Learning spatio-temporal trajectories from manifold-valued longitudinal data
Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, Stanley Durrleman

To cite this version:

HAL Id: hal-01245909
https://hal.archives-ouvertes.fr/hal-01245909
Submitted on 18 Dec 2015

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.

Public Domain
Learning spatio-temporal trajectories from manifold-valued longitudinal data

Jean-Baptiste Schiratti, Stéphane Allassonnière, Olivier Colliot, Stanley Durrleman
Inria Paris-Rocquencourt, Sorbonne Universités, UPMC Univ Paris 06 UMR S127, CNRS UMR 7225, ICM, F-75013, Paris, France
CMAP, Ecole Polytechnique, Palaiseau, France

Introduction

Aim: model the progression of neuro-degenerative diseases

- Understanding the progression of neuro-degenerative diseases, such as Alzheimer’s Disease (AD) is necessary for early diagnosis and care planning.
- We need to validate experimental hypothetical models of disease progression, such as [Clifford Jack et al., 2010].

Working with longitudinal data in the context of neuro-degenerative diseases raises two difficulties:

- Two individuals of the same age might be at very different stages of disease progression → statistical models based on the regression of measurements with age are inadequate to model disease progression and age should not be treated as a covariate but as a random variable.
- Longitudinal measurements sometimes belong to Riemannian manifolds (non-Euclidean spaces) → statistical models for such longitudinal data should be defined for manifold-valued measurements. Linear mixed-effects models ( Laird and Ware, 1982 ) are not defined for manifold-valued measurements.

Generic spatio-temporal model for longitudinal data

Summary: we propose a generic mixed-effects model for longitudinal manifold-valued data. The model allows to estimate an average trajectory as well as individual trajectories. Random effects allow to characterize changes in direction and pace at which individual trajectories are followed. This model is used to analyze the temporal progression of a family of univariate biomarkers.

\[ \langle B^t, \mathbb{R} \rangle \text{ smooth Riemannian manifold included in } \mathbb{R}^n \]
\[ \langle M, \mathbb{R} \rangle \text{ sub-Riemannian manifold of } M, \text{ assumed to be geodesically complete } \]
\[ p(t), \bar{p}(t), p \in T_pM, \text{ Exp}_p(\cdot) : \text{ Riemannian exponential in } M \text{ at } p \text{ of the tangent vector } v \]
\[ y \in \mathbb{R} \text{ or } M : \text{ geodesic of } M \]
\[ t \in [0, \tau] \text{ or } [0, \tau] \times \mathbb{R}_+ \text{ parallel transport in } M \text{ along } y \text{ from } y(0) \text{ to } y(t) \]
\[ t \rightarrow \text{Exp}_p (v(t)) : \text{ geodesic of } M \text{ which goes through } y \text{ at time } t \text{ with velocity } v \]

A hierarchical model!

Average trajectory:

\[ t \rightarrow \mathbb{E}[y(t)] = \mathbb{E}[\text{Exp}_p (v(t))] \]
\[ t \rightarrow \mathbb{E}[y(t)] = \mathbb{E}[\text{Exp}_p (v(t))] \]
\[ y_{t=1} = \mathbb{E}[y(t=1)] + \epsilon \]

1. The average trajectory \( t \rightarrow \mathbb{E}[y(t)] \) is chosen to be the geodesic \( t \rightarrow \text{Exp}_p (v(t)) \), \( p \in M, v \in T_pM \)

2. The trajectory of the \( t \)-individual is obtained in two steps. We start by constructing the parallel shift of the average trajectory by using a tangent vector \( v \), which we choose orthogonal to \( v \),

3. The trajectory \( t \rightarrow \mathbb{E}[y(t)] \) is then obtained by reparametrization in time the parallel shift \( \mathbb{E}[y(t)] \) using the affine time reparametrization \( q_{\lambda}(t) = \epsilon(t) + \epsilon \) in order to allow for the variability in stages of disease progression across the population.

The operation of parallel shifting, on the manifold \( M \), using a tangent vector, is defined as follows:

Definition: \( \text{Exp}_p (v(t)) \), \( v \in T_pM \). \( \theta \in \mathbb{R} \).

The curve \( s \rightarrow \mathbb{E}[y(s)] \) defined by:

\[ s \rightarrow \mathbb{E}[y(s)] = \text{Exp}_p (v(t)) \]

is said to be the \( s \rightarrow \mathbb{E}[y(s)] \) using \( v \).

By virtue of the tubular neighborhood theorem [Fuchs M.W., 2012], parallel shifting defines a local spatio-temporal coordinate system.

The model:

\[ y_{t=1} = \eta_1^t(y, p(t=1)) + \epsilon_1 \]

where \( \eta_1(t) = \alpha_1(t - t_0 - t_0) + \epsilon_2 \) and:

\[ \alpha_1 \text{ is a subject-specific acceleration factor } \alpha_1(t) = \exp(\lambda_1(t)) \text{ and } \lambda_1 \geq 0 \text{ for all } t \]

\[ \lambda_1 \text{ is a subject-specific time shift } \lambda_1 = \text{Node} \text{ of } M \text{ for all } t \]

\[ \epsilon_2 \text{ is a subject-specific random effects } \epsilon_2 \]

\[ \epsilon_0 \text{ is a subject-specific random effects } \epsilon_0 \]

\[ w_0 = \text{mean and } Y_1 \leq Y_1 \leq Y_2 \text{ } \epsilon_2 \text{ on } (0, 1) \]

In the spirit of Independent Component Analysis, the space shift \( w_0 \) appears as a linear combination of the independent components, namely the columns of the matrix \( A_0 \).

Three particular cases of our generic spatio-temporal model:

- A straight lines model [Schiratti et al., IPMI 2015]
- Geodesics are straight lines

- \( y_{t=1} = p_0 + \alpha_1(t - t_0 - t_2) + \epsilon_2 \]

Our approach

- \( Y_{t=1} = (a(t), b(t), c(t)) \)

- \( y_{t=1} = (a(t), b(t)) \)

Note that this model is not equivalent to a linear model on the logit of the observations: the logit transform corresponds to the Riemannian logarithm at \( p_0 = (1,0,0) \). The model written in the tangent space is still not linear due to the multiplication between the random effects \( a(t) \) and \( b(t) \).

A progression model for a family of univariate biomarkers:

Aim: we want to analyze the temporal progression of a family of \( N \) biomarkers.

We assume that the measurements of each biomarker belong to a one-dimensional Riemannian manifold \( \mathbb{R} \), geodesically complete and included in \( M \). As a consequence, \( M \) is a product of one-dimensional manifolds: \( M = \mathbb{R}^N \times \mathbb{R}_+^N \).

The average trajectory \( t \rightarrow \mathbb{E}[y(t)] \) is chosen among a parametric family of geodesics of \( M \):

\[ y_{t=1} = \left( \frac{1}{1 - (c_1(t) + 1)} \left( \exp \left( \frac{a(t) + b(t)}{1 - (c_1(t) + 1)} \right) \right) \right) \]

where the parameters \( \lambda_1 (1 \leq i \leq N) \) correspond to the relative delay between the biomarkers and \( \tau_0 (1 \leq i \leq N) \) is a geodesic of the one-dimensional Riemannian manifold \( \mathbb{R} \) (straight line, logistic curve, …)

If the space of observations is the open interval \( (0, 1) \) for all the biomarkers (we consider normalized measurements), the manifold trif \( M = \mathbb{R}^N \times \mathbb{R}_+^N \) with the product metric (see "Logistic curves model")

Writing the generic spatio-temporal model in this case leads to a progression model for normalized biomarkers named "multivariate logistic curves model". This model is given by:

\[ y_{t=1} = \left( \frac{1}{1 - (c_1(t) + 1)} \left( \exp \left( \frac{a(t) + b(t)}{1 - (c_1(t) + 1)} \right) \right) \right) \]

where \( y_{t=1} \) is the measurement of the i-th biomarker for individual i, at time \( t=1 \).

Experimental results

Data: Normalized cognitive scores grouped into four categories (biomarkers): memory (5 items), language (5 items), praxis (2 items), concentration (1 item). Data collected from the ADNI database for 248 MCI patients who converted to AD. Each observation is a point in \( M = \mathbb{R}^N \).

Estimation of the parameters of the model

The parameters of the generic spatio-temporal model are \( \theta = (\mu_0, \mu_u, \alpha_1, \lambda_1, \delta, \tau_0, v_0, \alpha \), where (A).

Summary: the parameters are estimated using a stochastic version of the EM algorithm [Dempster, Laird, Rubin, 1977]. This algorithm is the Monte Carlo Markov Chain Stochastic Approximation EM algorithm (MCMC-SAEM) [Delyon et al., 1999; Allassonnière et al., 2010]. Theoretical results regarding the convergence of the algorithm have been proved in [Delyon et al., 1999; Allassonnière et al., 2010].

Note that the MCMC-SAEM requires that the model belongs to the curved exponential family. However, the multivariate logistic curves model does not belong to this family. The model can be made exponential by considering each parameter as realizations of independent Gaussian random variables.

Overview of the MCMC-SAEM for the multivariate logistic curves model:

- Initialisation: \( b = \theta_{(0)} \), \( x_{(0)} \rangle \), random, \( S = 0 \), \( (c_1(i), \delta(i)) \) sequence of positive step-sizes \( \delta \) until convergence.

Simulation (Hasting-Metropolis with Gibbs sampler):

- Compute the sufficient statistics \( S_{k} \) for each \( k \) : \( S_{k} = (u_{k}, \nu_{k}, \mu_{k}, \eta_{k}, \sigma_{k}) \) : \( k \) ).

Stochastic approximation:

- \( S_{k+1} = S_{k} + a_k \left( x_{k+1} \right) - S_{k} \) for all \( k \).

- Maximization: \( \alpha^{(l+1)} = \arg \max_{\alpha} \left( \psi(\theta) + \psi(\theta) + \psi(\theta) \right) \).

Regarding the scale of the biomarkers, we use the relative scale (log scale).

The city block metric is used for the distance measure.