Identification supervisée et non supervisée de sous-groupes à l'aide d'un modèle de progression des maladies (Leaspy)

Sofia Kaisaridi, Gabrielle Casimiro, Juliette Ortholand, Sophie Tezenas Du Montcel

IDENTIFICATION SUPERVISÉE ET NON SUPERVISÉE DE SOUS-GROUPES À L'AIDE D'UN MODÈLE DE PROGRESSION DES MALADIES (Leaspy)

Sofia Kaisaridi [1], Gabrielle Casimiro [1], Juliette Ortholand [1] & Sophie Tezenas du Montcel [1]

[1] *ARAMIS, Sorbonne Université, Institut du Cerveau-Paris Brain Institute-ICM, CNRS, Inria, Inserm, AP-HP, Groupe Hospitalier Sorbonne Université, Paris, France*

sofia.kaisaridi@icm-institute.org, gabrielle.casimiro@icm-institute.org, juliette.ortholand@icm-institute.org, sophie.tezenas@aphp.fr

Résumé. Les modèles de progression des maladies sont des outils prometteurs pour analyser des données longitudinales présentant de multiples modalités. De tels modèles peuvent être utilisés pour estimer la progression de la maladie à long terme et reconstruire des trajectoires individuelles qui peuvent rendre compte de la variabilité entre les patients, mais aussi entre les modalités. Si les modèles classiques modélisent cette variabilité inter-patients comme des perturbations aléatoires autour d'une référence fixe, ceci n'est pas forcément réaliste si des facteurs externes expliquent une partie de la variabilité des trajectoires. Les mutations génétiques, le sexe, les antécédents familiaux, le niveau d'éducation, le statut socio-économique e.t.c. peuvent influencer l'évolution de certaines pathologies. De manière supervisé, nous pouvons répondre au défi de l'identification de sous-groupes d'évolution, par un conditionnement de notre modèle sur les covariables observés à l'inclusion. En revanche, dans certaines maladies chroniques où les mécanismes sous-jacents sont peu étudiés, nous ne disposons pas d'informations suffisantes sur les facteurs externes qui influencent la progression de la maladie. Par conséquent, l'identification de sous-groupes devient un défi plus complexe qui peut être abordé par une approche non-supervisée, basée sur un mélange probabiliste des paramètres du modèle. Dans cette étude nous utilisons un modèle de progression des maladies, non-linéaire à effets mixtes qui passe par une reparamétrisation du temps (*Disease Course Mapping Model* implémenté dans la librairie Python open-source Leaspy) et nous comparons les deux approches. Nous utilisons des donnés de la maladie CADASIL comme notre exemple motivant. Avec le modèle à covariables (approche supervisée) nous allons modéliser le niveau de scores cliniques en fonction de la position de la mutation. Avec le modèle de mélange (approche non-supervisée), différents sous-groupes de progression sont définis à partir uniquement de l' évolution des caractéristiques de la maladie. Une fois que les sous-groupes sont formés nous allons examiner l'effet de la position de la mutation sur leur formation.

Mots-clés. Modélisation de la progression de la maladie, Effets mixtes, MCMC-SAEM, Covariables, Mélange, CADASIL, Mutation

Abstract. Disease progression models are promising tools for analyzing longitudinal data with multiple modalities. Such models can be used to estimate long-term disease progression and reconstruct individual trajectories which can account for variability between patients, but also between modalities. While conventional models model this inter-patient variability as random variations around a fixed reference, this is not necessarily realistic if external factors explain some of the variability in trajectories. Genetic mutations, gender, family history,

level of education, socio-economic status and so on can influence the evolution of certain pathologies. In a supervised way, we can answer the challenge of identifying subgroups of evolution by conditioning our model on the baseline covariates. On the other hand, in certain chronic diseases where the underlying mechanisms are little studied, we do not have sufficient information on the external factors that influence disease progression. As a result, identifying subgroups becomes a more complex task, which can be tackled by an unsupervised approach, based on a probabilistic mixture of model parameters. In this study, we use a non-linear mixed-effects disease progression model that uses time reparametrisation (Disease Course Mapping Model implemented in the Leaspy open-source Python library) and we compare the two approaches. We use data from the CADASIL disease as our motivating example. with the covariate model (supervised approach) we will model the level of clinical scores as a function of mutation position. With the mixture model (unsupervised approach), different progression subgroups are defined solely on the basis of the evolution in disease characteristics. Once the subgroups have been formed, we will examine the effect of the position of the mutation on their formation.

Keywords. Disease Progression Modeling, Mixed-effects, MCMC-SAEM, Covariates, Mixture, CADASIL, Mutation

1 Introduction

Les modèles à effets mixtes sont prometteurs pour analyser des données répétées représentant plusieurs caractéristiques et ils sont souvent utilisés dans les études épidémiologiques. De tels modèles peuvent être utilisés pour estimer une progression à long terme d'une maladie et reconstruire des trajectoires individuelles qui tiennent compte de la variabilité entre les patients (effets aléatoires), mais aussi entre les modalités (effets fixes). Comprendre comment la maladie progresse et quelle est la variabilité attendue entre les individus est essentiel.

Si les modèles classiques modélisent cette variabilité inter-patients comme des perturbations aléatoires autour d'une référence fixe ceci n'est pas forcément réaliste car une partie s'explique parfois par des facteurs externes (et est donc difficilement expliquée uniquement par des perturbations aléatoires). Les mutations génétiques ou des facteurs externes tels que le sexe, les antécédents familiaux, le niveau d'éducation ou le statut socio-économique peuvent influencer l'évolution de certaines pathologies. D'un autre point de vue, dans certaines maladies chroniques où les mécanismes sous-jacents sont peu étudiés, nous disposons de peu ou pas d'intuition sur les covariables à l'inclusion qui influencent la progression de la maladie. Par conséquent, l'identification de sous-groupes devient un défi plus complexe, pouvant être abordé par un mélange probabiliste des paramètres du modèle.

Dans ce cadre, les modèles de progression de maladie (*Disease Progression Models*) sont des outils émergents qui reconstruisent les chronologies des maladies chroniques à long terme, fournissant ainsi un aperçu unique des processus pathologiques et de leurs mécanismes sous-jacents (Young et al. (2024)). Ils sont interprétables, facilitant ainsi la compréhension des maladies. Dans cette étude nous avons utilisé un modèle de progression des maladies, non-linéaire à effets mixtes, qui passe par une reparamétrisation du temps, appelé *Disease Course Mapping Model*, comme proposé par Schirrati et al. (2017). Les deux approches décrites ci-

dessus peuvent être considérées comme des extensions du modèle DCM, correspondant respectivement à un modèle tenant compte des covariables et à un modèle de mélange gaussien. Nous étendons le modèle de mélange existant pour inclure tous les aspects de la variabilité spatio-temporelle. L'objectif de ce travail est d'éclairer la manière dont ces deux techniques peuvent être modélisées dans le cadre du modèle DCM et de comparer leurs résultats. Les deux répondent à la même problématique : l'identification de différents sous-groupes de progression de la maladie ainsi que les facteurs qui les caractérisent. Nous utilisons comme exemple motivant les données de la cohorte de patients atteints de CADASIL recrutés au Centre national de référence français CERVCO.

2 Disease Course Mapping Model

2.1 Modèle générique

Nous considérons des données longitudinales, c'est-à-dire des patients avec des mesures répétées dans le temps. Le modèle de progression utilisé, c'est-à-dire le modèle DCM implémenté dans une librairie Python open-source (Leaspy), a déjà fait ses preuves pour décrire la progression des maladies de façon multivarié, et a déjà été appliqué aux diverses maladies neurodégénératives (Kaisaridi et al. (2025), Ortholand et al. (2023), Maheux et al. (2023), Poulet et Durrleman (2023), Moulaire et al. (2023) Koval et al. (2022)).

Nous supposons un ensemble de données longitudinales $(y_{ijk})_{1\leq i\leq n, 1\leq j\leq N_i, 1\leq k\leq d}$ qui decrivent les valeurs pour chaque patient i à la visite j de la caractéristique k. Le nombre total de visites N_i peut être différent pour chaque patient i. Les données y_{ijk} sont supposées être des points sur une variété riemannienne $\mathcal{M}$. Les paramètres de population (effets fixes) décrivent la trajectoire moyenne comme une géodésique γ_0 sur la $\mathcal{M}$ avec $\gamma_0(t_0) = \mathbf{p}$ et $\dot{\gamma_0}(t_0) = \mathbf{v}$. Ici nous utilisons la forme logistique du modèle :

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

avec une erreur résiduelle $\epsilon_{ij} \sim \mathcal{N}(\mathbf{O}_d, \sigma_k^2 \mathbf{I}_d)$. Pour simplifier, notre modèle construit la trajectoire moyenne de chaque score sous la forme d'une courbe logistique décrite par les paramètres p_0, v_0 et t_0, où p_0 et v_0 sont la position et la vitesse (dérivée de la courbe) au temps t_0, le point médian de la logistique. La reparamétrisation du temps individuel prend la forme suivante : $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ où τ_i est le décalage temporel et $\alpha_i = e^{\xi_i}$ est le facteur d'accélération. Le profil temporel des patients est décrit par deux paramètres : le décalage temporel τ_i qui correspond à l'âge estimé du début de la maladie, mesuré en années, et le taux de progression ξ_i qui indique si la progression globale est accélérée (valeurs positives) ou ralentie (valeurs négatives). Le profil spatial est défini par les paramètres d'espacement intermarqueurs ω_{ik} qui tiennent compte de l'ordre différent des événements au sein de notre population. Pour chaque patient, nous avons un ω pour chaque score clinique inclus dans le modèle avec des valeurs négatives indiquant qu'un score spécifique commence à se détériorer plus tôt pour un patient spécifique que pour la population moyenne. Néanmoins pour une interprétabilité accrue, le modèle utilise une analyse en composantes indépendantes avec N_s

sources indépendantes, ce qui conduit à des paramètres de décalage spatial $\mathbf{w}_i = As_i$ tels que les colonnes $\mathbf{A}_{:,l} = \sum_{m=1}^{d-1} \beta_{ml}\mathcal{B}_m$ sont une combinaison linéaire d'une base orthonormale $(\mathcal{B}_m)_{1\leq m\leq d-1}$ du sous-espace orthogonal à $Span(\mathbf{v})$. Le modèle statistique hiérarchique suppose que les paramètres de population $\mathbf{z}_{pop} = (\tilde{g_k}, \tilde{v_k}, \beta_{ml})$ et les paramètres individuels $z_i = (\xi_i, \tau_i, s_{il})$ sont latents et suivent des distributions gaussiennes :

$$\begin{cases} \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \\ \tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2) \\ s_{il} \sim \mathcal{N}(0, 1) \end{cases} \qquad (2) \qquad \begin{cases} \tilde{g_k} \sim \mathcal{N}(\overline{\tilde{g_k}}, \sigma_{\tilde{g}}^2) \\ \tilde{v_k} \sim \mathcal{N}(\overline{\tilde{v_k}}, \sigma_{\tilde{v}}^2) \\ \beta_{ml} \sim \mathcal{N}(\overline{\beta_{ml}}, \sigma_\beta^2) \end{cases} \qquad (3)$$

Enfin les paramètres du modèle statistique sont $\theta = (\overline{\tilde{g}}_k, \overline{\tilde{v}}_k, \overline{\beta}_{ml}, \overline{\tau}, \sigma_\tau, \sigma_\xi, \sigma_k)$ tandis que $\sigma_{\tilde{g}}$, $\sigma_{\tilde{v}}$, σ_β sont fixes.

2.2 Deux approches pour l'identification de sous-groupes

Le modèle à covariables définit les sous-groupes de manière déterministe et alors, il est susceptible de nous fournir des estimations plus stables et plus précises sur l'évolution de la maladie dans les différents groupes. En revanche, il exige une connaissance préalable des facteurs qui affectent la progression de la maladie, ce qui n'est pas toujours le cas. De l'autre côté le modèle de mélange, ne nécessitant aucune connaissance a priori, a un champ d'application plus large et conclut à l'identification de sous-types présentant des profils de progression similaires, après avoir examiné une série de répartitions possibles. Pourtant sa nature probabiliste s'accompagne d'un coût computationnel plus élevé, et pourrait nous amener à une convergence plus lente. Nous allons décrire les deux méthodes ci-dessous.

2.2.1 Approche supervisée : Modèle à covariables

Au lieu d'estimer des effets fixes (paramétrant l'évolution moyenne de la maladie) ainsi que des effets aléatoires, nous introduisons une fonction lien f_ϕ capable de prédire, à partir d'un ensemble de covariables c_i, une trajectoire attendue de la maladie conditionnée par ces covariables (Fournier et Durrleman (2023)). La principale différence par rapport à l'approche standard réside dans le fait que les effets fixes, précédemment introduits, sont désormais estimés comme une fonction déterministe de covariables $f_\phi(c_i)$. Dans ce cas, les covariables sont utilisées de manière supervisée lors de la calibration du modèle et servent à naviguer à travers un continuum de modèles de maladies, au lieu d'avoir des clusters définis.Nous paramétrons la fonction lien f_ϕ comme une correspondance linéaire entre les covariables et les paramètres de population :

$$f_{\boldsymbol{\phi}}(c_i) = \boldsymbol{\phi}_{slope} \cdot c_i + \boldsymbol{\phi}_{intercept}$$

C'est-à-dire pour chacun des paramètres $(\tilde{g_k}, \tilde{v_k}, , \overline{\tau})$ on calcule une paire de paramètres $\boldsymbol{\phi}$ correspondant à la pente et à l'intercept de la formulation linéaire. Ces paramètres $\boldsymbol{\phi}$ nous donnerons les valeurs de $(\tilde{g_k}, \tilde{v_k}, , \overline{\tau})$ pour chaque valeur possible des covariables c_i. Ainsi, les paramètres du modèle statistique sont $\theta = (\bar{\boldsymbol{\phi}}, \bar{\beta}_{ml}, \sigma_\xi, \sigma_\tau, \sigma)$ tandis que $\sigma_{\boldsymbol{\phi}}$, σ_β sont fixes.

2.2.2 Approche non-supervisée : Modèle de mélange

Pour créer un modèle de mélange, nous ajoutons une nouvelle couche au-dessus de la structure hiérarchique déjà construite et nous nous inspirons de la méthode décrite dans Poulet et Durrleman (2021). Notre contribution réside dans le fait que nous intégrons les paramètres individuels spatials (les sources s_{il}) dans le mélange probabiliste. Nous supposons que chaque cluster c est défini par ses paramètres $(\overline{\xi}^c, \sigma_\xi^c, \overline{\tau}^c, \sigma_\tau^c, \overline{s}^c)$ et décrit ainsi un sous-type avec une moyenne et un écart-type différents pour le début de la maladie et le rythme de progression, mais aussi des moyennes différentes pour les paramètres contrôlant l'ordre d'apparition des caractéristiques mesurées. Ainsi, les clusters formés se différencient à la fois en termes de progression spatiale et temporelle. Dans ce cadre les paramètres individuels suivent des distributions de mélange de gaussienne et donc l'équation 2 est remplacé par :

$$\begin{cases} \xi_i \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{\xi}^c, \sigma_\xi^c), & with \quad \sum_{c=1}^{n_c} \pi^c \overline{\xi}^c = 0 \\ \tau_i \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{\tau}^c, \sigma_\tau^c) & \\ s_{il} \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{s}^c, 1), & with \quad \sum_{c=1}^{n_c} \pi^c \overline{s}^c = 0 \end{cases} \tag{4}$$

avec π^c les coefficients de mélange qui correspondent aux probabilités d'occurrence de chaque cluster c. Par conséquent, nous considérons que chaque patient appartient au cluster c, avec une probabilité individuelle π_i^c dépendant de la vraisemblance des effets aléatoires. L'estimation conjointe des clusters et des paramètres du modèle est effectuée à l'aide d'un algorithme de mélange MCMC avec approximation stochastique (M-MCMC SAEM).

3 Cadre d'application

L'artériopathie cérébrale autosomique dominante à l'origine d'infarctus sous-corticaux et d'une leucoencéphalopathie (CADASIL) est la maladie héréditaire la plus courante touchant les petits vaisseaux cérébraux (SVD) et elle est causée par des mutations du gène NOTCH3 (Rutten et al. (2020)). Les patients atteints de CADASIL présentent un large éventail de symptômes tels que des crises de migraines avec aura, des accidents vasculaires cérébraux, des troubles du comportement, une incapacité motrice et des troubles cognitifs de la dysfonction exécutive jusqu'à une démence sévère (Chabriat et al. (2009)). L'étude de sous-groupes de patients CADASIL (Kaisaridi et al.(2025)) a montré que les patients présentant un début précoce et un rythme de progression rapide développent plus tôt des symptômes d'incapacité, de dépendance et de déficits focaux (éventuellement liés aux accidents vasculaires cérébraux). De l'autre côté les patients présentant un début plus tardif et un rythme de progression plus doux, développent d'abord des symptômes cognitifs. Globalement, le sexe, le niveau d'éducation, la position de la mutation dans les domaines de l'EGFr et certains facteurs de risque cardiovasculaire peuvent affecter la progression clinique de la maladie.

Avec le modéle à covariables nous allons évaluer l'intensité de scores cliniques en fonction de la position de la mutation de maniére supervisée. Nous supposons à priori que la position de la mutation est un facteur de risque qui affecte la progression de la maladie. De l'autre côté avec le modéle de mélange que nous avons implémenté dans Leaspy nous allons d'abord identifier des sous-groupes d'évolution et ensuite explorer l'effet de la position de la

mutation. Nous adoptons une approche non-supervisé où les différents profils de progression sont basés uniquement sur l' évolution des caractéristiques de la maladie. Une fois que les sous-groupes sont formés nous allons examiner l'effet de la position de la mutation sur leur formation. Nous comparerons ensuite les sous-groupes issus de l'approche supervisée et ceux obtenus par l'approche non-supervisée afin d'évaluer leur pertinence. Cette comparaison nous permettra de mieux comprendre dans quelle mesure la position de la mutation influence la progression de la maladie et si les sous-groupes identifiés de manière non-supervisée capturent des dynamiques similaires ou distincte.

Bibliographie

Chabriat, H., Joutel, A., Dichgans, M., Tournier-Lasserve, E. et Bousser, M.G. (2009), Cadasil. *The Lancet. Neurology*, 8(7), pp. 643-653.

Fournier N., Durrleman S. (2023). A Multimodal Disease Progression Model for Genetic Associations with Disease Dynamics. *Medical Image Computing and Computer Assisted Intervention, MICCAI, Oct 2023, Vancouver, Canada*, pp.601-610

Kaisaridi S., Herve D., Jabouley A., Reyes S., Machado C., Guey S., Taleb A., Fernandes F., Chabriat H. et Tezenas du Montcel S. (2025). Determining Clinical Disease Progression in Symptomatic Patients With CADASIL. *Neurology*, 104(1), e210193

Koval I., Dighiero-Brecht T., Tobin A. J., Tabrizi S. J., Scahill R. I., Tezenas du Montcel S., Durrleman S., et Durr A. (2022), Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials. *Scientific reports - Nature*, 12(1), 18928.

Maheux E., Koval I., Ortholand J., Birkenbihl C., Archetti D., Bouteloup V., Epelbaum S., Dufouil C., Hofmann-Apitius M., Durrleman S. (2023). Forecasting individual progression trajectories in Alzheimer's disease. *Nature Commmunications*, 14(1), pp. 761

Moulaire P., Poulet P.E., Petit E., Klockgether T., Durr A., Ashisawa T., Tezenas du Montcel S., for the READISCA Consortium (2023). *Movement Disorders*, 38(1), pp. 35-44

Ortholand J., Pradat P.F., Tezenas du Montcel S., et Durrleman S. (2023). Interaction of sex and onset site on the disease trajectory of amyotrophic lateral sclerosis. *Journal of Neurology*, 270(12), pp. 5903-5912.

Poulet P. E., Durrleman S. (2023). Multivariate disease progression modeling with longitudinal ordinal data. *Statistics in Medicine*, 42(18), pp. 3164-3183

Poulet P. E., Durrleman S. (2021). Mixture modeling for identifying subtypes in disease course mapping. *Information Processing in Medical Imaging, 27th International Conference*

Rutten, J. W., Hack, R. J., Duering, M., Gravesteijn, G., Dauwerse, J. G., Overzier, M., van den Akker, E. B., Slagboom, E., Holstege, H., Nho, K., Saykin, A., Dichgans, M., Malik, R., et Lesnik Oberstein, S. A. J. (2020), Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank : CADASIL to nonpenetrance. *Neurology*, 95(13), pp. e1835-e1843.

Schirratti J.B., Allassonniére S., Colliot O., et Durrleman S. (2017), A Bayesian Mixed- Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. *Journal of Machine Learning Research*, 18, pp. 1-33.

Young, A. L., Oxtoby, N. P., Garbarino, S., Fox, N. C., Barkhof, F., Schott, J. M., et Alexander, D. C. (2024), Data-driven modelling of neurodegenerative disease progression : thinking outside the black box. *Nature reviews.* Neuroscience, 25(2), pp. 111-130.