, the temporal reparametrization can be considered as a general diffeomorphism (Su et al., 2014) or as an affine reparametrization combining acceleration and offset coefficients (Bône et al., 2018).

All these methods however assumed that observations are drawn from an homogeneous population that may be summarized by a single representative trajectory. Several clustering methods have already been proposed to create atlases from cross sectional datasets in an unsupervised way (Allassonnière and Kuhn, 2010; Srivastava et al., 2005) or for longitudinal datasets of continuous trajectories in a supervised way (Abdelkader et al., 2011). However, if (Hong et al., 2015) proposes a test to detect if there is one cluster or more in a longitudinal population, there is, to our knowledge, no paper proposing a method to detect those clusters in an unsupervised way in the longitudinal framework while also creating the corresponding atlases. This will be one of the goals of this paper. Our algorithm should be able to detect sub populations that could be different from those expected and so highlight unexpected dynamics. Such a behaviour can be interesting to test different models or to highlight in a population some characteristics that were previously considered without influence on the phenomenon under study.

In this paper, we explain with more details and examples the work presented in (Debavelaere et al., 2019) where the population is supposed to contain a certain fixed number of unknown clusters. To tackle this problem, we construct a mixed-effect generative model. To estimate the different parameters, we choose to use a variant of the Expectation-Maximization algorithm called the Markov Chain Monte Carlo Stochastic Approximation Expectation Maximization algorithm

(MCMC-SAEM) (Delyon et al., 1999; Allassonnière et al., 2010). However, using those algorithms in a clustering context leads to the problem of trapping states: changing class assignment is often more costly than adjusting the parameters of the current clusters, resulting in very few updates of class assignment during optimization. Solutions have already been presented in the case of cross sectional data sets analysis but at very high computational costs (Allassonnière and Kuhn, 2010). Here, we choose to introduce temperate distributions in our Expectation-Maximization algorithm to avoid being trapped in the initial labelling.

In this paper, we will first explain in section 2 the geometrical framework allowing us to compute the representative trajectories and deformations towards the subjects. Because this framework allows us to define our model by a finite number of parameters, we will present in section 3 the statistical model and the algorithm used to estimate those parameters. Finally, we will apply our work to different data sets. We will quantitatively validate it on simulated 2D data. We will then perform experiments on real data: we will work with 1D RECIST score used to monitor the growth of a tumor (Therasse et al., 2000), with a data set of 3D faces expressing different expressions and with a 3D data set of hippocampi of patients with or without Alzheimer's disease.

2 Geometrical model

We will first present the geometrical model that allows us to compute the representative trajectory of each of our clusters as well as the deformations towards the subjects.

2.1 Construction of the representative trajectory

In the following, we consider a longitudinal data set of n subjects, each being observed k_i times: $(y_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$ at time $(t_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$, where each observation $y_{i,j}$ is a point of $\mathbb{R}^d$, $d\in\mathbb{N}$.

We first want to explain how to construct a longitudinal trajectory in a set of shapes that will, later on, define our group average. We choose to use the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework to define our shape deformations. Therefore, we can deform an initial shape using the flow of a velocity $v_t \in V$ for $t\in[0,1]$ and for V a fixed Hilbert space:

$$\begin{cases} \dfrac{\partial \phi_t^v}{\partial t} = v_t \circ \phi_t^v \\ \phi_0^v = Id\,. \end{cases} \tag{1}$$

Given velocities $(v_t)_{t\in[0,1]}$, this equation creates diffeomorphisms $(\phi_t^v)_{t\in[0,1]}$ that will deform the ambient space and so, in particular, our initial shape y_0. Hence, given velocities $(v_t)_{t\in[0,1]}$, $(\phi_t^v(y_0))_{t\in[0,1]}$ will define a longitudinal trajectory of shapes.

Each of those diffeomorphism ϕ_t^v belongs to the set $\mathcal{G}=\{\phi_1^v | v\in V\}$. This group of deformation maps is provided with a right invariant metric via

$$d(Id,\phi)=\sqrt{\inf\left\{\int_0^1 ||v_t||_V^2 dt | \phi=\phi^v\right\}}\,. \tag{2}$$

This exactly states that $\mathcal{G}$ is given the structure of a manifold on which distances are computed as the length of minimal geodesic paths connecting two elements. Given this structure, we will no longer allow any diffeomorphism to be our group average but only diffeomorphisms such that $t\mapsto\phi_t^v$ follows a geodesic path in $\mathcal{G}$.

We need now to ask ourselves how to choose velocities verifying this condition. Since we only study discrete shapes, we can place ourselves in the finite dimensional setting and suppose that our velocities $(v_t)_{t\in[0,1]}$ belong to a Reproducing Kernel Hilbert Space V with kernel K_g. V is in fact the set of squared integrable functions regularized by the convolution by the kernel K_g. A vector v in V can then be written using a set of n_{cp} control points $(c_i)_{1\leq i\leq n_{cp}}$ and momentum vectors $(m_i)_{1\leq i\leq n_{cp}}$ in $\mathbb{R}^d$: for $x\in\mathbb{R}^d$,

$$v(x)=\sum_{i=1}^{n_{cp}} K_g(c_i,x)m_i\,. \tag{3}$$

The value of v at a point x is obtained as the interpolation of the momenta at the control points.
Hence, to create a longitudinal trajectory, we now need to choose an initial shape and a set of control points and momenta defining the velocities $(v_t)_{t\in[0,1]}$ such that $(\phi_t)_{t\in[0,1]}$ defines a geodesic in $\mathcal{G}$.

It has been shown in (Miller et al., 2006) that if the initial velocity field v_0 is the interpolation of momentum vectors at control points as in Eq. (3), then the velocity field defining a geodesic path in $\mathcal{G}$ keeps the same form:

$$v_t(x)=\sum_{i=1}^{n_{cp}} K_g(c(t)_i,x)m(t)_i\,. \tag{4}$$

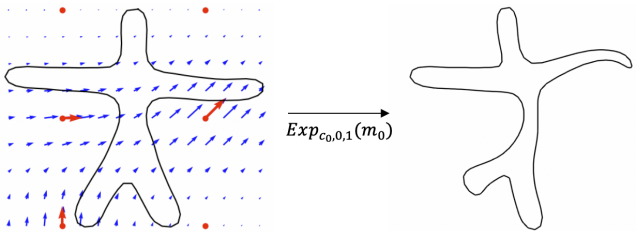

Fig. 1 The initial control points are the red points, the initial momenta, the red vectors. The blue vector field is created using the initial momenta and control points. Finally, we compute the deformation of the initial shape by this vector field.

Moreover, $m(t)$ and $c(t)$ are then time dependent momenta and control points solutions of the Hamiltonian equations:

$$\begin{cases} \dot{c}(t) = K_g(t)m(t) \\ \dot{m}(t) = \nabla_{c(t)} \left(m(t)^T K_g(t) m(t) \right) \end{cases} \tag{5}$$

with initial conditions $m(0) = (m(0)_k)_{1 \leq k \leq n_{cp}}$, $c(0) = (c(0)_k)_{1 \leq k \leq n_{cp}}$ and where $K_g(t)$ is the $n_{cp} \times n_{cp}$ kernel matrix $(K_g(c_i(t), c_j(t)))_{1 \leq i,j \leq n_{cp}}$.

To sum up, to define our longitudinal trajectory of shapes, we now only need to set an initial shape and an initial set of momenta and control points. By integrating the Hamiltonian equations (5), one can compute the evolution of those control points and momenta over time and obtain the velocity vector at any time t (Eq. (4)). By integrating the flow equation (1), we obtain diffeomorphisms $(\phi_t))_{t \in [0,1]}$ deforming the ambient space. By applying this diffeomorphism at a point cloud or mesh y_0, we are finally able to deform it.

We finally note $\text{Exp}_{c_0, t_0, t}(m_0) = \phi_t^v$ the diffeomorphism obtained above with the initial condition $\phi_{t_0}^v = Id$. This deformation process involving the Riemaniann Exponential is showed on an example figure 1.

However, in order to deal with possible change of dynamics in the population, we do not only want to consider geodesics but piecewise geodesics. Hence, we will modelize our group trajectories as a combination of K different geodesics following each other, generalizing the work done in (Allassonniere et al., 2017) in dimension 1. In particular, each of the geodesics defining γ_0 describes a dynamic of the population on a particular time segment, different from the others. The time at which the group average goes from one dynamic to the other will be called rupture times. The component of the piecewise geodesic following a rupture time will then be defined using the Exponential operator defined previously, applied at the value of the trajectory at that rupture time.

We now formalize this: we introduce a subdivision of $\mathbb{R}$: $(t_{R,1} < ... < t_{R,K-1} < t_{R,K} := +\infty)$ where $(t_{R,k})_{1 \leq k \leq K-1}$ are called rupture times i.e. times when the representative curve switches from one geodesic to another. It is at those times that the population switches from one dynamic to the other. Given a set of initial control points $c^1 \in \mathbb{R}^{n_{cp} \times d}$, of rupture times $t_R \in \mathbb{R}^{K-1}$, an initial shape x^1 and K momenta $(m^0, m^1, ..., m^{K-1})$, we define the representative trajectory as:

$$\begin{cases} \gamma(t)(x^1) = \text{Exp}_{c^1, t_{R,1}, t_{R,1}-t}(m^0) \cdot x^1 \mathbb{1}_{t \leq t_{R,1}} \\ \qquad + \sum_{k=1}^{K-1} \text{Exp}_{c^k, t_{R,k}, t-t_{R,k}}(m^k) \cdot x^k \mathbb{1}_{t_{R,k} \leq t \leq t_{R,k+1}} \\ \text{with, for } k \geq 2 : \\ \qquad c^k = \text{Exp}_{c^{k-1}, t_{R,k-1}, t_{R,k}-t_{R,k-1}}(m^{k-1}) \cdot c^{k-1} \\ \qquad x^k = \text{Exp}_{c^{k-1}, t_{R,k-1}, t_{R,k}-t_{R,k-1}}(m^{k-1}) \cdot x^{k-1} \end{cases}$$

Here, the c^k and x^k are respectively the position of the control points and the value of the representative curve at times $t_{R,k}$. For $k \geq 2$, they are fixed to assure the continuity of the trajectory.
It can be noticed that the first rupture time has a particular role as we must define a geodesic before it, determining the trajectory from $-\infty$ to the first rupture time and another after it, determining the trajectory from the first rupture time to the second. The control points c^1 and momenta m^0, m^1 are used to compute the velocities at the time $t_{R,1}$ defining the geodesic before and after it. The other momenta $m^2, ..., m^{K-1}$ and control points $c^2, ..., c^{K-1}$ define the subsequent geodesics.

The construction of a piecewise geodesic is applied on an example figure 2.

2.2 Deformations towards the subjects

We now know how to construct a longitudinal trajectory that will play the role of a representative trajectory. From this representative trajectory featuring the group characteristic path, we want to generate individual trajectories following different behaviours. To achieve this goal, we take into account both temporal and spatial differences by introducing a time reparametrization and a diffeomorphic spatial deformation.

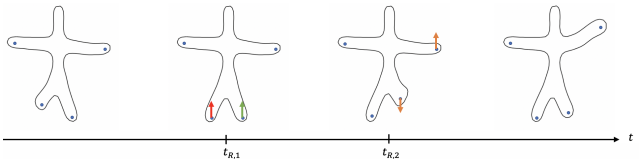

Fig. 2 Example of a piecewise geodesic with 3 parts. At the first rupture time $t_{R,1}$, the blue control points and red momenta code the exponential before it. The green momenta codes the exponential after the first rupture time. Both the control points and the shape are transported by this diffeomorphism until the second rupture time $t_{R,2}$. It is this transported shape and those transported control points that will be used, along with the orange set of momenta, to compute the deformation after the second rupture time.

2.2.1 Time reparametrization

Each individual can follow its own rhythm of progression, different from the representative curve and varying from one time segment to another, hence the need to introduce time reparametrizations.
For each subject i, let $\xi_{i,0}, ...\xi_{i,K-1}$ be acceleration coefficients and $\tau_{i,0}, ..., \tau_{i,K-1}$ time shifts. We write for every subject i:

$$\psi_{i,0}(t) = t_{R,1} - e^{\xi_{i,0}} \left(t_{R,1} - t + \tau_{i,0}\right) \tag{6}$$

and, for each time segment $k \geq 1$,

$$\psi_{i,k}(t) = t_{R,k} + e^{\xi_{i,k}} \left(t - t_{R,k} - \tau_{i,k}\right) . \tag{7}$$

$\psi_{i,k}$ codes the temporal reparametrization of the subject i on the time segment k. Once again, a first time reparametrization must be defined before the first rupture time.
The time shifts $\tau_{i,k}$ are offsets that allow the subjects to be at different stage of evolution while the acceleration factors $\xi_{i,k}$ allow an inter-subject variability in the pace of evolution on each geodesic (quicker evolution if $\xi_{i,k} > 0$, slower if $\xi_{i,k} < 0$). Both of those factors allow us to represent behaviors in the population observed by clinicians.

Different conditions must be verified to assure the continuity of the time reparametrizations. First, as the representative trajectory goes through a change of dynamics at the rupture times, each subject has its own rupture times $t_{R,i,k}$ such that $t_{R,k} = \psi_{i,k}(t_{R,i,k})$ i.e. $t_{R,i,k} = t_{R,k} + \tau_{i,k}$. Before the individual rupture time $t_{R,i,k}$, the time reparametrization is computed using $\psi_{i,k-1}$ and after it, using $\psi_{i,k}$. Hence, to assure the continuity of the global time reparametrization at each of those rupture times, we also fix all the time shifts

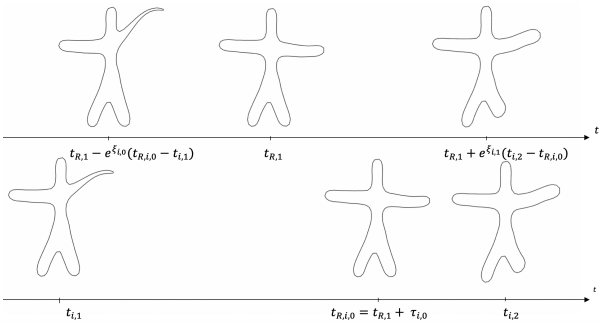

Fig. 3 Example of a time reparametrization. At the top, the representative trajectory. At the bottom, a time reparametrization towards the subject i observed at two times: $t_{i,1}$ and $t_{i,2}$. The individual rupture time of the subject i is obtained as a translation of the rupture time by $\tau_{i,0}$, here chosen positive. On the first time segment, $\xi_{i,0}$ is negative and the progression is slower than the one of the representative trajectory. On the second time segment, $\xi_{i,1}$ is positive and the progression is quicker.

but $\tau_{i,0}$ by continuity conditions: we impose for all k $\psi_{i,k-1}(t_{R,i,k}) = \psi_{i,k}(t_{R,i,k})$, i.e.: $\tau_{i,0} = \tau_{i,1}$ and, for $k \in [|2, K-1|]$,

$$\tau_{i,k} = \tau_{i,k-1} + (t_{R,k} - t_{R,k-1})(e^{-\xi_{i,k-1}} - 1) . \tag{8}$$

From now on, we note $\tau_i = \tau_{i,0}$.

It can be remarked that the choice of this particular temporal reparametrization simplifies the computations needed to assure the continuity of the final trajectory at each of the rupture time. Indeed, if we had chosen, on each component, a diffeomorphic temporal reparametrization without constraint (as done in Su et al. (2014) in the geodesic case), more complex equalities should have been imposed at each of the individual rupture times. This reparametrization has also the advantage to be easily interpreted.
Finally, we set:

$$\psi_i(t) = \psi_{i,0}(t)\mathbb{1}_{t \leq t_{R,i,1}} + \sum_{k=1}^{K-1} \psi_{i,k}(t)\mathbb{1}_{t_{R,i,k} \leq t \leq t_{R,i,k+1}} .$$

To summarize, those equations mean that the subject i at the instant t is obtained from the representative trajectory shifted by τ_i and accelerated on each time segment by $e^{\xi_{i,k}}$. The time reparametrization process is summarized figure 3.

2.2.2 Space deformations

Concerning the space deformations, as proposed in (Bône et al., 2018), we will account the space variability by using exp-parallelizations, i.e. the generalization of parallelism to geodesically complete manifolds (Schiratti et al., 2015). More precisely, we introduce for each subject i a space-shift momentum w_i. We note $P_\gamma^{(w)}$ the parallel transport which transports any vector $w \in \mathbb{R}^{n_{cp} \times d}$ along the trajectory γ. Practically, we compute it using the fanning scheme (Louis et al., 2017). Then, to code the deformation field at a time t, we transport the momentum w along the curve $\gamma(t)$ and then compute the flow given by this new momentum. The given trajectory is the exp-parallelization of γ by w_i. Hence, we define:

$$\eta_t(w) = \mathrm{Exp}_{\gamma(t)(c^1),0,1}(P_{\gamma(t)}(w)) \,.$$

Finally, given x^1 the value of the representative curve at the first rupture time, the deformation of the representative curve γ by the space shift w is given by:

$$\gamma_w(t) = \eta_t(w) \circ \gamma(t) \circ x^1 \,.$$

We give examples of the space deformation process first on Fig. 4 by computing the exp-parallelization of a trajectory on a sphere and then on Fig 5 by presenting an example in a space of shapes.

We model this space shift as a linear combination of n_s sources: we suppose that $w = As$ with A a $n_{cp} \times n_s$ matrix called the modulation matrix and $s \in \mathbb{R}^{n_s}$ the sources. This matrix plays the role of the source separation matrix also known as the modulation matrix in the Independent Component Analysis. This helps to reduce the dimension by highlighting the principal sources of deformation. By projecting all the columns of A on $(m^0, ..., m^{K-1})^\perp$ for the metric K_g, we impose orthogonality between the deformations towards the subjects and the velocity field defining our representative trajectory. It has been shown in (Schiratti et al., 2017) that this condition is necessary to assure the identifiability of the model by preventing the algorithm to consider an acceleration with respect to the representative trajectory as a space shift.
Finally, we deform the template $\gamma(t)(x^1)$ by setting:

$$\gamma_i(t) = \gamma_w(\psi_i(t)) \,.$$

2.3 Mixture and branching process

This construction builds a piecewise geodesic model of progression. Until now, it can only process homogeneous populations. We propose an extension for the

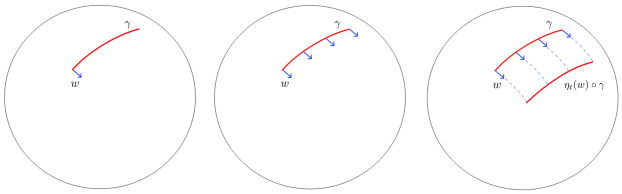

Fig. 4 Example of parallel transport on a sphere. On the left, we draw a trajectory γ and the momenta to transport w. On the center, we transport w along γ. On the right, we compute the exp-parallelization of γ by w.

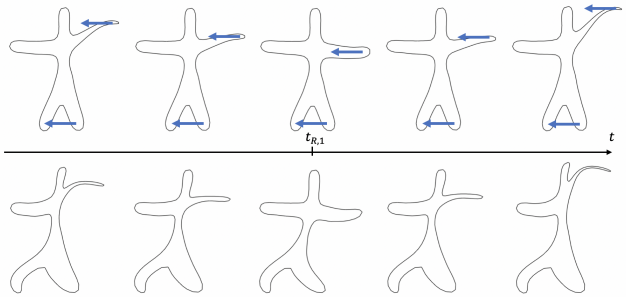

Fig. 5 Samples from a piecewise geodesic (top) and a parallel deformation (bottom). The blue momenta is first defined at the rupture time $t_{R,1}$. It is then transported along the piecewise geodesic and defines the deformation frame towards a subject.

analysis of heterogeneous populations. More precisely, we suppose there exists N different representative curves in a given population, each of the subjects i being in the cluster $cl(i)$ defined by the particular representative curve $\gamma^{cl(i)}$. This representative curve comes with its own set of rupture times $(t_{R,1}^{cl(i)} < ... < t_{R,K-1}^{cl(i)})$, initial shape $x^{1,cl(i)}$, control points $c^{1,cl(i)}$, momenta $(m^{0,cl(i)}, ..., m^{K-1,cl(i)})$ and modulation matrix $A^{cl(i)}$.

This mixture framework enables to compare and test hypothesis on the clusters. For instance, some of the time segments can be shared by several clusters. This imposes the representative curves of these clusters on these time segments to be the same. In particular, if we want some of the clusters to be equal on the first time segment, we impose $t_{R,1}^k$, $x^{1,k}$, $c^{1,k}$ and $m^{0,k}$ to be the same for these clusters. This allows us to handle populations forking or merging at the rupture times. The rupture times are then not only times when a change of dynamic occurs but also times when populations fork or merge.

Hence, we have presented a complex geometrical model allowing us to compute global trajectories and

the deformations towards subjects. Those global trajectories can take a wide variety of forms. But, in all cases, our model is parameterized by a finite number of parameters. Hence, the next step is to construct a statistical model to estimate the unknown variables. We will need to estimate the parameters defining the template as well as the clusters and the parameters defining the deformations towards the subjects. This is the goal of the next section: in section 3.1, we will present the statistical model considered while in section 3.2 we will explain how to estimate the parameters defining it.

3 Statistical Model and estimation

3.1 Statistical Model

We define a mixed effects statistical model allowing us to estimate those different parameters. We note:

$$z_{pop}^r = \left(m^{0,r}, (m^{k,r}, t_{R,k}^r)_{1\leq k\leq K-1}, x^{1,r}, c^{1,r}, A^r\right)$$

the population parameters of the cluster r and

$$z_i = ((\xi_{i,k})_{0\leq k\leq K-1}, t_{R,i,0}, s_i)$$

the deformation parameters of the subject i with ξ_i the acceleration parameters, s_i the sources and $t_{R,i,0}$ the first individual rupture time. As all the time shifts but the first one are fixed by continuity conditions (cf Eq. (8)), all subsequent individual rupture times are also fixed by an expression depending only of the first individual rupture time, the acceleration parameters and the global rupture times of the cluster.

We suppose that the subject i is obtained as a noisy deformation of the representative curve $\gamma^{cl(i)}$: $\forall i \in [|1, n|]$, $\forall j \in [|1, k_i|]$,

$$y_{i,j}|cl(i), z_{pop}^{cl(i)}, z_i \sim \mathcal{N}(\gamma_i(t_{i,j}), \sigma^2 Id)\,.$$

Such a notation implies that we are able to compute the distance between two different shapes. Depending on the application, the points constituting the shape will be labeled or not. In the first case we will be able to use a landmark distance. In the other, we will use the current (Vaillant and Glaunès, 2005) or varifold (Charon and Trouvé, 2013) distances.

We also suppose that the deformation parameters z_i verify:

$$z_i|cl(i) \sim \mathcal{N}(\mu^{cl(i)}, \Sigma^{cl(i)})$$

where for all cluster r, Σ^r is a positive-definite matrix and $\mu^r = (0, ..., 0, t_{R,0}^r)$. Unlike in (Debavelaere et al., 2019), we suppose that the first rupture time of each piecewise-geodesic $t_{R,0}^r$ is not a random variable but a parameter of our model, defined as the mean of the law of the individual rupture times. Thus, those individual rupture times are here considered as random variables. It allows to accelerate the computation time of each iteration while improving the stability of our algorithm.

The cluster r is drawn with a probability p^r i.e.

$$cl(i) \sim \sum_{r=1}^{N} p^r \delta_r$$

and finally, we suppose $z_{pop}^r \sim \mathcal{N}(\bar{z}_{pop}^r, v_{pop})$ where v_{pop} are small fixed variances so that our model belongs to the curved exponential family. Finally, our model is defined with parameters $\theta = \left((t_{R,0}^r, \Sigma^r, p^r, \bar{z}_{pop}^r)_{1\leq r\leq N}, \sigma\right)$.

For effectiveness in the high dimension low sample size setting, we work in the Bayesian framework and set the usual conjugate priors:

$$\begin{cases} t_{R,0}^r \sim \mathcal{N}(\overline{t_{R,0}^r}, v_{t_R}) \\ \Sigma^r \sim \mathcal{W}^{-1}(V, m_\Sigma) \\ \sigma \sim \mathcal{W}^{-1}(v, m_\sigma) \\ p \sim \mathcal{D}(\alpha) \\ \bar{z}_{pop}^r \sim \mathcal{N}(\bar{\bar{z}}_{pop}^r, \bar{v}_{pop}) \end{cases} \tag{9}$$

where $\mathcal{W}$ is the inverse Wishart distribution, $\mathcal{D}$ is the Dirichlet distribution and $\overline{t_{R,0}^r}$, v_{t_R}, V, m_Σ, v, m_σ, α, $\bar{\bar{z}}_{pop}^r$ and $\bar{v}_{pop}$ are hyperparameters of the model.

It is important to note that our model belongs to the curved exponential family and so allows us to define sufficient statistics. It will then be possible, in the next section, to estimate the parameters of our algorithm using only those sufficient statistics.

3.2 Estimation

To estimate the parameters θ, we want to compute a maximum a posteriori estimator by using a stochastic version of the Expectation Maximization algorithm known as MCMC-SAEM (Allassonnière and Kuhn, 2010). It consists in the following steps: (i) simulation of (z, z_{pop}, cl), (ii) stochastic approximation of the sufficient statistics of the curved exponential model and (iii) maximization using the updated stochastic approximation. We can remark that the joint distribution is in the curved exponential family which guaranties the convergence of the MCMC-SAEM algorithm, as proven in (Allassonnière et al., 2010).

Concerning the sampling, we simulate (z, z_{pop}, cl) as an iterate of an ergodic Monte Carlo Markov Chain with stationary distribution $q(z_{pop}, z, cl|y, \theta)$. More precisely, we use a symmetric random walk Monte-Carlo Markov Chain within Gibbs sampler with adapted variance. Once those variables are sampled, it is then possible to compute the sufficient statistics and to obtain the parameters maximizing the posterior distribution in a closed form.

However, using the algorithm as presented above yields to bad results in exploring the support of the conditional probability distribution. This issue is known as trapping states: once a label is given to an observation, the probability of changing to another is almost zero. This leads to no change of cluster after a few iterations. This problem has already been encountered in the clustering case, for instance in Allassonnière and Kuhn (2010) and Srivastava et al. (2005). In the first case, the authors chose to compute deformations from each template towards each subject leading to very high computational cost. In the second paper, the authors used tempered distributions but only determine the clusters without the associated representative curve and inter-subjects variability.
Here, to solve this problem, we use a tempered version of the MCMC-SAEM. Instead of targetting $q(c|y, \theta)$ in the MCMC step, we rather sample from an ergodic Markov Chain with density $\frac{1}{C(T_k)} q(c|y, \theta_k)^{\frac{1}{T_k}}$ where k is the current iteration of the algorithm, T_k is a sequence of temperature converging towards 1 and $C(T_k)$ is the normalizing constant. The higher the temperature, the flatter the distribution and the more the clusters are likely to explore the entire set.

Finding a good distribution of temperatures such that meaningful representative curves are found without immediately fixing the clusters nor forcing them to move throughout the whole algorithm is quite difficult. Several choices have been proposed in (Allassonnière and Chevallier, 2019) but we choose here a distribution that takes into account the current state of the algorithm. For each subject i and each cluster k, we set $\tau_i^k = \log\left(\frac{q(cl_(i)=j)}{q(cl(i)=k)}\right)$ where $cl(i)$ is the cluster of the subject i, j the index of that cluster during the previous iteration and q is the complete log likelihood. τ_i^k is in fact the logarithm of the acceptance rate of the MCMC-SAEM algorithm for the subject i to go from the cluster j to the cluster k. We then take:

$$T = \begin{cases} \dfrac{\text{Median}(\tau)}{\lceil \text{iter}/10 \rceil} \dfrac{5 - \text{iter}\%10}{5} + 1 - \dfrac{5 - \text{iter}\%10}{5} \\ \qquad\qquad\qquad\qquad \text{if iter}\%10 < 5 \\ 1 \text{ otherwise} \end{cases} \tag{10}$$

where % is the modulo operator and iter is the current iteration.

Such a distribution of temperature allows the representative curves to fix themselves when iter%10 ≥ 5 while forcing the clusters to explore the whole space when iter%10 < 5. Indeed, such a temperature distribution allows us to directly influences the acceptance rate of the clusters.

If this temperature scheme allows us to observe meaningful clusters, as showed later in section 4, it must be remarked that it depends of the acceptance rate τ and so of the previous state of the algorithm. The convergence of tempered SAEM algorithms has already been proven in (Allassonnière and Chevallier, 2019) and can easily be generalized in the case where the temperature depends of the previous state of the algorithm. However, for the MCMC-SAEM case used here, the geometric ergodicity of the Markov Chain should be proven in order to conclude that the algorithm converges.

The process is summarized on algorithm 1.

3.3 Initialization and influence of the hyperparameters

Now that we have presented the algorithm estimating θ, we interest ourselves in its initialization and in the influence of the choice of the hyperparameters.

Concerning the initialization, all the representative trajectories of the different clusters are chosen equally by building a constant trajectory equal to the first observation of the first subject at all times. Similarly, we initialize the individual parameters such that there is no initial deformation towards the subjects. Hence, at first, all individual trajectories are equals.

The different hyperparameters defining the priors influence the update of θ at each iteration. Indeed, all those updates can in fact be seen as barycenters between a quantity defined by the sufficient statistics and another depending on the prior. For instance, $\bar{z}_{pop}^r$ is

Algorithm 1: MCMC-SAEM algorithm

Data: $(y_{i,j})$, $(t_{i,j})$, total number of iterations K, $s_0 = 0$ and $(\Delta_k)_{k\in\mathbb{N}}$ a decreasing positive step size sequence

for $1 \leq k \leq K$ **do**

Sample (z_{pop}, z) using a single step of a Symmetric Random-Walk Metropolis Hastings within Gibbs sampler targeting the posterior distribution $q(z_{pop}, z|y, \theta_k)$.

Compute T_k using Eq. 10 and sample c using a single step of a Symmetric Random-Walk Metropolis Hastings within Gibbs sampler targeting the posterior distribution $\frac{1}{T_k}q(c|y, \theta_k)$.

Compute the stochastic approximation $s_k = s_{k-1} + \Delta_{k-1}(S(z, z_{pop}, y) - s_{k-1})$ where S are the sufficient statistics.

Update the parameters θ_k to maximize the posterior likelihood $q(\theta|y)$: $\theta_k = \hat{\theta}(s_k)$.

the barycenter between a sufficient statistic and $\bar{\bar{z}}^r_{pop}$ with respective weight $\frac{\bar{v}_{pop}}{\bar{v}_{pop}+v_{pop}}$ and $\frac{v_{pop}}{\bar{v}_{pop}+v_{pop}}$. Hence, we can choose the prior to influence the final value of $\bar{z}^r_{pop}$ and also choose the weight given to this a priori. Similar remarks can be done with all parameters.

Finally, we must also choose the kernel used to compute the deformations. Here, we take a Gaussian kernel: $K_g(x, y) = \exp\left(\frac{\|x-y\|_2^2}{\sigma_g^2}\right)$. We choose the kernel width σ_g in the range of the distance between the control points such that the whole shape can be deformed smoothly.

4 Results

4.1 2D simulated data

4.1.1 Creation of the dataset

We first test our algorithm on simulated data mimicking the shape of a dancing man. We create 100 subjects by deforming a branching piecewise-geodesic representative curve with two components. More precisely, we begin by creating the two branching representative trajectories by drawing three sets of random momenta that we apply on 16 control points equally spaced. We first apply one set of momenta on a fixed shape to obtain the first common component and then we apply the two other sets of momenta on the same fixed shape to obtain the two distinct components forking at the rupture time, set as 70. We then create our 100 individuals by sampling random accelerations, time shifts and space shifts from a gaussian distribution as well as random number of observation times before and after the rupture time. Those observation times are sampled using an exponential distribution. Finally, we add a gaussian noise of variance 0.02 to each subject, use the varifold distance and choose a kernel width equals to the distance between two adjacents control points.

4.1.2 Estimation of the parameters

We apply our algorithm to find the representative curves and the spatiotemporal deformations towards the data sequence of each subject, asking for two branching clusters. Results in Fig. 6 show that there is only little differences between the true and estimated representative trajectories (left), and no noticeable differences between the true and reconstructed observations. To quantify the reconstruction error, we compute the varifold norm of the errors for all subjects along the iterations on Fig. 7 (left).

97% of the subjects are classified in their right cluster. As for the others subjects, in most cases, no measurement is done after the rupture time or the second acceleration coefficient is so small that the shape practically does not vary after the rupture time, which explains why the algorithm cannot find the right cluster. We also show the necessity of using tempered distributions by plotting the error of classification with and without temperature on Fig. 7 (right). The oscillations we see on those figures are due to the oscillating evolution of the temperature. We can see that the classification and hence the final reconstructions are better with tempered distributions.

Finally, we launch the algorithm on the same data set 10 times to compute the errors on the estimation of the different parameters. On the table 1, we display the relative errors of the individual parameters. All those errors are below 10%, with particular good estimation for the individual rupture times. The high standard deviation observed is in fact due to the badly classified subjects. Indeed, for those subjects, the individual parameters often take absurd values: practically null accelerations, large rupture times, etc..
On the table 2, we present the errors of reconstruction for the varifold norm. We can remark that both the subjects and the templates are very well reconstructed. The error on the template is a bit higher due to the repercussion of the small errors in the temporal reparametrization. Indeed, the small errors in accelerations can cause

the time lines between the real template and the estimated one to differ causing small errors when comparing them at the same time point.

We also present the errors on our parameters table 3. Here, we can remark the very poor estimation of Σ. Once again, this is due to the presence of badly classified subjects having absurd individual parameters. Those outliers then induce a very high variance in the estimated individual parameters. However, if we try to compute the estimated Σ taking into account only the subjects in the correct cluster, we then find more correct results: an error of 8.12% with a standard deviation of 3.97. Hence, it seems impossible to have a correct estimation of Σ here.

$\xi_{i,0}$	$\xi_{i,1}$	$t_{R,i,0}$
$5.89\%_{\pm 7.01}$	$8.60\%_{\pm 10.7}$	$0.76\%_{\pm 1.61}$

Table 1 Mean and standard deviation of the relative errors for the temporal parameters.

Subjects	Templates
$1.23\%_{\pm 1.96}$	$5.56\%_{\pm 2.60}$

Table 2 Mean and standard deviation of the errors of reconstruction for the subjects and templates.

$t_{R,0}$	Σ	σ	p
$0.25\%_{\pm 0.17}$	$160\%_{\pm 223}$	$7.19\%_{\pm 4.01}$	2%

Table 3 Mean and standard deviation of the errors on the parameters θ.

4.1.3 Prediction of new data

Here, we test the ability of our model to predict new data by using cross validation. We create 100 new subjects deformed from the same representative curve as before. We then ask our algorithm to classify and reconstruct the trajectories while fixing the parameters θ and the representative curve by those learned previously. This time, 91% of the subjects are well classified and the error of reconstruction is only 0.84% with a standard deviation of 1.93. Hence, our model can process new data without a problem, proving that we have no problem of overfitting or selection bias.

4.1.4 Comparison of the clustering with a baseline

We now want to test the performance of the clustering of our model against a baseline. To do so, for each of the subjects, we compute the trajectory minimizing the distance with the observations using a geodesic regression. We obtain, for each subject, a set of momenta defining its trajectory. We then use the kmeans algorithm on the set of all momenta to classify the subjects. This algorithm will not create representative trajectories nor compute the variability of the population but will only classify the subjects without any time reparametrization.

In this easy example where only the global movement of the shapes is important in the clustering, the baseline gives us a perfect classification of the subjects. However, it is easy to create cases where our algorithm will outperform the baseline. Indeed the baseline only takes into account space deformations. Hence, it is unable to distinguish two different objects deformed the same way. For instance, a geodesic regression will give us the same set of momenta for squares and spheres following the same movement. Hence, the baseline will not be able to distinguish two different clusters. In contrast, our algorithm also takes into account the mean shape of each cluster and so is able to separate two such clusters.
Moreover, no time reparametrization is taken into account by the baseline. To highlight this fact, we create a new dataset of "dancing men" with two clusters, each containing 100 subjects. We obtain those subjects from the same representative curve but, for one cluster, the mean acceleration of the subjects e^{ξ} is 1.3 while the other has a mean acceleration of 0.7. This time, the baseline is unable to distinguish the two clusters as the momenta obtained by geodesic regression for the different trajectories are all collinear. All the subjects but 6 are placed in the same cluster and so only 51% of the subjects well classed. On the other hand, our algorithm is more successful in this clustering task: subjects are indeed classified according to their speed of progression: 84% of the subjects are classified as expected. As for those badly classified, their acceleration is close to 1.

Finally, when the only distinction between clusters is based on their space deformation, the baseline seems as precise as our algorithm. However, it is not able to distinguish differences in time and is more limited than

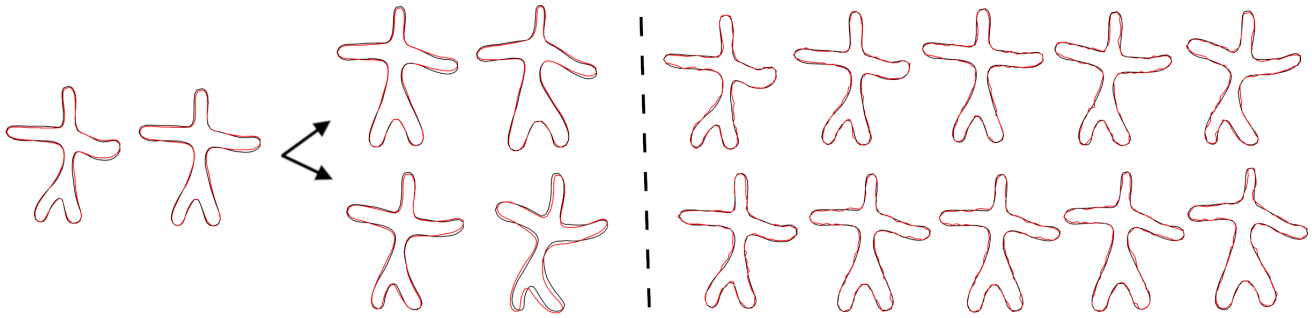

Fig. 6 In red, the exact simulated data, in black, the results given by our algorithm. On the left, the representative curves that split up at a certain rupture time. On the right side, two subjects given with their reconstructions.

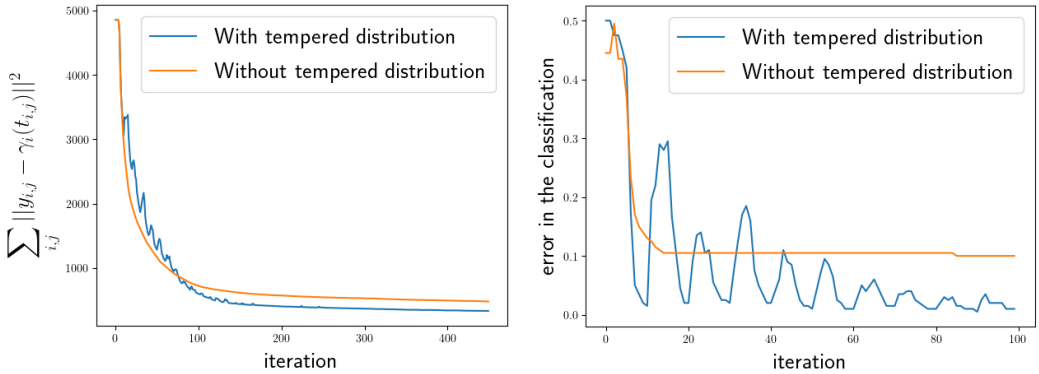

Fig. 7 Left: evolution of the varifold distances between the subjects and their reconstructions. Right: percentage of error in the classification along the first 100 iterations. With tempered distribution, the oscillating temperature coerces a lot of subjects to change classes. After 500 iterations, the error is 31.3% smaller.

our model. Those observations will be confirmed in the next examples.

4.1.5 Test of an hypothesis on the model

We want now to test hypothesis about the heterogeneity of the population. We run our algorithm on the dataset created section 4.1.1, supposing first that the two representative trajectories are different. We then run it again supposing that their first component is the same and that they fork at the rupture time. To select the model, we first compute the log-likelihood ratio test. However, in this case, this test is not enough to determine which model to choose. Indeed, with two independent representative curves, the algorithm can reconstruct the subjects as precisely as with branching representative curves. Hence, the difference between the likelihoods of the two models is too small to conclude and the test unstable between runs. To overcome this problem, we choose to compute the Bayesian Information Criterion (BIC):

$$\text{BIC} = ln(n)m - 2ln(q(y, z, \theta))$$

where m is the total number of parameters involved in the model and n the number of subjects.
This criterion takes into account the complexity of the model by adding a penalty proportional to the number of parameters involved. Hence, we will penalize the model with two independent trajectories (as it involves more parameters) even if the reconstruction is similar. This time, there is a difference of 2.98% between the two BIC criterions, leading us to choose, as expected, the model with branching representative curves.

4.2 1D RECIST scores

We test here the algorithm on a real 1D dataset. We consider a database of patients suffering from the metastatic kidney cancer and taking antiangiogenic drugs. They come on a regular basis at the hospital to check the tumor evolution. Two behaviours are expected in the population: for all patients, the tumor first regresses. But then, for some, it stabilizes while for others the tumor size increases again forcing to change the treat-

ment. The RECIST score is a feature that measures the tumor size and is used in the majority of clinical trials evaluating cancer treatments for objective response in solid tumors. Our dataset consists in the evaluation of the RECIST score for 176 patients with an average of 7 visits per subject and an average duration of 90 days between consecutive visits.

In this 1D case, shapes are just curves on $\mathbb{R}$ and we work with a logistic metric. The parallel transport is just a translation of the geodesic. That is why we rather considerate another space reparametrization, as done in (Allassonniere et al., 2017): for all classes i and all components l, we set:

$$\phi_{i,l}(x) = \gamma^{cl(i)}(t_R^{cl(i)}) + e^{\rho_i^l}\left(x - \gamma^{cl(i)}(t_R^{cl(i)})\right) + \delta_i^l\,.$$

ρ_i^l is a dilatation factor and δ_i^l is a translation factor. As with the time reparametrization, all the translation factors but the first one are fixed by continuity conditions and we note $\delta_i^0 = \delta_i$. Finally, our individual curve is defined by deforming spatially each component of $\gamma^{cl(i)}$ by $\phi_{i,l}$ and temporally by the same $\psi_{i,l}$ as previously.

With only two components, the piecewise geodesics for the logistic metric can be parameterized, for any class r, by:

$$\begin{cases} \gamma_1^r(t) = \dfrac{\gamma_{init}^r + \gamma_{escap}^r e^{a_r t + b_r}}{1 + e^{a_r t + b_r}} \\ \gamma_2^r(t) = \dfrac{\gamma_{fin}^r + \gamma_{escap}^r e^{-(c_r t + d_r)}}{1 + e^{-(c_r t + d_r)}} \\ \gamma^r(t) = \gamma_1^r(t)\mathbb{1}_{]-\infty, t_R^r]} + \gamma_2^r(t)\mathbb{1}_{]t_R^r, +\infty[}\,, \end{cases} \tag{11}$$

with $\gamma_{init}^r, \gamma_{escap}^r, \gamma_{fin}^r \in \mathbb{R}$. We fix a_r, b_r, c_r and d_r by asking the geodesics $\gamma_{0,r}^1$ and $\gamma_{0,r}^2$ to be ν-near their geodesics at an initial time t_0^r, at the rupture time t_R^r and at a final time t_1^r (see Allassonniere et al. (2017) for more details). Hence, rather than sampling momenta and control points, we will sample $z_{pop}^r = (\gamma_{init}^r, \gamma_{escap}^r, \gamma_{fin}^r, t_0^r, t_1^r)$. This whole process is summarized Fig. 8.

First, we launch our algorithm looking for two different representative curves. The result is displayed on the first line of figure 9. Our algorithm is indeed able to explain the variability of the population. However, it seems that our algorithm favours size over response dynamic as a clustering feature: small initial tumors (blue curve, 28% of the patients) are separated from big initial tumors (orange curve, 72% of the patients). For example, the orange reconstructed trajectory (top right plot) is classified with the blue template (top left plot) even if the treatment stays effective.

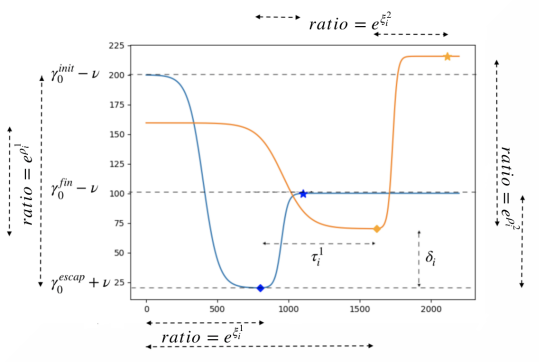

Fig. 8 Model description. In blue, the template with the different parameters defining it and in orange one subject obtained by deforming it. Here, $t_0 = 0$, the rupture points are represented by diamonds and the final times t_1 by stars.

To overcome this trivial differentiation based on the tumor initial size, we ask the two templates to be the same until the rupture time using a branching process. This time, on the second line of figure 9, we really see two different behaviours: for one of the template, the RECIST score increases a lot more (blue curve, 37% of the patients) than for the other (orange curve, 63% of the patients). As for the clustering, we see indeed that the subjects whose RECIST score do not increase after the rupture time are pooled together (green, red, orange and blue curves). Hence, we are able to separate the patients whose tumor becomes resistant to the treatment from the others. It can also be remarked that we have fewer time points for patients whose tumor becomes resistant because the clinicians change the treatment when this resistance is remarked and so the record of score for this patient stops.

4.3 3D faces

We now obtain shapes of subjects expressing different facial expressions from the Birmingham University 3D dynamic facial expression database (Yin et al.). This real database contains short videos from 101 subjects expressing happiness or surprise. We uniformly extract 8 frames, from the first to the 36-th one, which correspond to a subsampling of the first 1.4 seconds of each video. We do not work directly with the texture video, but with a set of 75 semi-automatically extracted landmarks, which were readily available along with this data set. Every set of 3D landmarks is registered to a reference one by Procrustes alignment.

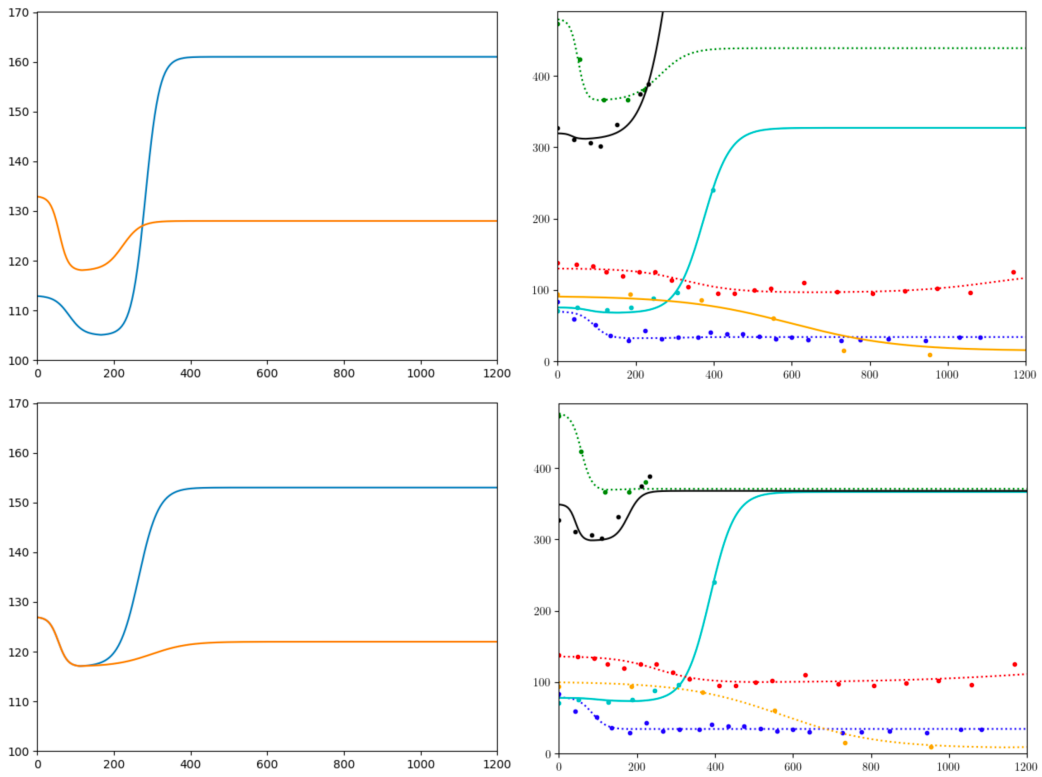

Fig. 9 At the top, the results given with two different templates, at the bottom, with two templates whose first component is the same. To the left, our templates. To the right, 6 subjects and their reconstructed trajectories. In dotted lines, subjects in the cluster of the orange template. In plain lines subjects in the cluster of the blue template.

We apply our algorithm, once again with the varifold distance, to find two clusters, with only one component geodesic for each template. As we can see Fig. 10 and 11, the faces are well reconstructed and we can recognize the two expressions of surprise and happiness on the two templates. In particular, for the surprise cluster, the mouth is more widely open, while the eyes are wide open and the eyebrows higher.

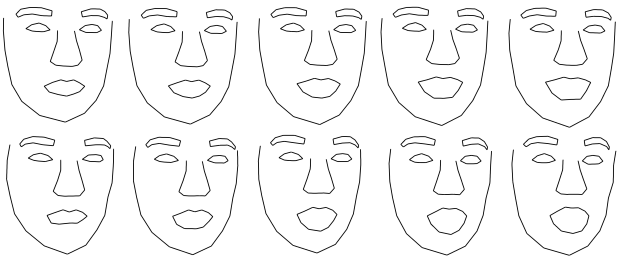

Fig. 10 Results of the algorithm when applied to a dataset of surprised or happy visages. At the top, the evolution of the template of the happiness cluster, at the bottom, the evolution of the template of the surprised cluster, one component for each template.

Hence, we can ask ourselves if the algorithm has really detected those two expressions or if another characteristic has been detected to distinguish two sub populations. In fact, 68.5% of the subjects are classified as expected (i.e. surprised subjects in the cluster with the template looking surprised and happy subjects in the one looking happy). There are different explications about the subjects classified differently. First, we can remark that some of them have a non neutral expression at the first image, for example smiling at the beginning while they should express surprise. For others, it is just really difficult (even for a human) to determine if they express happiness or surprise (see Fig. 12). Finally, we can also remark figure 10 that the left eyebrow is quite different from one template to another. And indeed, we find that same difference in several subjects misclassified. However, even if the clustering can be surprising, the algorithm fulfilled his role: we have been able to highlight two different dynamics in the population that can be explained by differences in the subjects considered.

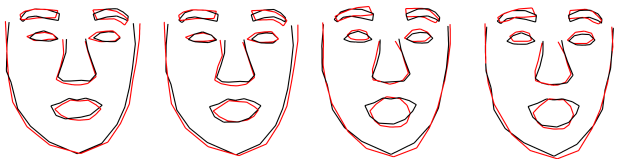

Fig. 11 Reconstitution of a subject expressing surprise. In red, the exact data, in black the reconstitution.

Fig. 12 Evolution of subject that has been asked to express happiness but seems to express surprise. It is indeed classed in the template looking surprised by our algorithm.

Concerning the baseline, we have a better classification in this case: 88% of the subjects are classified as expected. This better classification can be explained by the fact that the movement of the lips and eyebrows is the principal feature separating the two clusters. By not taking into account the initial shape of the subjects but only the deformation, the baseline is able to obtain a better classification result. In this case, if we are interested in separating the happy subjects from the surprised ones, it would thus be preferable to first compute the clusters using the baseline and only after to run our algorithm in a supervised way with the fixed clusters obtained previously to obtain the representative trajectories and the variability in each cluster.

4.4 Hippocampi dataset

We finally test the algorithm on 100 subjects obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). 50 of those subjects are control patients (CN) and 50 are Mild Cognitive Impairment subjects eventually diagnosed with Alzheimer's disease (MCIc). Meshes of the right hippocampus is segmented from the rigidly registered MRI.

We first run our algorithm with a forking model: we look for two clusters that fork at a certain rupture time. As there is no reason for the control subjects to have two different dynamics, we also ask one of the cluster (i.e. one of the evolution scenario) to follow the same geodesic before and after the rupture time. Finally, we choose to use the varifold distance. Our algorithm splits the patients in two clusters, one of them presenting a quicker and different pattern of atrophy (Fig. 15 and left side of Fig. 13 where the hippocampi volume is plotted along time). Moreover, 72% of the subjects are classified as expected: the CN in the cluster with a single dynamic showing a slower atrophy and the MCIc in the cluster with a faster atrophy after the rupture time.

We have also studied the relation between our rupture time and the age of diagnosis. The individual rupture times are strongly correlated to the diagnostic age, indicating that we have been able to detect a change of behaviour correlated with the date of diagnosis (Fig. 14).

We run again the algorithm, this time looking for two clusters with separate trajectories, one of them with only one dynamic. The results are presented Fig. 16 and on the right side of Fig. 13 for the hippocampi volumes evolution. It is interesting to remark that the cluster with only one dynamic also presents a slower atrophy, as expected with a normal ageing. We can also detect different patterns of atrophy before and after the rupture time for the cluster with two dynamics. This time, 70% of the subjects are classified as expected: CN in the cluster with one dynamic and MCIc in the cluster with two dynamics and a quicker rate of atrophy.

As we are given two possible evolution scenarii, it is natural to try to quantify the goodness of fit of each of them, allowing for a choice of a better explanation of the disease. As for synthetic data, we use the Bayesian Information Criterion. We find a difference of 2.92% between the two BIC values leading to choose the branching model. Hence, this suggests that the MCI subjects first follow a normal aging scenario but deviate from it at the rupture time. It must however be remarked that our model is quite complex with a lot of high dimensional variables, making model selection quite difficult.

Once again, we compare those results with the baseline. However, in this case, the difference between the two clusters is largely coded by the speed of atrophy and not the global dynamic. Hence, it is not surprising to note that practically all the subjects are grouped in the same cluster by the baseline and so, only 52% of the subjects are well classified. Thus, in this example, our algorithm has to be used to cluster the subjects.

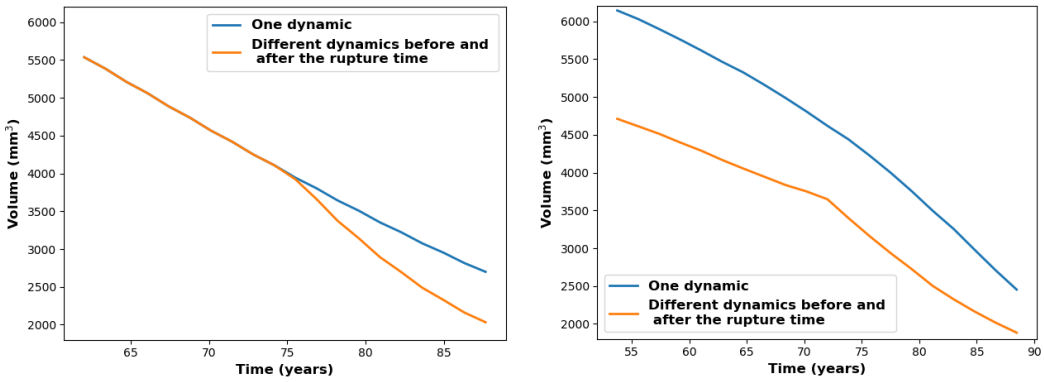

Fig. 13 Left: volume evolution for two branching clusters. Right: volume evolution for two clusters with separate trajectories.

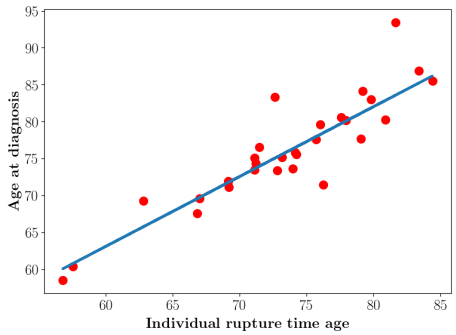

Fig. 14 Comparison of the age at diagnosis with the individual rupture time for the MCIc patients in the case of the branching model, $R^2 = 0.91$

5 Conclusion

We proposed a mixture model for longitudinal shape data sets where representative trajectories take the form of piecewise geodesic curves. Our model can be applied in a wide variety of situations to test whether subpopulations are independant from each other or fork or merge at different time-points. We showed on simulated examples that our tempered optimization scheme is key to achieve convergence of such a mixed effect model combining discrete variables with continuous variables of high dimension. It has also been noticed that taking only into account the individual trajectories is not always enough to obtain a meaningful clustering of the population. We have shown the versatility of our model by applying it to a lot of different cases: trajectories with one or several dynamics, branching or not after a rupture time, with one part of the population still following the same dynamic or not after the rupture time. Its application on 1D data allowed us to present results of the same model in another setting while the application with 3D faces showed that we can highlight different meaningful dynamics in a same population. Finally, the hippocampi data set allowed us to investigate the relationship between normal and pathological ageing.

Different questions still have to be answered. In particular, our scheme of temperature depends of the current state of the algorithm and a proof of convergence should be provided in this situation. Moreover, specific model selection criterion should be devised in this complex longitudinal setting. Those criterion should in particular help us to detect the optimal number of clusters and rupture times.

References

Mohamed F. Abdelkader, Wael Abd-Almageed, Anuj Srivastava, and Rama Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on Riemannian shape manifolds. *Computer Vision and Image Understanding*, 2011. ISSN 10773142. doi: 10.1016/j.cviu.2010.10.006.

Stéphanie Allassonnière and Juliette Chevallier. A new class of em algorithms. escaping local minima and handling intractable sampling. 2019.

Stéphanie Allassonnière and Estelle Kuhn. Stochastic algorithm for bayesian mixture effect template estimation. *ESAIM: Probability and Statistics*, 14:382–408, 2010.

Stéphanie Allassonnière, Estelle Kuhn, Alain Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.

Stéphanie Allassonnière, Stanley Durrleman, and Estelle Kuhn. Bayesian mixed effect atlas estimation

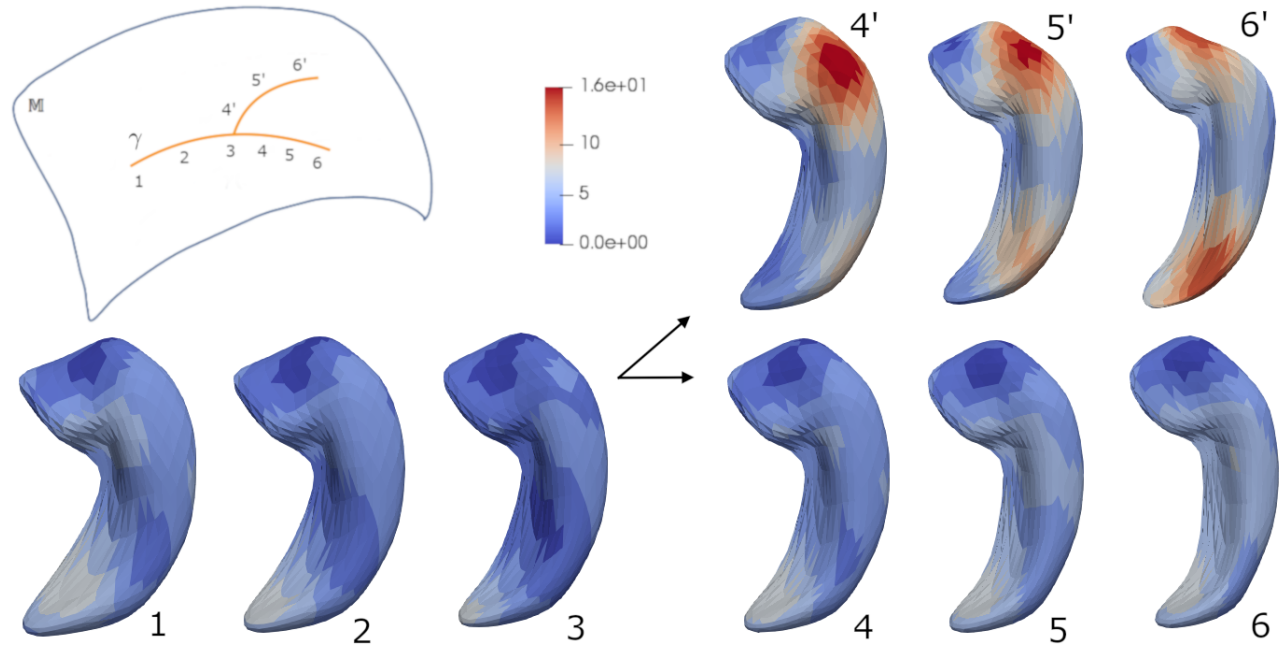

Fig. 15 Representative shape evolution at the ages 63.4y, 68.8y, 74.2y (i.e. rupture time), 75.5y, 80.9y and 86.3y. Bottom shapes: cluster with one dynamic. Top shapes : cluster with change of dynamic after rupture time. The color map gives the norm of the velocity field $\|v_t\|$ on the meshes.

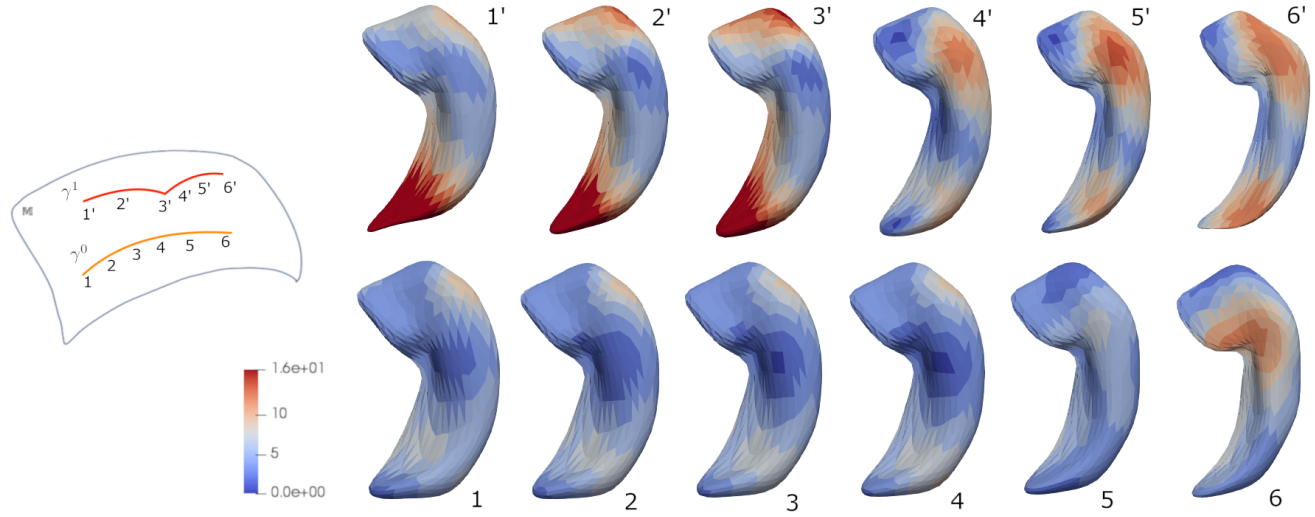

Fig. 16 Model with two separate clusters, one of them following only one dynamic. Representative shape evolution at the ages 64.7y, 68.3y, 74.3y (i.e. rupture time for the cluster with two dynamics), 77.5y, 80.9y and 86.3y. Top shapes : cluster with change of dynamic after rupture time. Bottom shapes: cluster with one dynamic. The color map gives the norm of the velocity field $\|v_t\|$ on the meshes.

with a diffeomorphic deformation model. *SIAM Journal on Imaging Sciences*, 8(3):1367–1395, 2015.

Stéphanie Allassonniere, Juliette Chevallier, and Stephane Oudard. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *Advances in Neural Information Processing Systems*, pages 1152–1160, 2017.

Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 9271–9280, 2018.

Rudrasis Chakraborty, Vikas Singh, Nagesh Adluru, and Baba C Vemuri. A geometric framework for statistical analysis of trajectories with distinct temporal spans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 172–181, 2017.

Nicolas Charon and Alain Trouvé. The varifold representation of nonoriented shapes for diffeomorphic registration. *SIAM Journal on Imaging Sciences*, 6 (4):2547–2580, 2013.

Vianney Debavelaere, Alexandre Bône, Stanley Durrleman, and Stéphanie Allassonnière. Clustering of longitudinal shape data sets using mixture of separate or branching trajectories. 2019.

Bernard Delyon, Marc Lavielle, Eric Moulines, et al. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1): 94–128, 1999.

Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Raman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5): S400–S410, 2014.

Stanley Durrleman, Stéphanie Allassonnière, and Sarang Joshi. Sparse adaptive parameterization of variability in image ensembles. *International Journal of Computer Vision*, 101(1):161–183, 2013.

P Thomas Fletcher. Geodesic regression and the theory of least squares on riemannian manifolds. *International journal of computer vision*, 105(2):171–185, 2013.

Yi Hong, Nikhil Singh, Roland Kwitt, and Marc Niethammer. Group testing for longitudinal data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9123, pages 139–151. Springer Verlag, 2015. doi: 10.1007/978-3-319-19992-4_11.

B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, J. L. Prince, et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.

David G. Kendall. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, mar 1984. ISSN 00246093. doi: 10.1112/blms/16.2.81. URL `http://doi.wiley.com/10.1112/blms/16.2.81`.

Hyunwoo J. Kim, Nagesh Adluru, Heemanshu Suri, Baba C. Vemuri, Sterling C. Johnson, and Vikas Singh. Riemannian nonlinear mixed effects models: Analyzing longitudinal deformations in neuroimaging. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 5777–5786. Institute of Electrical and Electronics Engineers Inc., nov 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.612.

Peter Lorenzen, Brad C Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–418. Springer, 2005.

Marco Lorenzi, Nicholas Ayache, and Xavier Pennec. Schilds ladder for the parallel transport of deformations in time series of images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 463–474. Springer, 2011.

Maxime Louis, Alexandre Bône, Benjamin Charlier, Stanley Durrleman, Alzheimers Disease Neuroimaging Initiative, et al. Parallel transport in shape analysis: a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, 2017.

Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006.

Prasanna Muralidharan and P Thomas Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1027–1034. IEEE, 2012.

Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412, 2015.

Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *The Journal of Machine Learning Research*, 18(1): 4840–4872, 2017.

Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. Hierarchical geodesic models in diffeomorphisms. *International Journal of Computer Vision*, 117(1):70–92, 2016.

Anuj Srivastava, Shantanu H. Joshi, Washington Mio, and Xiuwen Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):590–

602, apr 2005. ISSN 01628828. doi: 10.1109/TPAMI.2005.86.

Jingyong Su, Sebastian Kurtek, Eric Klassen, Anuj Srivastava, et al. Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552, 2014.

Patrick Therasse, Susan G Arbuck, Elizabeth A Eisenhauer, Jantien Wanders, Richard S Kaplan, Larry Rubinstein, Jaap Verweij, Martine Van Glabbeke, Allan T van Oosterom, Michaele C Christian, et al. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216, 2000.

Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 381–392. Springer, 2005.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*, 45(1 Suppl), 2009. ISSN 10959572. doi: 10.1016/j.neuroimage.2008.10.040.

L Yin, X Chenand Y Sun, T Worm, and M Reale. A high-resolution 3d dynamic facial expression database, 2008. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, volume 126.