Mixture modeling for identifying subtypes in disease course mapping

Pierre-Emmanuel Poulet, Stanley Durrleman

▶ **To cite this version:**

Pierre-Emmanuel Poulet, Stanley Durrleman. Mixture modeling for identifying subtypes in disease course mapping. Aasa Feragen; Stefan Sommer; Julia Schnabel; Mads Nielsen. Information Processing for Medical Imaging, Springer, pp.571-582, 2021, 10.1007/978-3-030-78191-0_44 . hal-03276811v1

HAL Id: hal-03276811

https://inria.hal.science/hal-03276811v1

Submitted on 2 Jul 2021 (v1), last revised 22 Oct 2021 (v2)

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.

Mixture modeling for identifying subtypes in disease course mapping

Pierre-Emmanuel Poulet[1 [0000−0002−4423−2749]] and Stanley Durrleman[1 [0000−0002−9450−6920]⋆], for the Alzheimer's Disease Neuroimaging Initiative

Inria, Aramis project-team, Paris Brain Institute, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France pierre-emmanuel.poulet@inria.fr, stanley.durrleman@inria.fr

Abstract. Disease modeling techniques summarize the possible trajectories of progression from multimodal and longitudinal data. These techniques often assume that individuals form a homogeneous cluster, thus ignoring possible disease subtypes within the population. We extend a non-linear mixed-effect model used for disease course mapping with a mixture framework. We jointly estimate model parameters and subtypes with a tempered version of a stochastic approximation of the Expectation Maximisation algorithm. We show that our model recovers the ground truth parameters from synthetic data, in contrast to the naive solution consisting in *post hoc* clustering of individual parameters from a one-class model. Applications to Alzheimer's disease data allows the unsupervised identification of disease subtypes associated with distinct relationship between cognitive decline and progression of imaging and biological biomarkers.

Keywords: Mixture model · Non-linear mixed-effect model · Disease course mapping · Alzheimer's disease subtypes

1 Introduction

In the wake of medical progress and general increase of life expectancy, neurodegenerative diseases have seen a dramatic surge in the population. In order to better understand these diseases, our goal is to build digital models of their progression, which may span decades in the life of the patients. Such models may be estimated from longitudinal data sets of several patients at different disease stages. Data may include cognitive or behavioral assessments, biological biomarkers or image-based biomarkers such as regional brain volumes.

The statistical analysis of longitudinal data is often done in the framework of mixed-effects models [1]. Linear mixed-effects models (LMEM) are widely used, yet their application to medical observations is not adapted because of the non-linearity of the disease progression over large periods of time [2]. Nonlinear mixed-effects models (NLMEM) have been proposed in recent years [3, 4].

⋆ This research has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06. This work was also funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Mixed-effects models separate fixed effects, assumed to be common to the population, and individual effects, which are random variables describing how the model should change to accommodate for inter-individual differences. They are often assumed to follow a unimodal distribution, thus assuming the homogeneity of the population. However, this is often not the case as neurodegenerative diseases are known to be heterogeneous, with various subtypes that are difficult to characterize. Uncovering the underlying clusters of population is a complex task to perform in addition to an already complex NLMEM.

Our work contribution is a framework allowing for both unsupervised clustering and estimation of an NLMEM for each cluster. We propose a mixture model based on a nonlinear Riemannian mixed-effect model which we will refer to as a disease course mapping model [5]. The joint estimation of clusters and model parameters is performed using a mixture Monte-Carlo Markov chain stochastic approximation Expectation Maximisation (M-MCMC SAEM) algorithm [6]. This end-to-end approach differs from a more naive approach based on *a posteriori* clustering the parameters of the individual effects after the estimation of an NLMEM. We applied the proposed mixture model to the Alzheimer's disease neuroimaging initiative (ADNI) cohort. Our main focus is the separation of two obvious clusters: controls and patients diagnosed with Alzheimer. We then included the mild cognitive impaired (MCI) patients to understand their position relatively to the two previously found clusters.

2 Related work

Analysis of longitudinal data can be performed with various types of models. Discrete models include event-based models which estimate the temporal ordering of sequence of pathological events [7–9]. This approach has been extended with the SuStaIn model [10] to identify clusters in the population.

Continuous models include the linear mixed-effects model [1]. Several NLMEM have been proposed, one being the generalized linear framework [11] where the observations result of simple non-linear transformation of an LMEM. Typically the link function can be the logit, allowing for a sigmoid-shaped model. NLMEM include univariate models with time-reparameterizing functions [12]. Multivariate approaches include DIVE [4], a voxel-based model which also clusters disease trajectories, and disease course mapping which combines variations in progression dynamics with phenotypic differences [5,13–15]. In this approach, the observations belong to a Riemannian manifold and each individual trajectory in time is a parallel to a geodesic curve representing the population trajectory. Other models are also based on differential equation models [3]. Non parametric models using deep learning were also explored [16].

Estimation of complex parametric models such as NLMEM corresponds to the maximization of the likelihood. Such optimization can be performed with the Expectation Maximization (EM) and more especially its stochastic approximation variant with Monte-Carlo Markov chain (MCMC SAEM) which has been proven to have good theoretical properties [17–19]. Moreover the EM algorithm

is also a staple for mixture model estimation. Combining the estimation of the mixture model and the NLMEM has already been studied [6]. However the theoretical properties are often not enough as the MCMC SAEM tends, in practice, to suffer from local maxima attraction. In particular, changing cluster assignments is known to be challenging in such methods; it is a problem referred to as trapping states. Adding a tempering scheme has been shown to alleviate this issue by flattening the target distribution and thus easing the exploration of the parameter space [20]. A mixture model on top of a disease progression model was also proposed in [21], the clustering during estimation was handled with hard labels and a probability for individuals to switch from one cluster to another.

3 Method

3.1 Disease course mapping model

We present here the disease course mapping model first mentioned in [5] and improved in [13, 15]. We introduce the notations and essential equations for the rest of the article. In this work we focus on the particular case where the model takes the form of a series of logistic curves for each biomarker.

We assume a longitudinal dataset $(y_{ijk})_{1\leq i\leq n, 1\leq j\leq N_i, 1\leq k\leq d}$ where each patient i has N_i visits, at time t_{ij}, and d features observed at each visit. The number of visits N_i may vary from one patient to another. Data y_{ijk} are assumed to be points on a Riemannian manifold $\mathcal{M}$. This model is a mixed effect model in the following sense: we define the population parameters or fixed effects as a set of parameters describing an average population trajectory as a geodesic γ_0 in the Riemannian manifold, with $\gamma_0(t_0) = \mathbf{p}$ and $\dot{\gamma}_0(t_0) = \mathbf{v}$. The individual effects take into account a temporal effect with an individual reparametrization of time and space-shifts, also called inter-marker spacing parameters. Following the hypothesis of sigmoids for neurodegenerative disease biomarkers [2], we use the logistic variant of the disease course mapping model:

$$y_{ijk} = \left(1 + (\frac{1}{p_k} - 1)exp(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1 - p_k)})\right)^{-1} + \epsilon_{ijk} \quad (1)$$

with the noise $\epsilon_{ij} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2\mathbf{I}_d)$. The individual time reparametrization takes the following form: $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$, where τ_i is called the time-shift and models a straight delay or advance one individual can have on the average trajectory, and $\alpha_i = e^{\xi_i}$ is called the acceleration factor. This time reparametrization captures two phenomena: the fact that a patient can have an early or late disease onset, and the fact that a patient can be a slow or fast progressor. The space-shifts $(\mathbf{w}_i)_i$ have the same dimension as the observations, but for more interpretability the model uses an ICA decomposition with N_s independent sources $(\mathbf{s}_i)_{1\leq i\leq N_s}$. This leads to a formulation $\mathbf{w}_i = A\mathbf{s}_i$ such that the columns $A_l = \sum_{k=1}^{d-1} \beta_{lk} B_k$ are a linear combination of an orthonormal basis $(B_k)_{1\leq k\leq d-1}$ of the orthogonal hyperplane to $Span(\mathbf{v})$.

The hierarchical statistical model assumes that the population and individual parameters are latent, and follow Gaussian distributions directly or after transformation: $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$, $\tau_i \sim \mathcal{N}(0, \sigma_\tau^2)$ and $\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}_{N_s}, \mathbf{I}_{N_s})$; $g_k = \frac{1}{p_k} - 1$ and $\mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, \sigma_g^2 \mathbf{I}_d)$, $v_k = e^{\tilde{v_k}}$ and $\tilde{\mathbf{v}} \sim \mathcal{N}(\bar{\mathbf{v}}, \sigma_v^2 \mathbf{I}_d)$, $t_0 \sim \mathcal{N}(\bar{t_0}, \sigma_t^2)$, $\beta_{lk} \sim \mathcal{N}(\bar{\beta_{lk}}, \sigma_\beta^2)$. The individual parameters are noted $z_i = (\xi_i, \tau_i, \mathbf{w}_i)$. The population parameters, i.e. the fixed effects, are noted $z_{pop} = (\mathbf{g,\ v}, t_0, A)$. All the latent variables are noted $\mathbf{z} = (z_{pop}, (z_i)_{1 \leq i \leq n})$. Finally the statistical model parameters are $\theta = (\sigma_\xi, \sigma_\tau, \bar{\mathbf{g}}, \bar{\mathbf{v}}, \bar{t_0}, \bar{\beta_{lk}})$, while $\sigma_g, \sigma_t, \sigma_v, \sigma_\beta$ are fixed.

MCMC-SAEM The estimation of the parameters in the disease course mapping model is performed with the Monte Carlo Markov chain stochastic approximation variant of the Expectation Maximization algorithm. The convergence of the MCMC-SAEM has been proven [18] for the curved exponential family, which is the family of distributions for which the log-likelihood can be written as: $\forall \theta \in \Theta, log\ q(\mathbf{y,\ z}, \theta) = -\Phi(\theta) + \langle S(\mathbf{y,\ z}), g(\theta) \rangle$ where Φ and g are smooth functions, S are called the sufficient statistics. The sufficient statistics are to be understood as a summary of the required information from the latent variables $\mathbf{z}$ and the observations $\mathbf{y}$. The algorithm alternates between two steps:

- **Expectation**: latent parameters are estimated by a Metropolis-Hastings within Gibbs sampler algorithm. First new values $\mathbf{z}^*$ are sampled from a proposal law. Proposal value $\mathbf{z}^*$ is accepted over current value $\mathbf{z}_K$ with probability $1 \wedge \frac{q(\mathbf{y}, \mathbf{z}^* | \theta)}{q(\mathbf{y}, \mathbf{z}_K | \theta)}$. This Metropolis-Hastings scheme guarantees that the new value z_{K+1} is asymptotically sampled from the target distribution $q(\mathbf{y}, \cdot | \theta_K)$. The Gibbs sampler is used to sequentially estimate the population parameters z_{pop} and the individual parameters z_i. Based on the latent variables z_{K+1}, the sufficient statistics are computed, giving an approximation of $log\ q$
- **Maximization**: the model parameters θ_K are updated by maximizing the expectation of the log-likelihood $\theta_{K+1} = \underset{\theta \in \Theta}{argmax}\ \log(q(\mathbf{y,\ z}_{K+1}, \theta))$, which is computed in closed form as a function of the sufficient statistics only

3.2 Mixture of disease course mapping models

The improvement we brought to the disease course mapping model is a new layer atop of the hierarchical structure already built. If we write $q(\mathbf{y,\ z}, \theta)$ the likelihood of the model parametrized by θ, with latent variables $\mathbf{z}$ and observations $\mathbf{y}$, then the likelihood of a mixture model with L clusters is: $Q_{\theta_1, \ldots, \theta_L}(\mathbf{y,\ z}) = \sum_{c=1}^{L} \pi^c q(\mathbf{y,\ z}, \theta_c)$ where π^c denotes the probability of cluster c. Then each cluster c has its own set of parameters θ_c. We assume that each individual has a true latent class π_i and a set of individual parameters $\mathbf{z}_i^{\pi_i}$ corresponding to the model attached to the cluster π_i. The estimation algorithm is described in Algorithm 1. It is a mixture version of the MCMC-SAEM, where we add the latent variables π_i^c for the probability of each individual i to be in each cluster c, and $(z_i^c)_i$ are the individual parameters for an individual i in cluster c. Contrarily to [21] where the latent variable is directly π_i, forcing individuals to attach to a cluster, our use of

soft cluster labelling avoids cluster freeze and trapping states. By conditioning, we rewrite the likelihood of individual i in cluster c $q(y_i, z_i^c|\theta^c, z_{pop}^c)$ as the product of $q(y_i|\theta^c, z_{pop}^c, z_i^c)$ the likelihood of the observations and $p(z_i^c|\theta^c, z_{pop}^c)$ the likelihood of the individual parameters. These two terms are akin to the classical terms of the L2 loss (resulting from the choice of the Gaussian noise) and the regularization respectively.

Algorithm 1: M-MCMC-SAEM estimation

Initialization: $\pi_0, (\theta_0^c), (\mathbf{z}_{0,i}^c)$

for $K = 0...N$ **do**

Compute probability of individual i to belong to cluster c:

$\pi_{K+1,i}^c = \frac{\pi_K^c q(y_i|\theta_K^c, z_{K,i}^c, z_{K,pop}^c))p(z_{K,i}^c|\theta_K^c, z_{K,pop}^c))}{\sum_j \pi_K^j q(y_i|\theta_K^j, z_{K,i}^j, z_{K,pop}^j))p(z_{K,i}^j|\theta_K^j, z_{K,pop}^j))}$

for $c = 1...L$ **do**

• **E step**

Population parameters estimation

Sample $z_{*,pop}^c$ and compute acceptance ratio

$\alpha = 1 \wedge \frac{q(\mathbf{y}|\theta_K^c, (z_{K,i}^c)_i, z_{*,pop}^c)p((z_{K,i}^c)_i|\theta_K^c, z_{*,pop}^c)}{q(\mathbf{y}|\theta_K^c, (z_{K,i}^c)_i, z_{K,pop}^c)p((z_{K,i}^c)_i|\theta_K^c, z_{K,pop}^c}$

Set $z_{K+1,pop}^c = z_{*,pop}^c$ with probability α else $z_{K+1,pop}^c = z_{K,pop}^c$

Individual parameters estimation

Sample $(z_{*,i}^c)_i$ and compute acceptance ratios

$\alpha_i = 1 \wedge \frac{q(y_i|\theta_K^c, z_{*,i}^c, z_{K+1,pop}^c))p(z_{*,i}^c|\theta_K^c, z_{K+1,pop}^c)}{q(y_i|\theta_K^c, z_{K,i}^c, z_{K+1,pop}^c))p(z_{K,i}^c|\theta_K^c, z_{K+1,pop}^c)}$

$\forall i$, set $z_{K+1,i}^c = z_{*,i}^c$ with probability α_i else $z_{K+1,i}^c = z_{K,i}^c$

Compute sufficient statistics $(S^c(y_i, \mathbf{z}_{K+1}))_i$

• **M step**

Update $\theta_{K+1}^c = \underset{\theta \in \Theta}{argmax}\ log\ q(\mathbf{y}, z_{K+1,pop}^c, (z_{K+1,i}^c)_i, \theta)$

Compute $\pi_K^c = \frac{1}{n}\sum_i \pi_{K,i}^c$

end

end

The challenging part is the update of model parameters within each cluster. In the case of a single model, the updates of the parameters θ of the model are computed in a closed form and are only a function of the total sufficient statistic, which writes $S(\mathbf{y}, \mathbf{z}_{K+1}) = \sum_i S(y_i, \mathbf{z}_{K+1})$ since the observations $(y_i)_i$ are supposed independent. When mixtures are involved, the log-likelihood changes to $log\ q(\mathbf{y}, \mathbf{z}^c, \theta^c) = \sum_i log\ (\sum_c \pi_i^c q(y_i, z_i^c, z_{pop}^c, \theta^c))$. However since the EM computes an expectation of the log-likelihood, we can condition on the true latent class π_i of each individual i such that $\pi_i^c = \mathbf{P}(\pi_i = c)$:

$$\begin{aligned}
\mathbf{E}_{\mathbf{z}}\Big(\log\big(\sum_c \pi_i^c q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)\Big) &= \mathbf{E}_{\pi}\Big(\mathbf{E}_{\mathbf{z}}\big(\log\big(\sum_c \mathbf{1}_{\pi_i=c} q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)|\pi_i\big)\Big) \\
&= \mathbf{E}_{\pi}\Big(\mathbf{E}_{\mathbf{z}}\big(\sum_c \mathbf{1}_{\pi_i=c} \log\big(q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)|\pi_i\big)\Big) \\
&= \mathbf{E}_{\mathbf{z}}\big(\sum_c \pi_i^c \log\ (q(y_i, z_i^c, z_{pop}^c, \theta^c))\big)
\end{aligned}$$

Thus we obtain the total sufficient statistics for each cluster $S^c(\mathbf{y}, \mathbf{z}_{K+1}) = \sum_i \pi_i^c S(y_i, \mathbf{z}_{K+1})$ which can be used to update θ^c. This formula is intuitive: each sufficient statistic for each cluster weights the contribution of the individual data by the probability of this individual being in the cluster. This result also proves that the mixture model still belongs to the curved exponential family. Therefore, the convergence of this algorithm is guaranteed by the MCMC-SAEM proof [18].

3.3 Tempered scheme

Even with theoretical guarantees, the convergence might be very slow in practice. The estimation of individual parameters in a non-linear single model may be challenging. Building a mixture on top of it further adds to the difficulty. Practical experiences show a high reliability on initialization. One of the bottleneck is the amount of cluster regularization contained in the term $p(z_i^c|\theta^c, z_{pop}^c)$ in the likelihood. Until the cluster population parameters stabilize, we do not want to restrict the exploration of the individual parameters space.

We thus propose to use a tempered scheme for the Gibbs sampler, mimicking simulated annealing. Tempered MCMC-SAEM has been shown to converge [20]. In a tempered scheme, inverse temperature comes as a multiplier of the log-likelihood of the model. We realized our model had overly constraining regularization, outweighing the data term in the likelihood. Thus we applied temperature only to the regularization term of the log-likelihood in the sample. In practice, we propose to simply replace p by $\tilde{p} = p^{1/T}$ in Algorithm 1.

For the temperature scheme, we opted for a sinusoidal pattern with a decreasing hull (sine cardinal) following [20]: $T(\kappa) = 1 + b\frac{sin(\kappa)}{\kappa}$, $\kappa = \Delta + 2\pi\frac{K}{p}$, where b can be seen as the amplitude of the oscillations, Δ as a phase delay, and p as a period, K is the iteration number. The use of this tempered scheme allows for an alternation between exploratory phases (high temperature) and exploitation phases (low temperature). The values of the hyperparameters were empirically set to $b = 1, \Delta = 0, p = N_{iter}/100$ after several attempts.

3.4 Initialization method

With the same practical considerations, we had to find an initialization method which would improve the odds of convergence towards the global optimum. Random or manual initialization of the model parameters $(\theta_0^c)_c$ was not satisfying, as we will see in the experiments.

We opted for an initialization as close to the clusters as possible. We first fit a disease course mapping model (without mixture) on the whole data. This yields a set of parameters $(\theta_{init}, z_{pop,init}, (z_{i,init}))$. Then we use a Gaussian mixture model (GMM) on the estimated individual parameters $z_{i,init}$. The GMM allows us to identify L clusters within the individual data. The parameters of each mode c in the GMM combined with the population parameters θ_{init} produce new population parameters for a disease course mapping model which represent the mode: for instance the new t_0^c is the shift of $t_{0,init}$ by the mean value of τ_i

in the mode. This provides initialization parameters for as many disease course mapping models as there are modes in the GMM, and we use these initialization parameters for the clusters of the mixture model.

Please note that the use of a *post hoc* clustering on the estimated parameters is not equivalent to the joint estimation in the mixture model, so this initialization method does not directly give the right cluster parameters. Indeed, we can only produce clusters such that $\forall c, \mathbf{v}^c \in Span(\mathbf{v}_{init})$, which is a strong limit to the posterior analysis of a one-class model.

4 Results and Discussion

4.1 Simulated data

Univariate model We first show that the mixture model accurately recovers the ground truth parameters on simulated data. We generated data by creating two unimodal disease course mapping models with chosen population parameters. We fix the first model with $t_0 = 70$, $\sigma_\tau = 2$, $\sigma_\xi = 0.1$, $\log(v_0) = -2$, $\log(g_0) = 1$. For the second model we change the values of t_0 and v_0. We use the Kullback-Leibler divergence to determine the difference between the two model distributions. We then assume the two clusters have equal prevalence and we create 512 individuals, which consists in randomly attributing the individual to a cluster and sample individual parameters according to the distribution of the cluster. Next we arbitrarily decide the number (7 in average) and time of the "visits" for each individual, and we compute the values associated to these timepoints. Finally we add a small Gaussian noise ($\sigma = 0.01$) to the output.

A simplified version of the model consists in assuming that there is no sources in the ICA, leading to an univariate model with no space-shifts. In this case, the only individual parameters left are the time-related ones: $(\xi_i)_i$ and $(\tau_i)_i$. This special case of the model converges more easily since there are less parameters to estimate, which allows us to initialize the model without having to first fit a single model. We evaluated the class estimation with the area under the ROC curve (ROC AUC) metric. For the evaluation of population and individual parameters, we computed the absolute error between the ground truth parameters of each cluster and the parameters of the closest estimated cluster, this error being normalized by the standard deviation of the parameter. Figure 1 shows the reconstruction metrics on simulated data as a function of the Kullback-Leibler divergence between clusters. With the ROC AUC, we are able to see that the mixture perfectly separates two clusters, once the two clusters are different enough. The estimation errors for population parameters increases slightly as the clusters get farther from each other, which is understandable since the algorithm needs to do more exploration. The mean error on individual parameters shows a very stable curve. The reconstruction is limited by the noise of the generated data. In all the univariate experiments, the estimated parameters allowed the model to reach a L2 loss close to the noise level. The consistent errors for individual parameters seemingly correspond to the range in which individual parameters differences can be mistaken for noise.

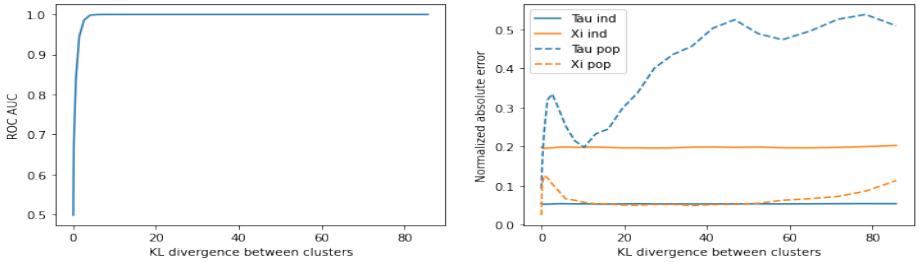

Fig. 1. Evaluation metrics for the reconstruction of true parameters. Left: ROC AUC for the estimated cluster probabilities of individuals π_i^c. Right: Absolute error between estimated parameters and ground truth, normalized by parameter standard deviation.

Multivariate model In the previous experiment, the results show consistent performance on a simplified univariate model. However when we add multiple sources, the number of parameters increases drastically and convergence might be more challenging in practice. We generated a dataset as in the first experiment but with two sources and three-dimensional observations. We first fitted a single disease course mapping model on the generated dataset, yielding $(\theta_{init}, z_{pop,init})$. We then fitted several mixture models to compare the different initialization methods, with or without the tempered scheme. Final results are shown in Table 1. Each model estimation takes about 5 minutes for 4,000 iterations which is the number of iterations required in this setup for the log-likelihood to stabilize.

The table shows that the tempered version allows for a better fit at the cost of regularization overall. Random initialization does not work. "Init" shows a better fit but is not able to cluster the individuals. "GMM" initialization is the best method in the absence of better heuristics.

Table 1. Performances of models. Single: single disease course mapping model; Random: random initialization; Init: initialization for both clusters at $(\theta_{init}, z_{pop,init})$; GMM: the initialization method described previously using a GMM on $(z_{i,init})_i$; True: perfect initialization at the ground truth parameters. Fit: log-likelihood related to the observations. Regularization: log-likelihood related to parameters' distribution.

Model	ROC AUC	Fit ($log\ q$)	Regularization ($log\ p$)	BIC
Single	-	42288	-3662	-38494
Random non-tempered	0.50	25922	-1439	-24212
Random tempered	0.50	36813	-3622	-32920
Init non-tempered	0.53	36523	-3637	-32615
Init tempered	0.56	40034	-3882	-35881
GMM non-tempered	**0.99**	**45787**	**-2914**	**-42602**
GMM tempered	**0.99**	**46871**	**-3068**	**-43532**
True non-tempered	1.00	53613	-2735	-50607
True tempered	1.00	53571	-2280	-51020

4.2 Applications on Alzheimer's disease data

Experiment without MCI We apply the method to the ADNI dataset [1]. We first build a model considering only stable Alzheimer's disease (AD) patients and stable controls. Our data was comprised of 400 AD (mean MMSE : 21.7, mean age : 75.9) and 695 controls (mean MMSE : 29, mean age : 75.8), with approximately 5 visits per patient. We select 8 features that are most relevant to the progression of AD based on a medical expert's advice, which included cognitive scores, MRI-derived regional volumes and biomarkers level. We then fit a single disease course mapping model on it. The posterior analysis of individual parameters highlights the need to take heterogeneity into account. The posterior distribution of the couple (τ_i, ξ_i) is shown in Fig. 2.

The GMM on the individual parameters provides two initial clusters for our mixture model. Optimal number of clusters was decided based on the Bayesian Information Criterion (BIC). The results of the mixture model are shown in Figure 3. The two clusters have very distinct average trajectories. The most obvious difference is the much earlier and rapid cognitive decline in cluster 2 compared to the cluster 1, while the progression of imaging and CSF biomarkers is only a few years earlier in cluster 2.

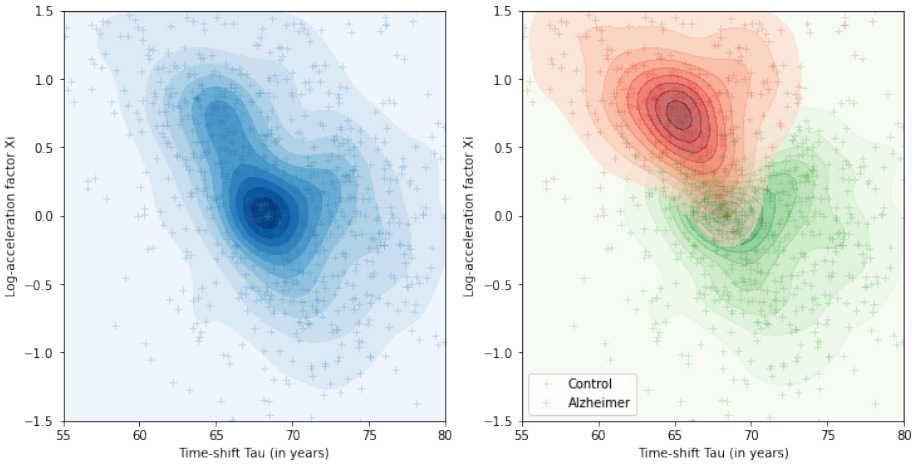

Fig. 2. Scatter plot of individual parameters, with kernel density estimation. Left: KDE on all AD and control patients. Right: KDE estimated separately for AD and controls.

Interestingly, the clustering does not correspond to AD and controls classes. Cluster 1 contains 88% of all controls and 65% of AD cases, meaning cluster 2 accounts for a minority of the data and contains mostly AD patients. Since average trajectory takes controls into account, we plotted the average trajectory

[1] http://adni.loni.usc.edu/

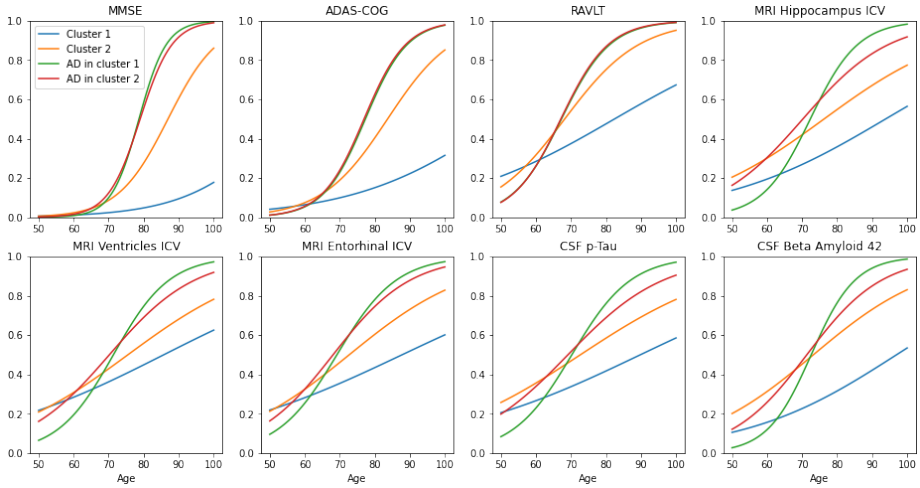

Fig. 3. Average trajectories in time. AD patients in both clusters have a similar cognitive evolution as shown by the cognitive scores (MMSE, RAVLT and ADAS-COG) whereas AD patients of cluster 1 (green) have a faster progression on MRI-based and CSF biomarkers than AD patients of cluster 2 (red).

of AD patients in each cluster in order to compare them. AD patients in both clusters show rapid cognitive degradation, while AD patients in the first cluster show a steeper increase of their imaging and CSF levels.

To understand why the mixture model did not separate AD patients from controls, we estimated one model on AD patients only and one model on controls only. We then computed the log-likelihood of our models. Results are shown in table 2. They confirm that the mixture model explains the variability seen in the data better than prior categorisation based on diagnosis. We further estimated a mixture with the initial clusters being AD and controls, and the algorithm still converged towards a similar version to the one presented in Figure 3.

Several studies report an association between the atrophy rates of particular brain regions and cognitive decline [22, 23], which lead to the identification of disease subtypes based on the differences in regional atrophies. Our analysis suggest that such associations are not systematic: similar pathological processes may lead to distinct pattern of cognitive decline, like similar cognitive decline may have distinct pathological processes.

Table 2. Log-likelihood of models

Model	Fit ($log\ q$)	Regularization ($log\ p$)	Total log-likelihood
Mixture	38100	-10391	27708
AD + Controls	37226	-9772	27454

Estimating MCI patients We performed individual personalization of MCI in the previous model, i.e. we estimated individual parameters including likelihood to belong to each cluster while keeping population parameters fixed. The distribution of MCI patients in the two clusters confirmed our hypothesis that cluster 2 is a specific subtype of AD: MCI associated to cluster 2 are mostly converters (66%) while 80% of non-converters are in cluster 1.

Conclusion We proposed a mixture for a disease course mapping model which has been validated on simulated data. We also introduced an heuristic initialization method to ensure convergence without extensive parameter search. The application to an Alzheimer's disease cohort suggests two subtypes of the disease associated with distinct relationship between cognitive decline and progression of imaging and CSF biomarkers.

References

1. Laird, N.M., Ware, J.H.: Random-effects models for longitudinal data. Biometrics **38**(4), 963–974 (Dec 1982)
2. Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, V.S., Donohue, M.C., Trojanowski, J.Q.: Update on hypothetical model of Alzheimer's disease biomarkers. Lancet Neurol **12**(2), 207–216 (Feb 2013)
3. Taddé, B.O., Jacqmin-Gadda, H., Dartigues, J.F., Commenges, D., Proust-Lima, C.: Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease. Biometrics **76**(3), 886–899 (2020)
4. Marinescu, R.V., Eshaghi, A., Lorenzi, M., Young, A.L., Oxtoby, N.P., Garbarino, S., Crutch, S.J., Alexander, D.C.: DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. NeuroImage **192**
5. Schiratti, J.B., Allassonnière, S., Colliot, O., Durrleman, S.: A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. The Journal of Machine Learning Research 18 (Jan 2017)
6. Lavielle, M., Mbogning, C.: An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models. Statistics and Computing **24**(5), 693–707 (Sep 2014)
7. Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., Alexander, D.C.: An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. Neuroimage **60**(3), 1880–1889 (Apr 2012)
8. Archetti, D., Ingala, S., Venkatraghavan, V., Wottschel, V., Young, A.L., Bellio, M., Bron, E.E., Klein, S., Barkhof, F., Alexander, D.C., Oxtoby, N.P., Frisoni, G.B., Redolfi, A.: Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. NeuroImage: Clinical **24**, 101954 (Jan 2019)
9. Bilgel, M., Jedynak, B.M.: Predicting time to dementia using a quantitative template of disease progression. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **11**(1), 205–215 (2019)

10. Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., van Swieten, J., Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M.C., Rowe, J.B., Graff, C., Tagliavini, F., Frisoni, G.B., Laforce, R., Finger, E., de Mendonça, A., Sorbi, S., Warren, J.D., Crutch, S., Fox, N.C., Ourselin, S., Schott, J.M., Rohrer, J.D., Alexander, D.C.: Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nature Communications **9** (Oct 2018)
11. McCullagh, P.: Generalized Linear Models. Routledge (Oct 2018)
12. Jedynak, B.M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B.T., Raunig, D., Jedynak, C.P., Caffo, B., Prince, J.L.: A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort. Neuroimage **63**(3), 1478–1486 (Nov 2012)
13. Couronné, R., Vidailhet, M., Corvol, J.C., Lehéricy, S., Durrleman, S.: Learning disease progression models with longitudinal data and missing values. In: ISBI 2019 - International Symposium on Biomedical Imaging (Apr 2019)
14. Louis, M., Couronné, R., Koval, I., Charlier, B., Durrleman, S.: Riemannian Geometry Learning for Disease Progression Modelling. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) Information Processing in Medical Imaging, vol. 11492, pp. 542–553. Springer International Publishing, Cham (2019)
15. Koval, I., Schiratti, J.B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., Durrleman, S.: Statistical Learning of Spatiotemporal Patterns from Longitudinal Manifold-Valued Networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) Medical Image Computing and Computer Assisted Intervention MICCAI 2017, vol. 10433, pp. 451–459. Cham (2017)
16. Mehdipour Ghazi, M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L.: Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. Medical Image Analysis **53**, 39–46 (Apr 2019)
17. Delyon, B., Lavielle, M., Moulines, E.: Convergence of a Stochastic Approximation Version of the EM Algorithm. The Annals of Statistics **27**(1), 94–128 (1999)
18. Kuhn, E., Lavielle, M.: Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM: Probability and Statistics **8**, 115–131 (2004)
19. Allassonnière, S., Kuhn, E., Trouvé, A.: Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study. Bernoulli **16**(1), 641–678 (2010)
20. Allassonniere, S., Chevallier, J., Oudard, S.: Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 1152–1160. Curran Associates, Inc. (2017)
21. Debavelaere, V., Durrleman, S., Allassonnière, S.: Learning the Clustering of Longitudinal Shape Data Sets into a Mixture of Independent or Branching Trajectories. Int J Comput Vis (Jun 2020)
22. Zhang, X., Mormino, E.C., Sun, N., Sperling, R.A., Sabuncu, M.R., Yeo, B.T.T., Initiative, t.A.D.N.: Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. PNAS **113**(42)
23. Risacher, S.L., Anderson, W.H., Charil, A., Castelluccio, P.F., Shcherbinin, S., Saykin, A.J., Schwarz, A.J., For the Alzheimer's Disease Neuroimaging Initiative: Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline. Neurology **89**(21), 2176–2186 (Nov 2017)