Construction de modèles mécanistes en grande dimension avec une approche par LASSO : application à la vaccination contre le virus Ebola

Auriane Gabaut [1+] & Mélanie Prague [1*]

[1] *Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center, SISTM Team ; Vaccine Research Institute, Créteil, France ;*
∗ *melanie.prague@inria.fr ;* + *auriane.gabaut@inria.fr*

Résumé. La construction de modèles non-linéaires à effets mixtes (NLMEM) enrichit notre compréhension des processus biologiques. L'estimation dans ces modèles est facilitée par des méthodes de maximum de vraisemblance, notamment l'algorithme Stochastic Approximation Expectation-Maximization (Kuhn & Lavielle, 2005). Toutefois, cette méthode requiert une importante charge de calcul, ce qui a mené à proposer des techniques automatisées de modélisation, particulièrement pour la sélection de covariables définissant les paramètres au niveau individuel (Svensson et Jonsson, 2022 ; Aural et al.,2021). À l'instar de ces autres méthodes, en optimisant un critère d'information de type BIC, l'algorithme Stochastic Approximation for Model Building (SAMBA - Prague & Lavielle, 2022) élabore le modèle de covariables en se fondant sur la simulation de réalisations selon la loi a posteriori des paramètres individuels. Initialement, SAMBA recourt à un algorithme stepAIC pour la sélection du lien de ces covariables sur ces réalisations. Dans ce travail, nous proposons l'utilisation de la méthode LASSO pour une meilleure gestion des covariables de grande dimension. Cette méthode inclut un processus de sélection par stabilité (Meinshausen & Bühlmann, 2010). Nous avons validé notre approche à travers des simulations reproduisant la dynamique de la réponse immunitaire humorale à un vaccin contre Ebola (Pasin, 2019) en lien avec des données de transcriptomes mesurées au temps d'inclusion. Notamment, notre méthode a permis de réduire le taux de faux positifs, tout en conservant un taux de faux négatifs comparable. Nous avons mis en œuvre notre méthode avec les données de l'essai Prevac/Prevac-UP (Prevac-UP Team, 2022), comparant deux vaccins contre Ebola autorisés en Afrique.

Mots-clés. Apprentissage statistique ; Biostatistique ; génomique ; santé ; Grande dimension ; Sélection de modèles : Modèles non linéaires à effets mixtes.

Abstract. The development of nonlinear mixed-effects models (NLMEM) deepens our understanding of biological processes. Estimation in these models is facilitated by maximum likelihood methods, notably the Stochastic Approximation Expectation-Maximization algorithm (Kuhn & Lavielle, 2005). However, this method requires significant computational resources, leading to the proposal of automated modeling building techniques, especially for selecting covariates that define parameters at the individual level (SCM-Jonsson, 1998 ; COSSAC-Ayral, 2021). Similar to other models, by optimizing a Bayesian Information Criterion (BIC) type of information criterion, the Stochastic Approximation for Model Building (SAMBA-Prague & Lavielle, 2022) algorithm develops the covariate model based on realizations of the posterior distribution of individual parameters. Initially, SAMBA uses a stepAIC algorithm for the selection of the link of these covariates on these realizations. In this work, we propose the use of the LASSO methods for better management of high-dimensional covariates. This method includes a stability selection process (Meinshausen & Bühlmann, 2010). We validated our approach through simulations that replicate the dynamics of the humoral immune response to an Ebola vaccine (Pasin, 2019) linked with transcriptome data measured at the time of inclusion. Notably, our method reduced the rate of false discoveries while maintaining a comparable rate of false negatives. We implemented our method with data from the Prevac/Prevac-UP trial (Prevac-UP Team, 2022), comparing two authorized Ebola vaccines in Africa.

Keywords. Statistical learning ; Biostatistics ; Genomics ; Health ; High dimensionality ; Model selection : Nonlinear mixed-effects models.

1. Introduction

En modélisation statistique, la sélection de modèle est une étape essentielle pour identifier le modèle le plus approprié afin de représenter au mieux les données et de faire des prédictions précises. Pour faciliter ce processus, plusieurs algorithmes ont été développés. La sélection de modèle et l'utilisation d'algorithmes appropriés, tels que SCM (Stepwise Covariate Modeling) (Svensson and Jonsson, 2022), COSSAC (COnditional Sampling use for Stepwise Approach based on Correlation tests) (Ayral et al., 2021) et SAMBA (Stochastic Approximation for Model Building Algorithm) (Prague and Lavielle, 2022), sont d'une grande pertinence pour la médecine personnalisée.

1.1. Modèle Non-Linéaires à Effets Mixtes - Notations

Nous considérons des observations $Y_i = (Y_{ij})_{j \leq n_i}$ observée aux instants t_{ij} pour l'individu $i \leq N$. Nous supposons que ces observations proviennent d'une dynamique individuelle $y_i = f(\cdot, \psi_i)$ dépendant du temps et des paramètres du modèle $\psi_i = (\psi_{il})_{l \leq m}$.

Dans le contexte des modèles non-linéaires à effets mixtes, on suppose que les paramètres individuels sont normalement distribués, $\psi_i \overset{iid}{\sim} \mathcal{N}(\psi_{pop}, \Omega)$ autour d'un paramètre de population $\psi_{pop} \in \mathbb{R}^m$. De plus, si l'individu i est caractérisé par n covariables $X_i = (X_{i1}, ..., X_{in})$, on peut inclure des effets de covariables dans la définition des paramètres individuels, de sorte que

$$\forall i \leq N, \ h(\psi_i) = h(\psi_{pop}) + X_i\beta + \eta_i \tag{1}$$

avec $\beta = (\beta_1, ..., \beta_m)$, où $\beta_l = (\beta_{l1}, ..., \beta_{ln})^T$ est l'effet des n covariables sur le paramètre $l \leq m$, à une transformation h près.

On peut finalement écrire le modèle non-linéaires à effets mixtes général comme

▸▸ **Modèle Structural** :

$$\forall i \leq N, j \leq n_i \quad \left\{ \begin{array}{rcl} y_i & = & f(\cdot, \psi_i) \\ h(\psi_i) & = & h(\psi_{pop}) + X_i\beta + \eta_i \end{array} \right. \qquad \text{(Mod Str)}$$

▸▸ **Modèle Statistique** :

$$\forall i \leq N, \quad \eta_i \overset{iid}{\sim} \mathcal{N}(0, \Omega) \qquad \text{(Mod Stat)}$$

▸▸ **Modèle d'Observation** :

$$\forall i \leq N, \quad \left\{ \begin{array}{rcl} u(Y_{ij}) & = & u(y_i(t_{ij})) + g(t_{ij}, \psi_i, \xi)\varepsilon_{ij} \\ (\varepsilon_{ij})_{j \leq n_i} & \overset{iid}{\sim} & \mathcal{N}(0, I_{n_i}) \end{array} \right. \qquad \text{(Mod Obs)}$$

(Mod)

Il est important de noter qu'il est possible de définir seulement un sous-ensemble des paramètres comme individuels, pas nécessairement tous les paramètres du modèle.

1.2. SAMBA

L'algorithme SAMBA (Prague and Lavielle, 2022) est itératif, il commence avec un modèle initial, qui est généralement vide $\mathscr{M}_0$ - ce qui signifie qu'il ne contient aucun effet de covariable et aucune corrélation. Ensuite, à chaque itération k, l'algorithme construit un nouveau modèle $\mathscr{M}_{k+1}$ basé sur le précédent $\mathscr{M}_k$. Pour ce faire, il estime les paramètres $\theta^{(k)} = \left(\psi_{pop}, \Omega, \beta, \xi\right)$ du modèle $\mathscr{M}_k$ par maximum de vraisemblance, puis échantillonne $\psi_i^{(k)}$ pour chaque individu selon la distribution à posteriori $p\left(\cdot \middle| Y, \theta^{(k)}\right)$ obtenue par Monte Carlo. À partir de ces paramètres, l'algorithme construit un modèle de covariables (ainsi qu'un modèle de correlation et d'erreur qui ne sont pas détaillés ici). Le modèle de covariable, noté $\mathscr{M}_{\mathbf{k+1}}^{\mathbf{COV}}$, se base sur les paramètres générés $\psi_i^{(k)} = (\psi_{i1}^{(k)}, ..., \psi_{im}^{(k)})$, $i \leq N$, et l'estimation $\psi_{pop}^{(k)} = (\psi_{pop\ l}^{(k)})_{l \leq m}$.

Des tests statistiques excluent les covariables qui n'affectent pas significativement un paramètre. Ensuite, un modèle de régression est construit pour le paramètre ψ_l, $l \leq m$:

$$h_l(\psi_{li}^{(k)}) = h_l(\psi_{lpop}^{(k)}) + c_i\beta_l + \eta_{il}^{(k)} \tag{2}$$

où $\eta_{il}^{(k)} \sim \mathcal{N}\left(0, {\omega_l^{(k)}}^2\right)$, et c_i est les covariables individuelles sélectionnées par une procédure de sélection séquentiel de covariable, stepAIC, s'il y a plus de 10 covariables, ou par une recherche exhaustive parmi tous les modèles s'il y en a moins.

2. Méthode

La méthode présentée propose de remplacer l'algorithme stepAIC pour la sélection de covariables dans SAMBA par une sélection par Lasso. Nous ajoutons un algorithme de sélection par stabilité pour réduire le problème d'instabilité de l'estimateur Lasso et éviter la sélection de covariables non significatives. L'algorithme initialement utilisé par SAMBA permet au critère d'information de diminuer au fil des itérations. Pour reproduire ce comportement, nous incorporons une recherche parmi plusieurs modèles de covariables, visant à minimiser le critère d'information.

2.1. Sélection par Lasso

Nous considérons un modèle non-linéaires à effets mixtes, comme introduit dans Mod précédemment, avec $N \in \mathbb{N}^*$ individus, $n \in \mathbb{N}^*$ covariables mesurées, et un total de $m \in \mathbb{N}^*$ paramètres dans le modèle.

À chaque itération $k \in \mathbb{N}$, la procédure SAMBA effectue d'abord, pour le modèle construit précédemment $\mathcal{M}_k$, l'estimation des paramètres de population $\theta^{(k)}$ et l'échantillonnage des paramètres individuels $(\psi_i^{(k)})_{i \leq N}$. Nous sommes alors en mesure d'écrire le modèle de régression :

$$h_l(\psi_{il}^{(k)}) = h_l(\psi_{popl}^{(k)}) + X_i\beta_l + \eta_{il}^{(k)} \tag{2}$$

Dans la nouvelle méthode, nous proposons d'effectuer la sélection par Lasso sur le modèle de régression pour chaque paramètre $l \leq m$:

$$Y_l^{(k)} = \left(h_l\left(\psi_{il}^{(k)}\right)\right)_{i \leq N} = \left(h_l\left(\psi_{popl}^{(k)}\right)\right)_{l \leq m} + X\beta_l + \eta_l^{(k)} \tag{3}$$

avec $X = (X_1, ..., X_i)^T \in \mathcal{M}_{Nn}(\mathbb{R})$; $Y^{(k)}$, $Y_{pop}^{(k)} \in \mathbb{R}^m$; $\eta_l^{(k)} \in \mathbb{R}^m$; $\beta_l \in \mathbb{R}^n$.

2.1.1. Algorithme de sélection par stabilité

L'algorithme de sélection par stabilité (Meinshausen and Bühlmann, 2010) est une méthode développée pour rendre la sélection par Lasso plus stable et robuste. L'idée centrale de l'algorithme de sélection par stabilité est de répéter le processus de sélection par Lasso un grand nombre de fois, sur des batchs de données de taille $\lfloor N/2 \rfloor$ tirés au hasard l'ensemble des données d'observation, de taille $N \in \mathbb{N}^*$. Pour chaque batchs de données, les covariables sont sélectionnées par Lasso. Après avoir effectué ces nombreuses sélections, le nombre de fois où chaque covariable est sélectionnée dans le modèle final parmi tous les batchs est comptabilisé. Une variable est finalement conservée si sa fréquence de sélection dépasse un seuil prédéfini t_{SS}.

Nous proposons dans la nouvelle méthode d'utiliser soit la sélection par stabilité telle qu'elle est proposée par Meinshausen et Bühlmann, sur des batchs de données des paramètres individuels obtenus lors de l'étape de simulation ; soit de substituer les batchs de données par les réplicats obtenus lors de l'étape d'échantillonnage dans SAMBA. Pour gérer le paramètre de seuil t_{SS}, il est soit fixé par l'utilisateur au début de l'algorithme SAMBA, ou en le cherchant sur une grille de valeurs de 60% à 95% de sorte de minimiser le critère d'information. Nous proposons également l'utilisation de l'algorithme de calibration automatique sharp (Bodinier et al., 2023).

3. Illustration

Nous proposons une application des méthodes sur les données recueillies lors de l'étude PREVAC-UP du consortium PREVAC (Badio et al., 2021). Cette étude vise à évaluer la sécurité et la durabilité de la réponse immunitaire de trois stratégies de vaccins contre Ebola, dont notamment la stratégie vaccinale Ad26.ZEBOV/MVA-BN-Filo sur laquelle nous nous focaliserons.

Lors de cette étude, 1400 adultes et 1401 enfants de Guinée, Libéria, Mali et Sierroa Leone, ont reçue un vaccin contre Ebola. Ils ont ensuite été surveillés : pour chaque individu, des mesures d'anticorps, comme illustré dans la figure 1, ont été prises le jour de la première dose, puis à environ 7, 14, 28, 56 jours après. La deuxième dose est administrée 56 jours après la première. Les patients ont ensuite été suivis aux jours 63, et à 3, 6 et 12 mois après l'inclusion.

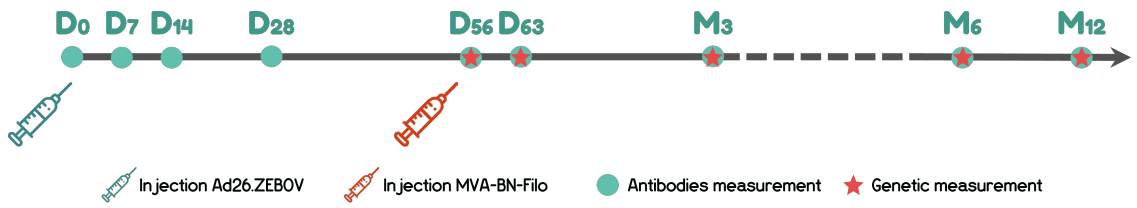

FIGURE 1 – Schéma des essais PREVAC-UP pour les personnes ayant reçu le schéma de vaccination Ad26.ZEBOV/MVA-BN-Filo.

Une sous-étude immunologique, menée spécifiquement en Guinée, rassemble 95 participants adultes, dont 30 ont reçu le schéma Ad26/MVA. Pour chacun d'eux, en plus des mesures d'anticorps, l'expression génique de plus de 28 000 gènes a été réalisée après le rappel du vaccin. Notre objectif ici est de sélectionner parmi les gènes ceux prédisant la réponse immunitaire.

La réponse immunitaire humorale a déjà été modélisée par un modèle non linéaire à effets mixtes (Pasin et al., 2019). Nous modélisons la production d'anticorps en considérant deux cellules sécrétrices d'anticorps (ASC), notées S -*pour courte durée*- et L -*pour longue durée*-, et caractérisées par leur demi-vie. Nous supposons que ces ASC produisent respectivement des anticorps Ab aux taux θ_S et θ_L. Les taux de décroissance de toutes ces entités biologiques sont respectivement notés δ_S, δ_L et δ_{Ab}. Ainsi, nous simulons la production d'anticorps définie par l'EDO suivante :

$$\frac{d}{dt}Ab(t) = \varphi_S e^{-\delta_S t} + \varphi_L e^{-\delta_L t} - \delta_{Ab} Ab(t)$$

Nous ajoutons ensuite des effets aléatoires sur δ_{Ab}, φ_S et φ_L et testons les différents expression de gènes comme effets de covariables. Concernant le modèle d'observation, seul une mesure bruitée des niveaux d'anticorps est mesurée.

Pour illustrer les avantages de la nouvelle méthode développée par rapport à la méthode originale, nous présenterons une étude de simulation. Notre objectif est alors de tester si les méthodes proposés permettent d'obtenir un meilleur taux de faux positifs en conservant des niveaux comparables de faux négatifs parmis les covariables sélectionnés dans le modèle final.

Nous présenterons aussi les résultats de l'étude sur les données de PREVAC.

Références

Ayral, G., Si Abdallah, J.-F., Magnard, C., and Chauvin, J. (2021). A novel method based on unbiased correlations tests for covariate selection in nonlinear mixed

effects models : The cossac approach. *CPT : Pharmacometrics & Systems Pharmacology*, 10(4) :318–329.

Badio, M., Lhomme, E., Kieh, M., Beavogui, A., Kennedy, S., Doumbia, S., Leigh, B., Sow, S., Diallo, A., Fusco, D., Kirchoff, M., Termote, M., Vatrinet, R., Wentworth, D., Esperou, H., Lane, H., Pierson, J., Watson-Jones, D., Roy, C., and Yazdanpanah, Y. (2021). Partnership for research on ebola vaccination (prevac) : protocol of a randomized, double-blind, placebo-controlled phase 2 clinical trial evaluating three vaccine strategies against ebola in healthy volunteers in four west african countries. *Trials*, 22.

Bodinier, B., Filippi, S., Nøst, T. H., Chiquet, J., and Chadeau-Hyam, M. (2023). Automated calibration for stability selection in penalised regression and graphical models. *Journal of the Royal Statistical Society Series C : Applied Statistics*, 72 :1375–1393.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72 :417–473.

Pasin, C., Balelli, I., Van Effelterre, T., Bockstal, V., Solforosi, L., Prague, M., Douoguih, M., and Thiébaut, R. (2019). Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine : Quantification and Sources of Variation. *Journal of Virology*, 93(18) :e00579–19.

Prague, M. and Lavielle, M. (2022). Samba : A novel method for fast automatic model building in nonlinear mixed-effects models. *CPT : Pharmacometrics & Systems Pharmacology*, 11(2) :161–172.

Svensson, R. J. and Jonsson, E. N. (2022). Efficient and relevant stepwise covariate model building for pharmacometrics. *CPT : Pharmacometrics & Systems Pharmacology*, 11(9) :1210–1222.