Learning disease progression models with longitudinal data and missing values

Raphäel Couronné, Marie Vidailhet, Jean-Christophe Corvol, Stéphane Lehéricy, Stanley Durrleman

LEARNING DISEASE PROGRESSION MODELS WITH LONGITUDINAL DATA AND MISSING VALUES

Raphael Couronne [1,2]*, Marie Vidailhet* [2,3]*, Jean Christophe Corvol* [2,3]
Stephane Lehericy [2,4], *Stanley Durrleman* [1,2], *for the Alzheimers Disease Neuroimaging Initiative*

[1] Inria, Aramis project-team
[2] ICM, Inserm U 1127, CNRS UMR 7225, Sorbonne Universite, F-75013, Paris, France
[3] Department of Neurology, ICM
[4] Centre for Neuroimaging Research (CENIR), ICM

ABSTRACT

Statistical methods have been developed for the analysis of longitudinal data in neurodegenerative diseases. To cope with the lack of temporal markers - i.e. to account for subject-specific disease progression in regard to age - a common strategy consists in realigning the individual sequence data in time. Patient's specific trajectories can indeed be seen as spatiotemporal perturbations of the same normative disease trajectory. However, these models do not easily allow one to account for multimodal data, which more than often include missing values. Indeed, it is rare that imaging and clinical examinations for instance are performed at the same frequency in clinical protocols. Multimodal models also need to allow a different profile of progression for data with different structure and representation.

We propose to use a generative mixed effect model that considers the progression trajectories as curves on a Riemannian Manifold. We use the concept of product manifold to handle multimodal data, and leverage the generative aspect of our model to handle missing values. We assess the robusteness of our methods toward missing values frequency on both synthetic and real data. Finally we apply our model on a real-world dataset to model Parkinson's disease progression from data derived from clinical examination and imaging.

***Index Terms*—** Longitudinal, Missing values, Non-Linear Mixed Effect Model, Riemmanian Geometry, Multimodal, Disease modeling

1. INTRODUCTION

Linear mixed effect model estimated via EM have been introduced for the analysis of longitudinal data [1], and later were extended for more flexibility to the non-linear [2] case. Well adapted with an objective time (e.g. relative to an event), they are less adapted to data that do not include such consistent time event, such as neurodegenerative disease progression. In [3, 4] the concept of Time Warps is introduced to account for age variability at onset, and in [5] a morphological age-shift. However these Time Shifts are not estimated in the context of a statistical model. Generalization of LME to Riemannian manifolds were proposed [6, 7], that allows to consider features defined by smooth constraints, such as images or mesh [8]. In [9], a generic spatio-temporal model is introduced in the Bayesian framework, modeling the course of biomarker's progression as a geodesic, as well as individual variations via parallel transport, travelled at subject-specific onset and speed with an affine time reparametrization.

Although this approach allows mutivariate data, it assumes the same profile of progression (e.g. linear, logistic, exponential, etc..) for all coordinates, and does not account for missing values, leading to the removal of all visits with at least one missing values from the analysis, or to the use of ad-hoc data imputation procedure. This can be problematic for multimodal data where missing values (denoted NAs) occur by design of the experiment. We propose to build on this model to extend its application range, assuming missingness is unrelated to the data (Missing Completely at Random) [10]. We allow the modeling of the joint progression of features that are assumed to offer different evolution profile, and handle the missing values in the context of a generative model.

2. METHODS

2.1. The general model

In [9] each data point is seen as a point in a Riemmannian manifold, denoted $y_{i,j}$, observation of the i-th subject at its j-th visit. These points are then considered as noisy samples along an individual trajectory, namely a curve on the manifold, which in turn is seen as a random spatiotemporal transformation of a reference geodesic on the manifold. The model $y_{i,j}$ projected on the kth modality can be written as:

$$(y_{i,j})_k = (\eta^{w_i} (\boldsymbol{\gamma_0}) (\psi_i(t_{i,j})))_k + (\varepsilon_{i,j})_k \qquad (1)$$

where

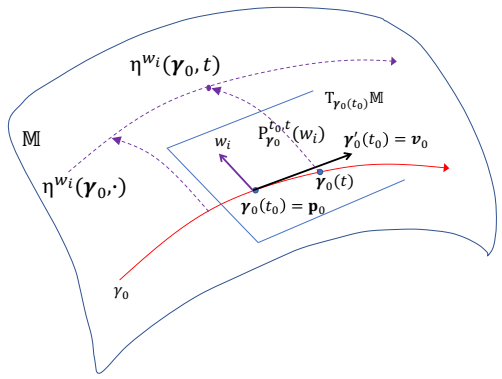

Fig. 1. Representation of the model on a schematic manifold. Each point $\eta^{w_i}(\boldsymbol{\gamma_0}, t)$ is obtained via the continuous transportation of the tangent vector w_i along the reference geodesic $\boldsymbol{\gamma_0}$ (in red). $\eta^{w_i}(\boldsymbol{\gamma_0}, .)$ (in purple) is a "parallel" of $\boldsymbol{\gamma_0}$. The model for patient i consists then in the trajectory $\eta^{w_i}(\boldsymbol{\gamma_0}, .)$ travelled at the subject-specific time $\psi_i(t)$.

- $\boldsymbol{\gamma_0} : t \to \text{Exp}_{p_0}^{t_0,t}(v_0)$ is the population average trajectory in the form a the geodesic passing at point p_0 with velocity v_0 at time t_0 (Exp denotes the Riemannian exponential as a concise way to write geodesics),

- $\eta^{w_i}(\boldsymbol{\gamma_0}) : t \to \text{ExpP}_{\boldsymbol{\gamma_0}}^{t_0,t}(w_i) = \text{ExpP}_{\boldsymbol{\gamma_0}}\left(P_{\boldsymbol{\gamma_0}}^{t_0,t}(w_i)\right)$ is the exp-parallelisation of the geodesic $\boldsymbol{\gamma_0}$ in the subject-specific direction w_i, called space-shift, as depicted in Fig 1. $(P_{\boldsymbol{\gamma_0}}^{t_0,t}(w_i)$ denotes the parallel transport of the vector w_i along the curve $\boldsymbol{\gamma_0}$ from $\boldsymbol{\gamma_0}(t_0)$ to $\boldsymbol{\gamma_0}(t))$,

- $\psi_i : t \to \alpha_i(t - t_0 - \tau_i) + t_0$ is a time-reparameterizing function, where α_i is a subject-specific acceleration factor and τ_i a subject-specific time-shift.

η^{w_i} and ψ_i define a spatiotemporal transformation of the average trajectory. To assure a unique decomposition through both the spatial and temporal transformation, the w_i are chosen orthogonal to $\mathbf{v_0}$ in the tangent space at $\mathbf{p_0}$. A spatiotemporal transformation of the reference trajectory to the ith patient trajectory is then parametrized by the individual parameters τ_i, α_i and w_i. A time-shift τ_i represents the delay at onset relative to t_0 for the individual i, to distinguish between individuals with early or late onset. The α_i models the speed at which the trajectory of individual i is travelled. Then the space-shifts w_i accounts for variations in position of the individual trajectory, and model difference in patters of disease progression between individuals. Normal distributions are chosen as priors for τ_i, w_i and ξ_i with $\alpha_i = \exp(\xi_i)$. These parameters are the random effects of the model, whereas $\boldsymbol{\gamma_0}$ is the fixed effect, parametrized by $\mathbf{p_0}$, $\mathbf{v_0}$ and t_0.

2.2. Manifold Product for multivariate Data

Dealing with a longitudinal and multimodal dataset, we wish to analyze at once the temporal progression of a family of N features, with possibly different evolution profile. Thus at the difference of [9] we consider manifold product that are not necessarily the product of the same univariate manifold M. Each feature k is described by repeated univariate observations $y_{i,j,k}$ on M_k that are considered as random perturbations along each trajectory. For each feature we choose a that defines a user-defined profile of progression (e.g. straight line, exponential decay, logistic). Ignoring missing values at the moment, each individual observations can be represented as a N-dimensional vector $(y_{i,j})_{1 \leq i \leq p}$, that is considered as random perturbation of quantities lying on the product manifold $\mathbb{M} = M_1 \times M_2 \times ... \times M_N$ equipped with the product metric. The product manifold gives geodesics of the form $\{\boldsymbol{\gamma} : t \in \mathbb{R} \to (\gamma_1(t), \gamma_2(t), ..., \gamma_N(t))\}$ on $\mathbb{M} = M_1 \times M_2 \times ... \times M_N$ equipped with the product metric. γ_k is the (univariate) geodesic which goes through point $p_k \in M_k$ at time t_0 and velocity v_k.

2.3. Missing Data

When missing values occur at time $t_{i,j}$, only a subset of $y_{i,j}$ if visible, we note $m_{i,j}$ these modalities. We decide to handle missing data in the context of our generative model, and compute the likelihood with visible data. The goodness of fit at a given visit $v_{i,j}$, at time $t_{i,j}$ for the kth modality writes $\|y_{i,j,k} - (\eta^{w_i}(\boldsymbol{\gamma_0})(\psi_i(t_{i,j})))_k\|^2$, and for the entire goodness of fit :

$$\mathcal{L}_{fit} = \sum_{i=0}^{p} \sum_{j=0}^{l_i} \sum_{k \in m_{i,j}} \|y_{i,j,k} - (\eta^{w_i}(\boldsymbol{\gamma_0})(\psi_i(t_{i,j})))_k\|^2 \quad (2)$$

with l_i the number of visits of the ith patient. We see in Eq 2 that the likelihood is informed only by available data while taking into account all the information available and without imputing missing values with ad-hoc procedures.

2.4. Estimation

Estimation of model parameters is done via the use of a stochastic version of the Expectation-Maximization Algorithm, namely the MCMC-SAEM algorithm [11], that seeks to maximize the likelihood $\mathcal{L} = \mathcal{L}_{fit} + \mathcal{L}_{prior}$. MCMC-SAEM iterates in 3 steps : simulation, approximation and maximization. It simulates first candidate individual variables, that are then accepted or rejected according to a probability function of the likelihood ratio. Then sufficient statistics are extracted from the current variables. Finally the current estimates of the parameters are maximized.

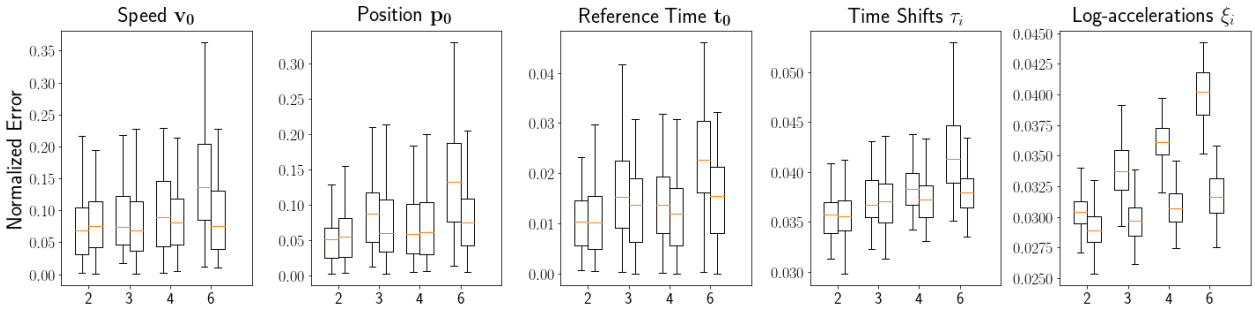

Fig. 2. Bootstrap distribution (b=100) of errors of estimation for population parameters ($\boldsymbol{p_0}$,$\boldsymbol{v_0}$,t_0) and individual parameters τ_i and ξ_i according to subsampling frequency on an artificial dataset. Parameter values are taken close to estimation on real-world data. The second modality is assigned to NAs at various frequency to compare performance worsening between the naive method (boxplots on the left), and the generative modeling method (boxplots on the right).

3. RESULTS

3.1. Experiment methodology

We propose to evaluate the method by pruning existing datasets and comparing the performance in the estimation between removing all visits with at least one missing value (naive method) or taking into account these missing values via our generative model (generative modeling method). From our experience that NAs occur mainly by design in neurodegenerative diseases datasets, we decide to prune the datasets by assigning chosen modalities to missing values at various (visit) frequencies.

In the 2 following experiments we use normalized scores on $M =]0, 1[$ with a metric ensuring that geodesics take the form of a logistic curve for each coordinate, so with $p_k = (\mathbf{p_0})_k$ and $v_k = (\mathbf{v_0})_k$ the multivariate model writes :

$$(y_{i,j})_k = (1+(\frac{1}{p_k}-1)\exp(-\frac{v_k\alpha_i(t_{i,j}-t_0-\tau_i)+(w_i)_k}{p_k(1-p_k)})^{-1}+(\epsilon_{i,j})_k \tag{3}$$

3.2. Synthetic Data

We produce synthetic data by simulating random-effects from their prior distribution and generating sample with the model. We generate a synthetic cohort of p=300 patients with 12 visits of 2 modalities each, occurring regularly on 4 years. Patient's age at beginning of the study are chosen arbitrarily as samples from a $\mathcal{N}(78, 5)$. We choose as initial parameters $\boldsymbol{p_0^*} = [0.4, 0.3]$, $t_0^* = 78$, $\boldsymbol{v_0^*} = [0.03, 0.04]$, $\sigma = 0.1$, $\sigma_\xi = 1$, $\sigma_\tau = 5$.

For each period in $[2, 3, 4, 6]$ we prune the dataset by assigning the second modality to a missing value every period of time, yielding datasets with patients that have respectively 6, 4, 3 and 2 visits with NAs. For each one of the obtained dataset, we bootstrap at the patient level the estimation procedure (6000 iterations) to obtain bootstrap distribution of relative estimation errors for both methods. The relative error is computed for each step of the bootstrap as followed : $\frac{\|\boldsymbol{v_0}-\boldsymbol{v_0^*}\|}{\|\boldsymbol{v_0^*}\|}$, $\frac{\|\boldsymbol{p_0}-\boldsymbol{p_0^*}\|}{\|\boldsymbol{p_0}*\|}$, $\frac{\|t_0-t_0^*\|}{\|t_0*\|}$ for the main population parameters, and $\frac{\|\xi-\xi^*\|}{\|\xi\|}$ and $\frac{\|\tau-\tau^*\|}{\|\tau\|}$ for individual parameters.

Results are reported in Fig 2. We observe that population parameter's estimation is quite robust to pruning, with a significant difference in performance only visible from period = 6 (2 visits per subjects). On the individual parameters the difference is more striking, the generative modeling approach showing more robustness toward pruning already with only 1 over 2 visits removed.

3.3. Real Data

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database. For up-to-date information, see www.adni-info.org.

We perform a similar experiment to a real dataset from ADNI1 cohort, consisting in 4 normalized neuropsychological test scores extracted from the ADAS-Cog, respectively associated with memory, language, praxis and concentration. Criteria for patient selection in ADNI1 was mild cognitive impairment at baseline and conversion to AD during the course of the study (MCI-converter), which led to 248 individuals. Patients are followed for an average of 6 visits and the dataset does not include any missing values. True parameters are not known, so we use as a proxy parameters estimated from a run of the estimation procedure on the entire dataset. Similarly to the previous experiment (6000 iterations, same estimation parameters), we subsample patients with NAs at various frequencies, discarding patients that are left with less than 2 visits without NAs and observe the bootstrap distribution of the resulting estimation error in Fig 3. Results show the same trend as with synthetic data, although estimation error is higher.

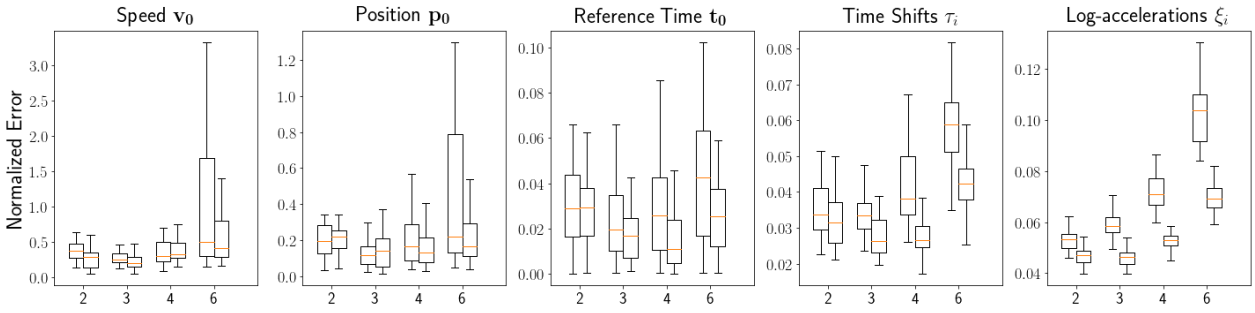

Fig. 3. Bootstrap distribution (b=100) of errors of estimation for population parameters ($\boldsymbol{p_0}$,$\boldsymbol{v_0}$,t_0) and individual parameters τ_i and ξ_i according to subsampling frequency on a real (ADNI) dataset. The 4 modalities used are subscores of the ADAS-COG accounting respectively for memory, praxis, language and concentration. 2nd, 3rd and 4th modalities are assigned to NAs at various frequency to compare performance worsening between the naive method and the generative modeling method.

3.4. Application to PPMI

Data used in the preparation of this article were obtained from the Parkinsons Progression Markers Initiative (PPMI) database. (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. From the PPMI cohort we extract 362 parkinsonian patients followed in average for 12 visits spread out on 4.6 years, yielding a total of 4441 visits. We model the joint progression of Parkinson's disease for 2 biomarkers, a motor score, namely the MDS-UPDRS part 3 (MDS) and an imaging score, the Right Caudate Striatal Binding Ratio (SBR). We normalize the data between 0 and the theoretical max for the motor score MDS-UPDRS part 3, and the empirical max for the Striatal Binding Ratio. We choose to prescribe a logistic profile for the motor score, as such assessments are designed to be sensitive to the transition from normal to disease state. By contrast, there is no such assumption for imaging data, that we assume to decay in a linear fashion. MDS contains 151 NAs, while SBR includes 3202 NAs. We run our multivariate sigmoid model to model the progression of these modalities at once (6000 iterations), which we represent in Fig 4, and obtain a resulting noise variance $\sigma_{MDS,SBR}$ is 0.00341, the same magnitude than the noise variance on univariate features only, with $\sigma_{MDS} = 0.00318$ and $\sigma_{SBR} = 0.00502$. We observe a positive correlation between the acceleration factor ξ_i and age at diagnostic ($p = 3.010^{-2}$), meaning subjects with later onset will progress faster. Furthermore, studying correlations with biological covariables we find the alpha-synuclein mean level to correlate with subject's onset ($p = 4.710^{-4}$).

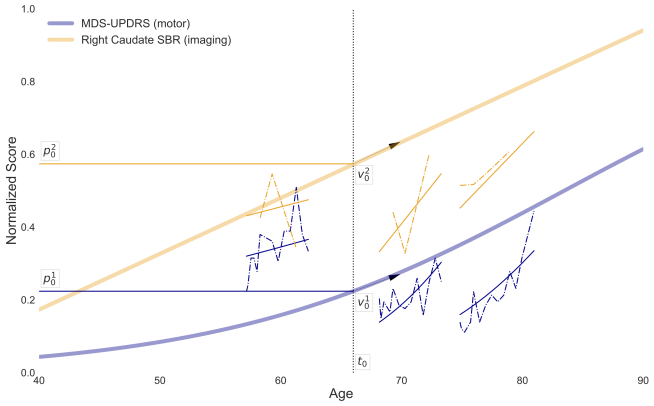

Fig. 4. In wide plain lines the mean geodesic estimated for PPMI PD patients on 2 modalities : MDS-UPDRS Score and Right Caudate SBR obtained from DatScan, and described by population parameters $\boldsymbol{p_0} = [p_0^0, p_0^1]$, $\boldsymbol{v_0} = [v_0^0, v_0^1]$ and t_0. The observations (in dotted lines) and individual models (in narrow lines) of 3 patients are also plotted.

4. CONCLUSION

We extended on a Bayesian non-linear mixed-effect model to allow the joint estimation of disease progression model on data with heterogeneous evolution profile. In practice such multimodal data include missing values by design. Instead of using ad-hoc method for data imputation, the generative statistical modeling allows to estimate model parameters by comparing generated data with observations only when they are available. Robustness analysis of the method is performed via the increasing pruning of existing dataset, while the variance of the performance is represented as a bootstrap distribution. The proposed method shows lower performance error in both synthetic and real-world data. This advocates for an extended use of the model, applicable to multimodal data with sparse design. We use thus our model to analyse the main PPMI modalities (motor, non-motor, imaging), and find that individidual parameters correlates with age of diagnosis and alpha-synuclein levels. **Acknowledgments.** This work has been partly funded by the European Research Council with

grant 678304, European Unions Horizon 2020 research and innovation program with grant 666992, and the program Investissements davenir ANR-10-IAIHU-06.

5. REFERENCES

[1] Nan M. Laird and James H. Ware, "Random-Effects Models for Longitudinal Data," *Biometrics*, vol. 38, no. 4, pp. 963, dec 1982.

[2] Mary J. Lindstrom and Douglas M. Bates, "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics*, vol. 46, no. 3, pp. 673, sep 1990.

[3] Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache, "Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 22–59, nov 2012.

[4] Yi Hong, Nikhil Singh, Roland Kwitt, and Marc Niethammer, "Time-Warped Geodesic Regression," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, pp. 105–112. Springer International Publishing, 2014.

[5] Marco Lorenzi, Xavier Pennec, Giovanni B. Frisoni, and Nicholas Ayache, "Disentangling normal aging from Alzheimers disease in structural magnetic resonance images," *Neurobiology of Aging*, vol. 36, pp. S42–S52, jan 2015.

[6] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P. Thomas Fletcher, "An efficient parallel algorithm for hierarchical geodesic models in diffeomorphisms," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. apr 2014, IEEE.

[7] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P. Thomas Fletcher, "A Hierarchical Geodesic Model for Diffeomorphic Longitudinal Shape Analysis," in *Lecture Notes in Computer Science*, pp. 560–571. Springer Berlin Heidelberg, 2013.

[8] Alexandre Bone, Olivier Colliot, and Stanley Durrleman, "Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms," in *CVPR 2018 - Computer Vision and Pattern Recognition 2018*, Salt Lake City, United States, June 2018.

[9] J.-B. Schiratti, , S. Allassonnière, A. Routier, O. Colliot, and S. Durrleman, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," in *Lecture Notes in Computer Science*, pp. 564–575. Springer International Publishing, 2015.

[10] Joseph G. Ibrahim and Geert Molenberghs, "Rejoinder on: Missing data methods in longitudinal studies: a review," *TEST*, vol. 18, no. 1, pp. 68–75, feb 2009.

[11] Estelle Kuhn and Marc Lavielle, "Coupling a stochastic approximation version of em with a mcmc procedure," vol. 8, 02 2004.