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.
Introduction

- Understanding the progression of neuro-degenerative diseases
- We need to validate experimentally 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.

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 generic model is used to analyze the temporal progression of a family of univariate biomarkers.

The model:

\[ y_i(t) = \theta + \alpha_i + \tau_i(t) + \varepsilon_i \]

where \( y_i(t) \) is the observation for individual \( i \) at time \( t \), \( \theta \) is the population mean, \( \alpha_i \) is the individual random effect, \( \tau_i(t) \) is a subject-specific trajectory, and \( \varepsilon_i \) is the measurement error.

A hierarchical model:

\[ y_{ij} = \theta + \alpha_j + \tau_{ij} + \varepsilon_{ij} \]

Three particular cases of our generic spatio-temporal model:

- **Straight lines model** [Schirotte et al., 2015]
  \[ y_i(t) = \beta_0 + \beta_1t + \varepsilon_i \]
  where \( \beta_0 \) is the intercept, \( \beta_1 \) is the slope, and \( \varepsilon_i \) is the error term.
- **Logistic curves model** [Schirotte et al., 2015]
  \[ y_i(t) = \frac{1}{1 + e^{-\gamma t}} + \varepsilon_i \]
  where \( \gamma \) controls the steepness of the curve.\( y_i(t) \) is the observed value at time \( t \).

Estimation of the parameters of the model

The parameters of the generic spatio-temporal model are \( \theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4) \).

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

Note that the MCMC-EM 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-EM**
  \[ \theta^{(n+1)} \approx \theta^{(n)} \text{ random, } S = 0 \text{, } (\varepsilon_i) \text{ sequence of positive step-sizes} \]
  repeat until convergence

Simulation (Hastings-Metropolis within Gibbs sampler):

- Initialization: \( \theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4) \approx \text{random, } S = 0 \), \( (\varepsilon_i) \) sequence of positive step-sizes
- Repeat until convergence

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 \( M = [0,1]^4 \).
- Right: Barplot of the average quadratic error between the true labels and the labels predicted by the model at the \( n \)-th iteration. The Y-axis represents the number of times the model was chosen as the best over the \( K \)-fold cross-validation.
- Left: Plot of the individual quadratic error on the test set at iteration \( n \). The Y-axis represents the true label error. The X-axis represents the iteration index. The results are averaged over the \( K \)-fold cross-validation. The best models are those with the lowest average error.