Covariate-Aware Longitudinal Modelling for Neurodegenerative Diseases

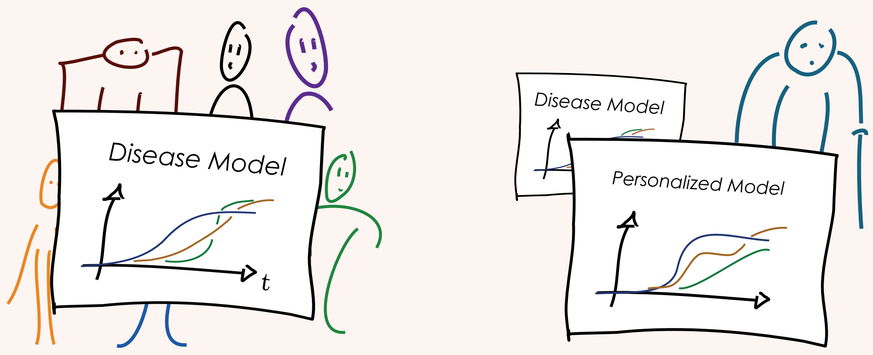
Nemo Fournier, Stanley Durrleman

Covariate-Aware Longitudinal Modelling for Neurodegenerative Diseases

Nemo Fournier, Stanley Durrleman
Aramis project-team, Inria, Sorbonne Université, Paris Brain Institute, CNRS, Inserm, AP-HP

Introduction

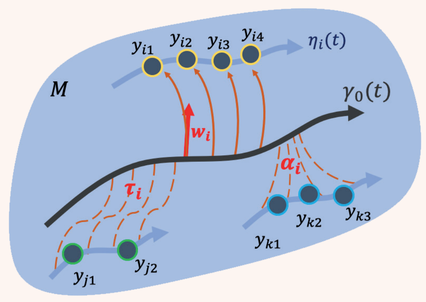
Longitudinal modelling is of pivotal interest for the study of neurodegenerative diseases. The Disease Course Mapping[1] is a multivariate Bayesian mixed-effect progression model that is able to recover the course of a disease from a cohort with multimodal longitudinal observations (imaging variables, cognitive and clinical scores) and to extract interpretable parameters to describe each patient. It has been validated on multiple diseases and on multiple applications settings (cohort study, trial enrichment, data simulation, ...)
However, its statistical formulation relies only on time-dependent observations. It thus fail to integrate time-independent information (gender, education levels, genetic factors, ...) in its modelling, even though such covariates are known to modulate clinical disease courses. We propose a mixed-effect formulation that captures the influence of such covariates over the dynamic of the disease.



Leaspy synthetizes longitudinal observations from a cohort into a descriptive model (left). Once presented with observations from a new patient, it finds descriptive parameters that register the population model onto the observed data, thus obtaining a personalized progression model. Priors learned during the initial calibration over the cohort help regularizing the model.
While performing this registration of the population model onto the personalized model, no information about covariates are taken into account. We propose to change this behaviour.

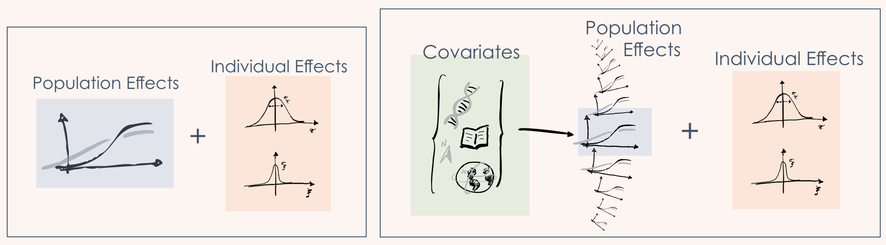
Methods

During calibration, the original model learns a set of population parameters characterizing the average progression of the pathology. Averaging is defined as learning a geodesic in a multi-dimensional Riemannian space which minimizes the cost of non-linear identifiable registrations of the average curve onto all individual observations. A Bayesian framework is used in order to include regularize registration costs: each registration is driven both by individual data attachment and information learned from the whole cohort.



Learning a geodesic minimizing the cost of registration toward observations (dots). The identifiable registration involve a time reparametrization of the patients age (τ and α) and spatial effects (w). The ambient metric is chosen so that geodesics respect clinical hypothesis behind biomarker evolution — monotonicity, logistic-like dynamic, etc. (Figure from the Leaspy gitlab repository)

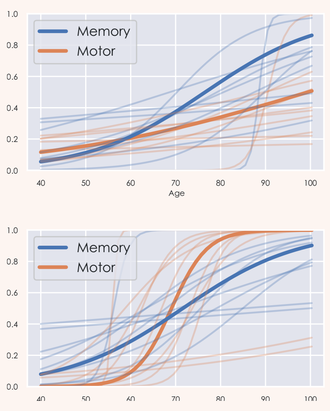
We adapt this model to explicit the dependancy between covariates and biomarker evolution. We learn a parametrized function linking covariates to population parameters. While the original model learned a fixed average trajectory, we now learn a function that associates a set of covariates, to a proposal average trajectory. Individual effects are retained to account for patient variability even within a "similar covariate class".
This link function is modelled as a linear map between covariates and the geodesic parameters. We thus obtain an interpretable link between each covariate and the dynamic of the pathology (such as progression speed for each biomarker, onset age, etc). The total likelihood of the model thereby formulated can be shown to lie in the curved exponential family. This allows using algorithms of the MCMC-SAEM family for calibration.



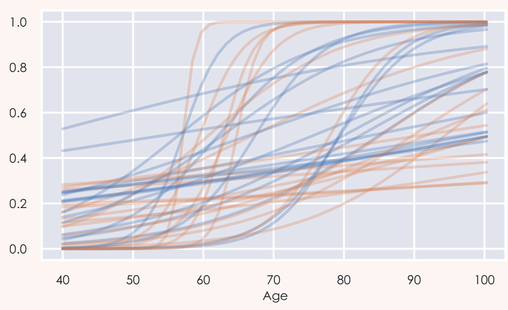
*Comparison between the original approach (**left**) and our new formulation (**right**).*
The link function of the new model diagrammatically corresponds to the arrow between the green and the blue areas.

Validation on Synthetic Data

Our model has been thoroughly validated on synthetic data. We generated data via synthetic disease models. The influence of covariates was simulated by instanciating multiple models with slightly different dynamics. For each subject we first sampled its time-fixed covariates, and then used them to select the model for data generation. Continuous covariates were modeled by combining slightly different models. We also included covariates irrelevant to the disease dynamics to validate the robustness of our approach.



*For example, we consider a fictive disease having two different progression patterns, depending on the status of a given mutation (binary covariate). We use two distinct generative models (bold line) to model a standard form (**top**) and a motor form of the disease (**down**). We consider a binary covariate (think: mutation status) that defines whether a patient suffers from the standard (no mutation) or the motor form (mutation) of the pathology.*

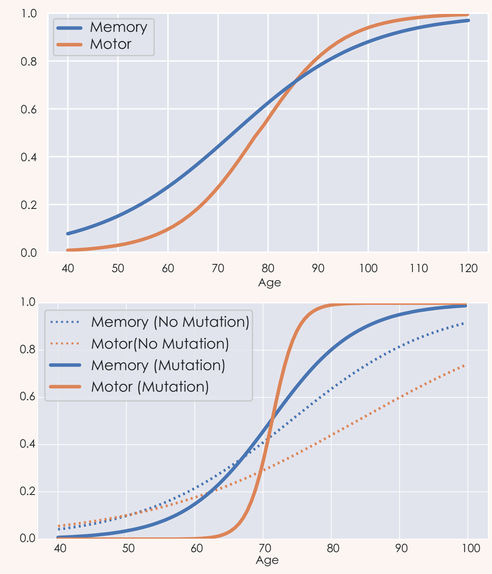


Trajectories of patients are pooled into a single synthetic cohort, where two disease modes coexist and depends on a mutation status.

We then calibrated our new model on such synthetic datasets and made sure that we could recover the link between covariates and dynamic effects. The example presented here is a simple binary setting, but multiple, continuous and slighter dynamics effects were also considered with success. Covariates that were artificially added but without any dynamic effects were also correctly discarded by our model.

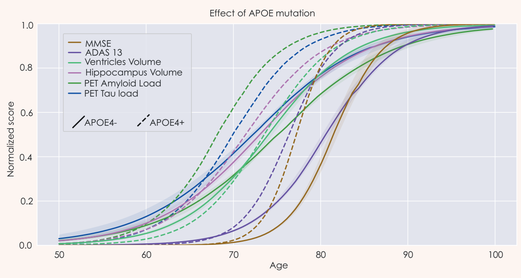
***Top**: average progression from a standard model calibrated on this heterogenous cohort.*
***Bottom**: average progression from our covariate aware model.*

The standard model averaged the cohort onto an intermediate dynamic that does not account for the heterogeneity of the cohort, while our new model was able to recover the dynamic effect that was associated to the mutation status.

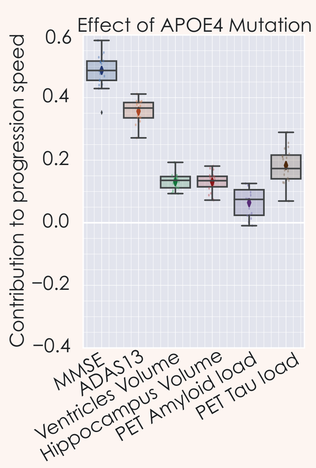


Results on Clinical Data

We ran our model on the ADNI dataset, which is an Alzheimer's Disease cohort. We showed that we can recover clinically established effects of covariates such as sex, APOE mutation.

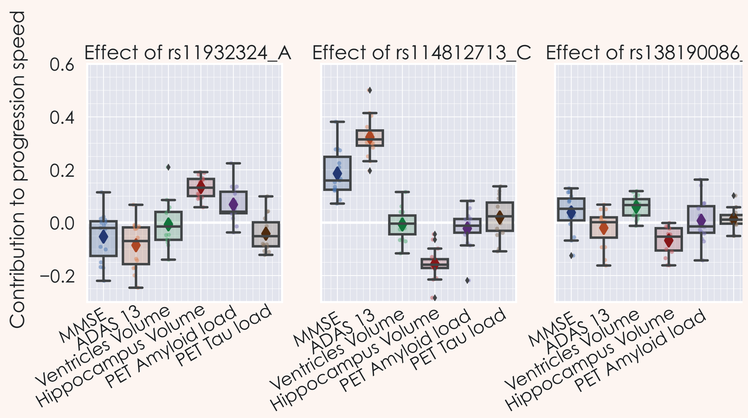


*Calibration over ADNI. Here we emphasize the effect of the APOE4 mutation. In **solid line** the average progression for patients without the APOE4 mutation. The **dotted line** shows the expected progression of a patient homozygous for the APOE4 mutation. We recover the clinically known effect of the mutation with patients affected earlier and faster.*



Analysis of the link function learned during calibration. Here is represented the effect of the APOE4 mutation on the progression speed of some biomarkers and cognitive scores. A positive value is associated to an «accelerating» effect, while a negative value stands for a «slowing» effect.

We then ran our model on a wider number of SNP selected using a reference GWAS[2]. Our findings suggest that even though those SNPs have some of the highest effect size in the GWAS, their association with dynamics effect of the disease — such as speed of progression — is not always significant. This is because in most standard GWAS, SNP are associated with diagnosis and not with dynamic aspects of pathologies.
This suggests that polygenic risk scores derived from regular GWAS do not efficiently target and inform about disease dynamics and that our approach could be a first step toward designing finer risk scores.



*Analysis of SNP using the same principle as for the APOE4 example. Those SNP are selected among some of the most associated SNP from a reference GWAS. The **first** and **third** SNP show no significant effect on the disease progression speed, while the **second** one is significantly associated to a faster cognitive decline.*

[1] *[Schiratti et al. — 2017]*
A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. JMLR
https://gitlab.com/icm-institute/aramislab/leaspy/

[2] *[Kunkle et al — 2019]*
Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nature genetics.