

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE
DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX
ECOLE DOCTORALE SOCIÉTÉS, POLITIQUE, SANTÉ
PUBLIQUE

SPÉCIALITÉ SANTÉ PUBLIQUE, OPTION BIOSTATISTIQUE

Par **Auriane Gabaut**

Méthodes de régularisation pour l'intégration de données de
grande dimension dans les modèles mécanistes : application pour
le développement de vaccins

Sous la direction de : **Mélanie Prague**

Co-directrice : **Cécile Proust-Lima**

Soutenue le 25 novembre 2025

Membres du jury :

M. CHIQUET Julien	Directeur de Recherche, Univ. Paris-Saclay, AgroParisTech, INRAE	Rapporteur
Mme. COMETS Emmanuelle	Chargée de Recherche, INSERM Paris	Rapporteuse
Mme. JACQMIN-GADDA Hélène	Directrice de Recherche, INSERM Bordeaux	Présidente
M. BENZEKRY Sébastien	Chargé de Recherche, INRIA-INSERM Aix-Marseille	Examineur
M. NEUVIAL Pierre	Directeur de Recherche, CNRS Toulouse	Examineur
Mme. PRAGUE Mélanie	Chargée de Recherche, INRIA Bordeaux	Directrice de thèse
Mme. PROUST-LIMA Cécile	Directrice de Recherche, INSERM Bordeaux	Co-directrice de thèse

Table des matières

Remerciements	5
Liste des abréviations	9
Production Scientifique et Enseignement	11
Introduction et Objectifs	15
1 Contexte Biologique	19
1.1 Le système immunitaire	19
1.1.1 L'immunité innée	20
1.1.2 L'immunité adaptative	21
1.1.3 L'immunité adaptative humorale	24
1.1.4 L'immunité adaptative à médiation cellulaire	26
1.2 Essais cliniques vaccinaux	27
1.2.1 Cadre expérimental des essais vaccinaux	28
1.2.2 Données immunologiques d'intérêt dans cette thèse	29
1.2.2.1 Dosages d'anticorps (Luminex, ELISA)	29
1.2.2.2 Profil d'expression génique (RNA-seq)	30
1.2.3 Jeux de données utilisés	32
1.2.3.1 Essai PREVAC-UP : vaccination contre Ebola	32
1.2.3.2 Essai sur le vaccin ZOSTAVAX contre le zona	33
1.2.3.3 Essai sur le vaccin ARNm contre le SARS-CoV-2	34
2 Modèles non-linéaires à effets mixtes et Méthodes de régularisation	37
2.1 Modèle statistique et estimation	38
2.1.1 Modélisation mécaniste et modèles non-linéaires à effets mixtes	38
2.1.1.1 Modèles linéaires à effets mixtes	38
2.1.1.2 Modèles non-linéaires à effets mixtes	39
2.1.1.3 Modèles mécanistes	40
2.1.2 Méthodes d'estimation dans les modèles non-linéaires à effets mixtes	42
2.1.2.1 Identifiabilité	43
2.1.2.2 Vraisemblance du modèle	43
2.1.2.3 Approximations de la vraisemblance	44
2.1.2.4 Optimisation de la vraisemblance	47

2.1.2.5	Algorithmes EM et SAEM	48
2.1.2.6	Monolix et l'implémentation de l'algorithme SAEM	50
2.2	Sélection de modèles et régularisation dans les modèles NLEM	51
2.2.1	Critères d'information pour la sélection de modèles	51
2.2.1.1	Critères AIC et BIC	52
2.2.1.2	Critères corrigés : le BICc	53
2.2.2	Pénalisation	53
2.2.2.1	Pénalisation Lasso	55
2.2.2.2	Stratégies de sélection et de calibration	57
2.2.2.3	Algorithme de sélection stable	58
2.2.3	Pénalisation Lasso dans les modèles non-linéaires à effets mixtes	60
2.3	Utilisation des modèles mécanistes pour l'immunologie	62
2.3.1	Modélisation du système immunitaire par Bonin <i>et al.</i>	63
2.3.2	Modélisation de la réponse immunitaire humorale par Dari <i>et al.</i>	66
2.3.3	Modélisation de la réponse humorale post-vaccinale par Balelli <i>et al.</i> et Clairon <i>et al.</i>	68
2.3.4	Modèle à deux compartiments de plasmocytes par Andraud <i>et al.</i>	71
2.3.5	Discussion sur l'identifiabilité et la complexité des modèles	72
3	Sélection de covariables par une approche Lasso et de sélection stable dans un algorithme de construction de modèles non-linéaires à effets mixtes	75
3.1	Positionnement et contribution	75
3.2	Article : A Stability-Enhanced Lasso Approach for Covariate Selection in Non- Linear Mixed Effects Model	80
3.3	Travaux additionnels : Analyse des données de l'essai clinique PREVAC-UP	96
3.4	Perspectives	100
4	Estimation régularisée dans les modèles non-linéaires à effets mixtes en présence de biomarqueurs répétés de grande dimension	105
4.1	Positionnement et contribution	106
4.2	Intégration des données transcriptomiques dans un modèle mécaniste de la réponse humorale	108
4.2.1	Modèle mécaniste et modèles d'observations	108
4.2.2	Méthode	109
4.2.3	Résultats	111
4.2.4	Discussion	114
4.3	Article : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers	115
4.4	Vignette : REMixed : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers in R	138
4.5	Perspectives	163

5	Discussion générale	167
	Matériels Supplémentaires	171
A	Matériels Supplémentaires associés au Chapitre 3	172
A.1	Simulation generation process	172
A.1.1	Pharmacokinetic model	172
A.1.2	Humoral Immune Response	173
A.1.3	Covariates Generation	175
A.2	Additional analysis for the real-data application	178
A.2.1	Initial model	178
A.2.2	Final Model	184
A.3	Additional graphs for results	189
A.4	Figures additionnelles concernant l'analyse sur les données PREVAC-UP .	191
B	Matériels Supplémentaires associés au Chapitre 4	193
B.1	Matériels Supplémentaires associés à la Section 4.3	193
B.1.1	NLMEM log-likelihood	194
B.1.2	Updating the biomarker contributions	195
B.1.3	Simulation Setup	196
B.1.4	Evaluation metrics	199
B.1.5	Additional Tables and Figure	200
B.1.6	Application to COVID-19 mRNA vaccination data	211
B.2	Matériels Supplémentaires associés à la Section 4.4	217
B.2.1	Regularization step details	217
B.2.2	Structure of the project	220
B.2.3	Mlxtran model file for demo monolix project	222
B.2.4	Summary file for initialization task	222
B.2.5	Summary file for grid search	224
B.2.6	Summary file for REMix algorithm	231
B.2.7	Summary file for final tests and estimation	237
	Bibliographie	240

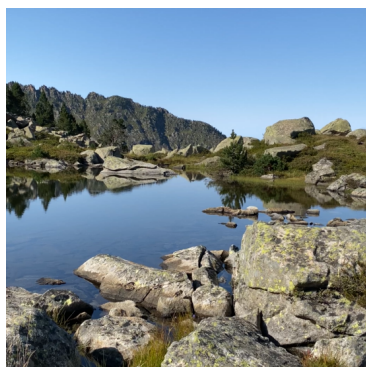
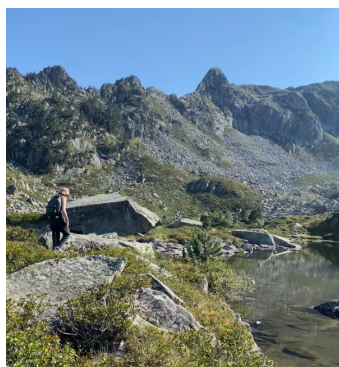
Remerciements



Riri (et Thilde) sur les sentiers du lac de Bastampe.

Voilà enfin, en ce jour de novembre, que s’achève cette thèse. Et c’est avec un certain plaisir que vos yeux attentifs cherchent, peut-être, votre nom au fil de ces quelques pages. À travers ces *quelques* lignes, j’espère avoir su rendre hommage à l’affection, au soutien, à la bienveillance avec lesquels vous m’avez portée ces dernières années. Cette aventure fut un parcours sinueux, parsemé de rebondissements, de rencontres, de larmes, de joies, de découvertes et d’enseignements.

Ces deux dernières années ont ressemblé, pour moi, à une randonnée improvisée vers le lac de Bastampe : un sentier raide, parfois presque impraticable sans de quoi escalader, et un timing... disons peu raisonnable, avec une fin précipitée pour ne pas se retrouver plongées dans le noir de la nuit. On courait sur des (*pseudo-*)sentiers, on tombait, on hésitait... mais on avançait, pressées par le temps qui filait. Et pourtant, une fois au sommet, la vue sur le lac était à couper le souffle ! Soudain, plus rien de tout ce chaos n’avait d’importance. Dans cette course un peu folle qu’a été ma thèse, vous avez été la corde (ou la frontale...) qui me permettait d’avancer. Celle qui me retenait quand je tombais, qui me guidait et m’éclairait quand je ne voyais plus clair, et surtout, qui me rassurait et me permettait d’avancer. Et si, par malheur, votre nom ne figure pas *vert* sur blanc ici, s’il s’est égaré entre deux paragraphes... il est tout de même bien là, ancré dans cette aventure et dans ma mémoire, à la place qu’il mérite.



Magnifique lac de Bastampe.

Je tiens d’abord à exprimer toute ma gratitude à mes deux directrices de thèse, **Mélanie** et **Cécile**. Leur écoute, leur patience et leur disponibilité m’ont accompagnée tout au long de ces années et elles ont été un soutien moral constant, même lorsque je sais que je ne leur ai pas toujours facilité la tâche. **Mélanie**, merci de m’avoir accordé ta confiance dès la fin de mon master, et de m’avoir accueillie avec tant de chaleur à Bordeaux. Ton aide précieuse m’a permis de m’installer sereinement et de trouver ma place dans le monde de la recherche. **Cécile**, merci pour tes encouragements constants, glissés en fin de mails ou au détour d’un commentaire dans un écrit. Ces mots, parfois discrets, m’ont donné une énergie incroyable pour continuer. Je suis heureuse d’avoir eu la chance de travailler à vos côtés, et profondément reconnaissante pour tout

le temps que vous m'avez consacré, que ce soit pour me préparer aux conférences, relire mes écrits, ou simplement m'écouter avec attention.

J'adresse mes remerciements à l'ensemble des membres de mon jury de thèse, **Sébastien Benzekry**, **Julien Chiquet**, **Emmanuelle Comets**, **Hélène Jacqmin-Gadda** et **Pierre Neuviel**, pour avoir accepté de consacrer de leur temps à l'évaluation de ce travail. J'exprime une reconnaissance particulière à **Julien Chiquet** et **Emmanuelle Comets**, pour avoir accepté la tâche exigeante de rapporteur et rapportrice, ainsi qu'à **Hélène Jacqmin-Gadda**, pour avoir accepté d'assurer la présidence de ce jury.

Je remercie vivement l'ensemble de l'équipe **SISTM** pour l'ambiance stimulante et conviviale dans laquelle j'ai eu la chance d'évoluer. Les échanges scientifiques, les moments de partage et l'entraide au quotidien ont fait de ces années une période particulièrement enrichissante. Je tiens également à remercier l'équipe **BIOSTAT**, dont l'accueil et les échanges ont été tout aussi précieux. Au sein de ces deux équipes, je souhaite remercier chaleureusement les doctorant.e.s, titulaires, ingénieur.e.s et stagiaires, qui ont chacun et chacune contribué à rendre cette expérience unique.

Je souhaite adresser tout particulièrement un immense merci à **Sandrine**, dont la gentillesse et le sourire ont illuminé mon quotidien. Son efficacité, sa disponibilité et son aide précieuse ont été d'une grande valeur, et je lui suis profondément reconnaissante pour tout ce qu'elle a apporté avec tant de bonne humeur. Un grand merci également à **Julia**, pour son aide précieuse dans la logistique de cette fin de thèse. Enfin, j'ai eu la chance de travailler, dans le cadre de ma thèse, avec **Ariane**, **Boris**, **Lisa** et **Rodolphe**, que je remercie très sincèrement. Merci tout particulièrement à **Boris** et **Rodolphe**, pour leur disponibilité et leurs conseils dans la préparation de ma soutenance, et à **Boris**, plus spécifiquement, pour son accompagnement tout au long de mon parcours et dans le cadre de mon CSI.

Je souhaite vivement remercier le bureau **John Snow**, ainsi que tous les bureaux voisins (y compris ceux de différents étages), qui sont devenus au fil des années bien plus qu'un simple espace de travail : un lieu de partage, de complicité, et surtout, de traditions qui me sont chères : nos cartes postales, nos goûters d'anniversaire, nos éblouissantes décorations... Tant de petites choses qui ont illuminé mon quotidien, et qui vont, sans aucun doute, terriblement me manquer. Un grand merci à **Hélène**, **Marie** et **Quentin** : vous avez été d'une aide inestimable dans mes moments de doute. Votre écoute et vos conseils m'ont souvent aidée à remettre les choses en perspective et à prendre des décisions plus éclairées. À **Mélanie H.**, notre Bretonne repartie trop tôt sur ses terres : ton amitié et tes conseils m'ont accompagnée tout au long de la thèse, et m'ont été infiniment précieux. Tu as apporté une vraie chaleur humaine à cette aventure. À **Corentin** : ta bonne humeur contagieuse a rythmé tant de repas. Ta présence enjouée a toujours su ramener le sourire, même dans les journées les plus grises. À **Benjamin**, mon voisin de bureau : pour nos discussions et, bien sûr, ta présence pleine de vie ! Tes chants et danses improvisés, les moments jeux dans le bureau, ont rendu le quotidien tellement plus jovial. À **Antonin** : tes récits d'aventures (et tes fabuleux dessins !) ont apporté une touche de légèreté au bureau. À **Ariane** : je suis heureuse d'avoir partagé avec toi quelques parenthèses

artistiques qui m’ont permis de changer d’aire. À **Marius** : pour ton humour et tes facéties redoutable, qui ont animé tant de journées et de voyages... mais également pour ton humour. Et ton humour, aussi, bien sûr ! Mais plus sérieusement, merci pour... À **Justine**, ma camarade d’applications “en cours” : au moins, nous n’étions pas seules dans notre galère et c’était déjà un soulagement en soi ! À **Sara**, ma seconde voisine de bureau : nos innombrables échanges m’ont souvent apaisée et remotivée, tu m’as tant de fois aidée à prendre du recul dans les périodes chargées. Je suis heureuse que notre aventure se poursuive à Toulouse, et tellement impatiente d’y partager encore rires et cafés. À **Adrien**, compagnon de galère de thèse (à distance) : nos discussions et les confs passées ensemble (Rome reste un moment fort !) ont été un vrai plaisir. À **Aurore** : toujours là quand j’avais besoin. Tu m’as sortie de tant de galères avec générosité et sans jamais hésiter. Et promis, nous prendrons le temps bientôt de finir cette partie d’échecs. À **Alice** : arrivée plus récemment, tu as su très vite trouver une place toute particulière pour moi. Merci pour ton écoute, ton humanité et ton empathie : nos discussions me font toujours un bien fou. À **Lisa** : je suis enchantée d’avoir travaillé ensemble ces derniers mois et d’avoir partagé un peu de ma galère avec toi. Je n’oublie pas non plus **Arthur**, **Quentin (Q2)**, **Nicolas**, **Daniela**, **Anne-Andrée**... et toutes les personnes que je n’ai pas citées ici, mais avec qui j’ai partagé des moments qui comptent, qu’ils aient été humains, scientifiques, ou un peu des deux.

Un immense merci aux amitiés nées au fil de cette thèse, en particulier à **Mélanie**, **Coco**, **Benji**, **Marius**, **Sara** et **Alice** ; mes très chers “amigollègues”. Vous avez rendu le quotidien bien plus léger et joyeux, et j’ai hâte de remonter à Bordeaux pour vous y retrouver, tout comme j’ai hâte de vous accueillir à Toulouse. J’espère que nous continuerons à partager ces moments qui nous rassemblent, nos discussions qui s’éternisent en terrasse, ces rires volés, nos petites pauses au labo qui rendaient les journées plus légères, offrant une parenthèse bienvenue... finalement, tout ce qui a donné à ma thèse sa vraie dimension humaine. J’ai aussi une pensée toute spéciale pour **Camille**, mon véritable coup de foudre bordelais. Vivre avec toi tout au long de cette thèse a été une chance immense. Tu as toujours su m’écouter avec une patience infinie, et ta présence a fait de notre colocation un lieu aussi vivant que rassurant. Merci également à **Lou**, arrivée plus récemment dans la colloc’, et qui a su, avec naturel et brio, compléter cette dynamique. Tu as apporté encore plus de chaleur et de rires dans un foyer qui en était déjà plein. “Chez moi” restera avec vous, tant dans les souvenirs que dans les retrouvailles à venir.

Je voudrais remercier mes amis de toujours, qui ont été un soutien précieux bien au-delà de cette thèse. À **Mathilde**, mon véritable alter ego, présente depuis toujours et qui a facilité chaque étape de mon parcours. Tu n’as jamais bronché devant mes appels à l’aide, et tu y as toujours répondu avec un soin incroyable. Je suis profondément reconnaissante de te compter près de moi. À **Clémentine**, dont le soutien inestimable et les conseils avisés m’ont tant apportée. Tu as été d’une aide décisive dans cette aventure, toujours là au bon moment, avec la parole juste. À **Anita**, un bain de fraîcheur et une amie précieuse : j’ai eu le bonheur de te retrouver régulièrement pour sortir la tête de tout ça... en faisant des maths ! Tu as même sauvé ma thèse en la compilant quand l’informatique m’a lâchée : *par quelle sorcellerie m’as-tu transmis ta malchance informatique ?* À **Sam**, sans qui je n’aurais jamais été aussi loin : merci pour ton affection et ton soutien indéfectible. Nos discussions scientifiques m’ont passionnée, et ton intérêt

sincère pour mon travail, quel qu'il soit, m'a toujours profondément touchée. À **Anaïs** : je me réjouis d'avance de partager encore trois années d'aventure. C'est une aventure un peu absurde, dans laquelle tu me comprends et m'accompagnes, et cela n'a pas de prix. À tout ce beau monde, merci pour nos moments partagés (avec **Pablo** et **Tim** aussi !), nos vacances, nos pauses, nos moments de télétravail. À **Chloë**, dont les petits messages réguliers ont eu un impact bien plus grand sur mon moral que tu ne l'imagines. Merci de m'accompagner de loin, et de me soutenir assidûment depuis le tout début. Merci également aux “copains” : **Théodore**, **Mitch**, **Rémi**, **Juliette**, **Guillaume n°5**... Vous êtes les premiers spectateurs de mes péripéties, et vous y avez toujours apporté humour et soutien. Merci à vous tous d'avoir toujours été là, malgré la distance ou les occupations, et d'avoir rendu ce parcours infiniment plus léger et joyeux.

Je veux adresser un merci immense, sincère, à ma famille. Vous avez été un soutien précieux, solide et inconditionnel tout au long de ce parcours. Merci **Papa** et **Maman**, vous avez toujours été là pour m'encourager, me rassurer et m'accompagner dans chacun de mes choix. Votre présence a été un repère constant. Merci à mes sœurs et à mon frère, dont les encouragements m'ont portée, même à distance. **Maurine**, merci d'avoir veillé sur moi à ta manière, toujours présente, rassurante et prête à tendre la main quand tout s'embrouillait. **Garance**, merci pour ta gaieté, ton écoute attentive et tes conseils, qui m'ont accompagnée à chaque étape. Merci également pour ton aide au cours de ces derniers mois, dans toute la logistique de cette fin de thèse. **Virgile**, merci pour ton insouciance, qui a su alléger les moments les plus lourds, et pour ta manière bien à toi de me rappeler que la persévérance est clé pour avancer. **Papi** et **Mamie**, merci pour votre affection, votre soutien à chaque étape, et votre patience à m'écouter déblatérer de tout, avec pour seule remarque : “parle moins vite, s'il te plaît!”. Merci aussi **Tonton Alain**, **Tata Danielle**, **Pépé Martial** et **Véronique**, pour vos encouragements constants, à chaque appel, échange ou retrouvaille. Enfin, merci **Laurine** et **Mathias**, pour votre présence bienveillante et votre aide logistique au cours de cette aventure.

Je tiens à remercier toutes les personnes qui ont accepté de relire les typos et fautes de morceaux de cette thèse : **Maman**, **Papa**, **Laurine**, **Maurine**, **Garance**, **Clémentine**, **Anita**, **Mélanie**, **Anaïs**, **Guillaume n°5**, **Sem** et **Mathilde**. Vos regards attentifs et vos suggestions m'ont été d'une grande aide. Un merci tout particulier à **Mélanie**, qui a eu le courage et la patience de relire l'intégralité de ce manuscrit, et à **Anita** pour sa relecture si minutieuse, allant jusqu'à corriger un indice dans un système de plus de dix équations.

Enfin, merci à **moi-même** d'être allée au bout de cette thèse : un parcours semé d'obstacles qui ont mis à rude épreuve mes nerfs et mon calme. Il a fallu de la patience (parfois feinte) et de la motivation (souvent fluctuante), mais j'y suis arrivée. Je suis heureuse d'avoir persévéré et fière d'être arrivée jusqu'au bout avec votre aide à **tous**.

De nouveau, merci à toutes les personnes qui ont accompagné ces années de thèse et qui ont contribué, de près ou de loin, à cette aventure. Un nouveau projet de recherche m'attend désormais, que j'aborde avec beaucoup d'enthousiasme. Je remercie enfin **Juliette Chevallier** et **François Bachoc** de me permettre d'ouvrir ce nouveau chapitre.

Liste des abréviations

AALasso	: <i>Adjusted Adaptative Lasso</i>
AIC	: <i>Akaike Information Criterion</i>
Alasso	: <i>Adaptative Lasso</i>
ASC	: cellule sécrétrice d'anticorps (<i>Antibody Secreting Cell</i>)
BCR	: <i>B Cell Receptor</i>
BDF	: <i>Backward Diffentiation Formulas</i>
BIC	: <i>Bayesian Information Criterion</i>
BICc	: <i>corrected BIC</i>
COSSAC	: <i>Conditional Sampling use for Stepwise Approach based on Correlation tests</i>
CPA	: Cellule Présentatrice d'Antigènes
EBE	: Estimateur empirique <i>a posteriori</i>
EDO	: Equation Différentielle Ordinaire
ELISA	: <i>Enzyme-Linked ImmunoSorbent Assay</i>
EM	: <i>Expectation-Maximization</i>
FDP	: <i>False Discovery Proportion</i>
FDR	: <i>False Discovery Rate</i>
FN	: faux négatifs
FNR	: <i>False Negative Rate</i>
FOA	: <i>First Order Approximation</i>
FOCE	: <i>First Order Conditional Estimation</i>
FP	: faux positifs
FPR	: <i>False Positive Rate</i>
Ig	: Immunoglobuline
Lasso	: <i>Least Absolute Shrinkage and Selection Operator</i>
LLPC	: plasmocytes à longue durée de vie (<i>Long-Lived Plasma Cell</i>)
MAMP	: <i>Microbe-Associated Molecular Patterns</i>
MAP	: Maximum <i>a posteriori</i>
MCMC	: Monte Carlo par chaîne de Markov
MCP	: <i>Minimax Concave Penalty</i>
NK	: cellules tueuses naturelles (<i>Natural Killer</i>)
NLEM	: Non-Linéaire à Effets Mixtes
PFER	: <i>Per-Family Error Rate</i>
REMixed	: <i>Regularized Estimation for nonlinear Mixed effect models</i>
SAEM	: <i>Stochastic Approximation Expectation-Maximization</i>
SAEMVS	: <i>SAEM Variable Selection</i>

SAMBA	: <i>Stochastic Approximation for Model Building Algorithm</i>
SCAD	: <i>Smoothly Clipped Absolute Deviation</i>
SCM	: <i>Stepwise Covariate Modeling</i>
SLPC	: plasmocytes à courte durée de vie (<i>Short-Lived Plasma Cell</i>)
TCR	: <i>T Cell Receptor</i>
TN	: vrais négatifs
TP	: vrais positifs
VPC	: <i>Visual Predictive Checks</i>

Production Scientifique et Enseignement

Publications scientifiques

- ▶ Gabaut, A., Thiébaud, R., Proust-Lima, C. & Prague, M. (2025). A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effects Model. *En révision pour Statistics in Medicine.* <https://doi.org/10.1101/2025.05.15.25327667>
- ▶ Prague, M., Gabaut, A., Ba, K., Alexandre, M., Hejblum, B.P. & Thiébaud, R. **Integrating gene expression from whole blood into mechanistic modeling of the immune system : illustration in a model of the antibody response to COVID vaccination.** *En préparation pour être soumis.*
- ▶ Gabaut, A., Hejblum, B.P., Bercu, A., Thiébaud, R., Proust-Lima, C. & Prague, M. **Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers.** *En préparation pour être soumis dans Biostatistics.*
- ▶ Gabaut, A., Proust-Lima, C. & Prague, M. **REMixed : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers in R.** <https://cran.r-project.org/web/packages/REMixed/vignettes/doc.pdf>.

Communications

Communications orales en conférences internationales

- ▶ Gabaut, A., Prague, M., **High-dimension Mechanistic Model Building using LASSO Approaches : Application to Ebola Vaccination.** 32^{ème} meeting Population Approach Group Europe (PAGE), 26 au 28 juin 2024, Rome, Italie.
- ▶ Gabaut, A., Bercu, A., Prague, M., Proust-Lima, C., **Regularized estimation of non-linear mixed effect models in the presence of high-dimensional biomarkers.** 10^{ème} Channel Network Conference (CNC), 19 au 21 mai 2025, Liège, Belgique.

Communications orales en conférences nationales

- ▶ Gabaut, A., Prague, M., **High-dimension Mechanistic Model Building using LASSO Approaches : Application to Ebola Vaccination.** Journées de Biostatistique 2023, 16 et 17 novembre 2023, Institut de Mathématiques de Toulouse, France.

- ▶ Gabaut, A., Prague, M., **Construction de modèles mécanistes en grande dimension avec une approche par LASSO : application à la vaccination contre le virus Ebola**. 55^{ème} Journée de Statistique de la SFdS, 27 au 31 mai 2024, Université de Bordeaux, France.
- ▶ Gabaut, A., Prague, M., **Construction de modèles mécanistes en grande dimension avec une approche par LASSO : application à la vaccination contre le virus Ebola**. Journée Scientifique Annuelle du RRI PHDS (Réseau de Recherche Impulsion “Public Health Data Science”), 3 juillet 2024, Université de Bordeaux, France.
- ▶ Gabaut, A., Bercu, A., Prague, M., Proust-Lima, C., **Regularized estimation of non-linear mixed effect models in the presence of high-dimensional biomarkers**. 56^{ème} Journée de Statistique de la SFdS, 2 au 6 juin 2025, Université Aix-Marseille, France.

Poster en conférence nationale

- ▶ Gabaut, A., Prague, M., **High-dimension Mechanistic Model Building using LASSO Approaches : Application to Ebola Vaccination**. 2^e journée des Jeunes Chercheurs du Bordeaux Population Health (BPH), 25 septembre 2023, Université de Bordeaux, France.

R Paquet

- ▶ LSAMBA : Lasso-SAMBA Algorithm. Disponible sur le CRAN.
<https://cran.r-project.org/web/packages/REMixed/index.html>
- ▶ REMixed : Regularized Estimation in Mixed Effect Model. Disponible sur le CRAN.
<https://cran.r-project.org/web/packages/LSAMBA/index.html>

Enseignement

- ▶ Master 1 Santé Publique :
 - 2023-2024 : EPISTA101 : Approche quantitative en Santé publique, TD, 10h.
 - 2023-2024 : Mise en application sur R, TP, 4h.
 - 2023-2024 : STG101 : Encadrement d’un projet tutoré, TD, 3h.
 - 2024-2025 : STG101 : Tutrice-Lectrice.
- ▶ License 3 Sage-Femme :
 - 2023-2024 : UE de Statistique, Cours/TD, 6h.
 - 2024-2025 : UE de Statistique, Cours/TD, 6h.

- ▶ 1^{ère} année d'internat de médecine :
 - 2023-2024 : UER Modélisation Statistique et Introduction au logiciel R, Cours/TD, 9h.
- ▶ 2^{ème} année d'Orthoptiste
 - 2024-2025 : Responsable de la partie Statistique de l'UE21 Statistique et Epidémiologie, CM et TD, 20h.

Introduction et Objectifs

La vaccination constitue l'un des outils les plus puissants en santé publique pour protéger les populations contre les maladies infectieuses. Des campagnes vaccinales historiques ont conduit à la quasi-élimination, voire à l'éradication, de certaines infections graves, comme par exemple la variole, à ce jour la seule maladie humaine éradiquée grâce à la vaccination d'après l'Organisation Mondiale de la Santé. Plus récemment, la pandémie de Covid-19 a rappelé l'importance du développement rapide de vaccins efficaces et sûrs, et a mis en lumière les défis associés à leur évaluation scientifique ainsi qu'à leur déploiement à grande échelle. Cette importance de la vaccination a stimulé le développement de nombreuses approches statistiques pour analyser les données recueillies lors d'études vaccinales (essais cliniques, études observationnelles), afin d'optimiser les stratégies de vaccination contre des maladies telles que le VIH, Ebola ou encore la Covid-19. Il apparaît dès lors crucial de mieux comprendre et analyser les mécanismes immunologiques sous-jacents à la vaccination, dans le but d'améliorer la protection des individus et la gestion des épidémies. Dans ce contexte, les outils mathématiques et statistiques jouent un rôle fondamental pour décrypter les mécanismes complexes d'interaction entre un agent pathogène, qu'il s'agisse de virus ou de bactéries, et le système immunitaire de l'hôte, ainsi que pour mieux comprendre le rôle de la vaccination dans ces interactions.

Parmi ces outils, les modèles non-linéaires à effets mixtes (NLEM), initialement très utilisés en pharmacométrie, se sont imposés comme une approche puissante pour la modélisation de la réponse vaccinale. Ces modèles statistiques, souvent qualifiés de mécanistes, s'appuient sur des systèmes d'équations différentielles ordinaires (EDO) permettant de décrire des dynamiques complexes. Ils permettent dans le contexte de la vaccination de capturer les mécanismes sous-jacents à l'infection et à l'injection de vaccin, par exemple les interactions cellulaires lors de la réponse immunitaire, ou encore la production d'anticorps, tout en tenant compte de plusieurs niveaux de variabilité : la variabilité inter-individuelle via des effets aléatoires, et la variabilité intra-individuelle ou résiduelle. Grâce à cette structure hiérarchique, les modèles à effets mixtes peuvent intégrer les données longitudinales de nombreux individus et dissocier les tendances moyennes de la variabilité biologique entre les sujets. Ils constituent ainsi un cadre robuste pour analyser les données issues d'essais vaccinaux et pour simuler différents scénarios de vaccination. Cependant, ces modèles traditionnels peinent à incorporer directement des informations issues de données biologiques très riches.

Or, l'essor récent des technologies de biologie à haut débit a permis de générer des données de plus en plus riches et complexes sur la réponse immunitaire à la vaccination. En particulier, l'accès facilité à des mesures d'expression génique offre de nouvelles perspectives pour caractériser finement les processus cellulaires activés après une vaccination. Ces données de grande dimension fournissent une image dynamique de l'activité biologique post-vaccination et peuvent, par exemple, être exploitées afin de prédire précocement la réponse immunitaire d'un individu à partir des variations d'expression génique observées peu après la vaccination. Cependant, l'utilisation de ces données à grande échelle s'accompagne de défis méthodologiques importants :

leurs mesures sont souvent bruitées et corrélées tant dans le temps, que entre gènes et entre sujets. De plus, leur grande dimension, avec des milliers de gènes mesurés pour quelques centaines voire dizaines de sujets parfois, complique l'intégration directe de ces informations dans des modèles statistiques classiques. Ces constats mettent en évidence un nouveau besoin méthodologique : développer des méthodes statistiques capables d'intégrer efficacement ces données de grande dimension au sein des modèles mécanistes de la réponse vaccinale, afin d'exploiter tout leur potentiel informatif.

Les travaux présentés dans cette thèse s'inscrivent précisément dans cet objectif. Ils visent à proposer des méthodes statistiques pour l'intégration de données de grande dimension dans les modèles non-linéaires à effets mixtes décrivant la réponse à une vaccination. Pour ce faire, nous adoptons une approche combinant les atouts des méthodes statistiques de régression pénalisée pour la sélection de variables avec des modèles dynamiques mécanistes basés sur des EDOs. Autrement dit, cette thèse propose d'enrichir la modélisation statistique des phénomènes biologiques observés lors d'une réponse vaccinale en y incorporant de nouvelles sources d'information de grande dimension, telles que les profils d'expression génique, tout en contrôlant la complexité du modèle grâce à des techniques de pénalisation adaptées. Ce faisant, nous cherchons à identifier, parmi un grand nombre de biomarqueurs, ceux qui sont les plus informatifs pour expliquer la variabilité observée des paramètres du modèle mécaniste, tout en maintenant une bonne qualité d'ajustement du modèle aux données.

Concrètement, deux axes méthodologiques complémentaires ont été développés au cours de cette thèse. Premièrement, nous nous sommes intéressés à l'intégration de biomarqueurs génétiques mesurés à des temps fixes, par exemple différents marqueurs par individu mesurés juste après la vaccination, et intégrés comme variables explicatives des modèles à effets mixtes non-linéaires. Nous proposons pour cela une méthode de sélection de covariables en grande dimension, fondée sur des techniques de régression pénalisée de type Lasso (*Least Absolute Shrinkage and Selection Operator*). Plus spécifiquement, nous avons développé un algorithme itératif de construction de modèle, baptisé **Lasso-SAMBA** (*Lasso Stochastic Approximation for Model Building Algorithm*), qui, à partir d'un système d'EDO mécaniste décrivant le phénomène biologique d'intérêt, permet d'estimer les paramètres du modèle et d'identifier quelles covariables génétiques expliquent au mieux la variabilité inter-individuelle observée dans les paramètres du modèle.

Deuxièmement, nous nous sommes intéressés à l'intégration de données transcriptomiques longitudinales, c'est-à-dire des trajectoires d'expression génique mesurées à plusieurs temps pour chaque individu, comme biomarqueurs mesurant les compartiments du système d'EDO dans les modèles non-linéaires à effets mixtes. Pour relever ce défi, un nouvel algorithme d'estimation pénalisée, appelé **REMixed** (*Regularized Estimation for non-linear Mixed effects models*), a été développé. Il vise à maximiser la vraisemblance pénalisée du modèle global tout en réalisant une sélection de variables parmi les trajectoires génomiques candidates.

La suite du manuscrit est organisée comme suit. Le Chapitre 1 présente le contexte biologique de ce travail : les principes de la vaccination et les bases de l'immunologie, en décrivant les

principaux mécanismes de la réponse immunitaire ainsi que les types de données couramment recueillies pour évaluer la réponse à un vaccin. Le Chapitre 2 introduit le cadre statistique général des modèles à effets mixtes, en particulier dans le cas non-linéaire, et passe en revue les méthodes d'estimations pour de tels modèles et les approches de régularisation, telles que le Lasso, qui seront utilisées par la suite. Les deux chapitres suivants sont consacrés aux développements méthodologiques réalisés durant la thèse. Le Chapitre 3 décrit en détail la méthode de sélection de covariables en grande dimension (Lasso-SAMBA) permettant d'intégrer des mesures génétiques individuelles à temps fixe dans un modèle non-linéaire à effets mixtes, et présente son application aux données des essais vaccinaux contre le virus Ebola et varicelle-zona. Le Chapitre 4 expose la méthode d'intégration de trajectoires longitudinales d'expression génique (REMixed), en formulant l'hypothèse de l'existence d'un compartiment latent relié à ces données omiques. Dans ce chapitre, l'algorithme d'estimation pénalisée ainsi que son application aux données de l'essai vaccinal contre le virus SARS-CoV-2 y sont présentés ; les aspects liés à l'implémentation de la méthode y sont également détaillés, notamment les choix algorithmiques permettant de traiter des modèles de grande dimension. Ceci inclut l'utilisation de techniques d'approximation et d'optimisation numérique adaptées aux modèles à effets mixtes de grande dimension, afin de garantir la faisabilité et la fiabilité des estimations obtenues et d'assurer la reproductibilité par la communauté scientifique. Enfin, le dernier Chapitre 5 conclut ce travail de thèse en synthétisant les résultats obtenus et en discutant les perspectives ouvertes, tant sur le plan méthodologique que sur le plan biologique.

– Chapitre 1 –

Contexte Biologique

Table des matières

1.1	Le système immunitaire	19
1.1.1	L'immunité innée	20
1.1.2	L'immunité adaptative	21
1.1.3	L'immunité adaptative humorale	24
1.1.4	L'immunité adaptative à médiation cellulaire	26
1.2	Essais cliniques vaccinaux	27
1.2.1	Cadre expérimental des essais vaccinaux	28
1.2.2	Données immunologiques d'intérêt dans cette thèse	29
1.2.2.1	Dosages d'anticorps (Luminex, ELISA)	29
1.2.2.2	Profil d'expression génique (RNA-seq)	30
1.2.3	Jeux de données utilisés	32
1.2.3.1	Essai PREVAC-UP : vaccination contre Ebola	32
1.2.3.2	Essai sur le vaccin ZOSTAVAX contre le zona	33
1.2.3.3	Essai sur le vaccin ARNm contre le SARS-CoV-2	34

Dans ce chapitre, nous présentons les concepts clés du système immunitaire qui ont motivé les modélisations employées dans les Chapitres 3 et 4. Ce chapitre constitue une introduction ; une présentation plus détaillée des mécanismes sous-jacents à la réponse immunitaire est disponible notamment dans *Cellular and molecular immunology* par Abbas *et al.* (2021). En particulier, les mécanismes d'intérêt ici sont l'évolution des dynamiques d'anticorps suite à l'injection de vaccins telles qu'elles sont observées dans les applications contre la varicelle et le virus Ebola dans le Chapitre 3 et contre le virus SARS-CoV-2 dans le Chapitre 4.

1.1 Le système immunitaire

L'immunité désigne la protection contre les maladies, notamment infectieuses. Le système immunitaire implique différents types de cellules et de molécules qui, lorsqu'un pathogène est détecté dans l'organisme, vont se coordonner et réagir afin de l'éliminer lors de la réponse immunitaire (Chaplin, 2010). Deux types de réponses immunitaires sont au centre de l'immunité (Figure 1.1) : la réponse immunitaire innée, réponse immédiate chez tout individu sans immunité

préalable, et la réponse immunitaire adaptative, permettant une réponse plus efficace grâce à une mémoire immunitaire lorsque le pathogène a déjà été rencontré (Hoebe *et al.*, 2004). L'immunité innée, bien que fondamentale comme première ligne de défense, ne sera pas modélisée en détail dans le cadre de cette thèse. Elle est néanmoins présentée dans ce chapitre introductif afin de situer son rôle et permettre de bien introduire ensuite l'immunité adaptative, qui constitue le cœur de la vaccination. Les développements méthodologiques et applications de cette thèse se concentreront en particulier sur la modélisation mécaniste de la réponse immunitaire adaptative, et plus spécifiquement de sa composante humorale.

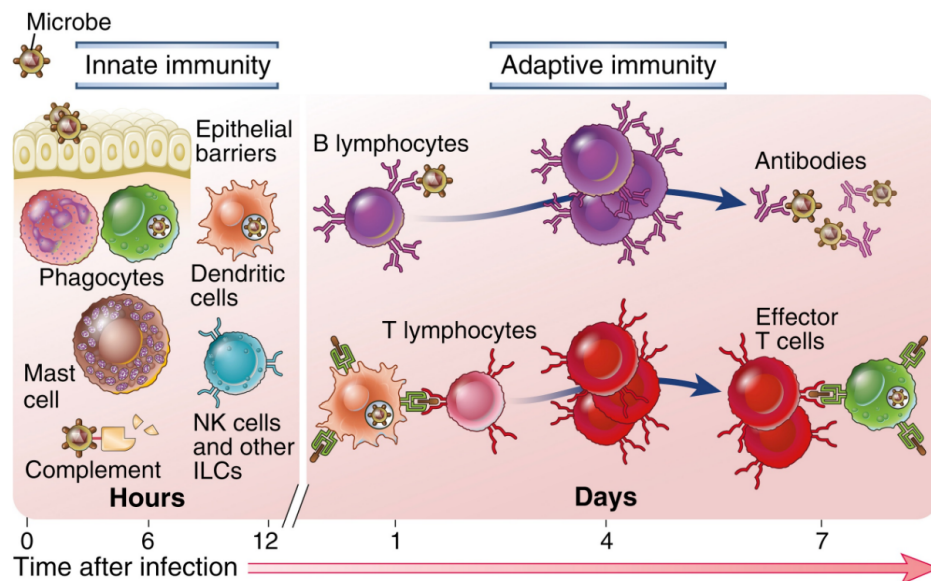


FIGURE 1.1 – **Immunité innée et adaptative** (extrait de Abbas *et al.* (2021)).

Certaines cellules représentées sur la figure, comme les cellules lymphoïdes innées (ILC) ou les mastocytes, ne sont pas détaillées ici mais participent également à la réponse innée via la production de médiateurs inflammatoires.

NK : cellules tueuses naturelles (*Natural Killer*)

1.1.1 L'immunité innée

L'immunité innée, ou naturelle, de l'organisme regroupe les mécanismes de défense immédiats et non spécifiques contre les infections. Présente dès la naissance, elle constitue une première ligne de défense qui agit rapidement pour contenir les agents infectieux. Par contraste, l'immunité adaptative se déploie plus lentement, de façon spécifique à chaque pathogène, et génère une mémoire immunologique durable (Hoebe *et al.*, 2004). Bien que dépourvue de mémoire, l'immunité innée est indispensable pour freiner l'infection dès les phases initiales et pour orienter la mise en place de la réponse adaptative. Elle repose sur la reconnaissance de pathogènes par des récepteurs immunitaires préexistants, ce qui permet une réaction immédiate. On distingue classiquement deux composantes majeures de l'immunité innée : d'une part des barrières physiques et chimiques qui empêchent l'entrée des pathogènes telles que la peau et les muqueuses, et d'autre part des mécanismes cellulaires chargés d'éliminer les agents ayant franchi ces barrières.

La seconde composante citée de l'immunité innée est assurée par divers leucocytes, ou globules blancs. Les cellules phagocytaires y jouent un rôle central : les neutrophiles, les monocytes (et leurs dérivés tissulaires, les macrophages) ainsi que les cellules dendritiques reconnaissent les agents infectieux, en détectant des motifs qui leur sont typiques (appelés MAMP, *Microbe-Associated Molecular Patterns*) grâce à des récepteurs spécifiques présents à leur surface. Ils détruisent ensuite ces microbes par phagocytose, processus cellulaire par lequel certaines cellules sont capables d'ingérer des particules étrangères, des micro-organismes pathogènes ou, comme dans le cas d'infection virale, des cellules infectées. L'activation de ces phagocytes induit la libération de médiateurs comme les cytokines et chimiokines, des molécules de signalisation permettant de moduler et coordonner les réponses immunitaires, qui déclenchent une inflammation locale et attirent d'autres leucocytes vers le site infecté. En parallèle, les lymphocytes NK (*Natural Killer*) détruisent les cellules infectées en libérant des substances comme la perforine et les granzymes qui percent la membrane de la cellule infectée et déclenchent sa mort programmée, appelée apoptose. L'action coordonnée de l'ensemble de ces composants permet de contenir l'infection dès les premières phases, avant même l'entrée en jeu de l'immunité adaptative.

Parmi ces cellules, les cellules dendritiques ou cellules présentatrices d'antigènes (CPAs) occupent une place particulière à l'interface avec l'immunité adaptative. Présentes dans les tissus, elles capturent les agents pathogènes par phagocytose puis migrent vers les ganglions lymphatiques, où elles présentent les antigènes, *i.e.* des fragments de ces agents pathogènes capables d'être reconnus par le système immunitaire, aux lymphocytes T, ou cellules T, pour permettre une réponse adaptée au pathogène rencontré. Ce processus de présentation d'antigène s'accompagne de l'activation des cellules dendritiques, qui sécrètent des cytokines influençant la réponse adaptative. Les cellules dendritiques font ainsi le lien entre immunité innée et adaptative en activant les lymphocytes spécifiques. Plus globalement, ces deux branches du système immunitaire coopèrent étroitement pour garantir une réponse complète et efficace. Durant la phase initiale d'une infection, les cytokines libérées par les cellules innées, en particulier par les phagocytes activés, créent un environnement inflammatoire favorable au recrutement et à l'activation d'autres leucocytes, facilitant ainsi la mise en place d'une réponse immunitaire adaptative ciblée. À l'inverse, l'immunité adaptative contribue également à renforcer l'efficacité de la réponse innée grâce à la production d'anticorps spécifiques, protéines capables non seulement de neutraliser directement les pathogènes mais aussi d'améliorer leur reconnaissance par les cellules phagocytaires via un processus appelé opsonisation. Enfin, l'activation du système du complément, une cascade enzymatique complexe initiée notamment par ces anticorps, conduit à la destruction directe des pathogènes et renforce l'inflammation locale, soulignant ainsi la complémentarité fonctionnelle essentielle entre ces deux formes d'immunité (Vivier *et al.*, 2008).

1.1.2 L'immunité adaptative

Après la réaction immédiate et non spécifique de l'immunité innée, l'organisme mobilise une deuxième ligne de défense plus ciblée : la réponse immunitaire adaptative, également appelée

immunité acquise, ou spécifique. Contrairement à l'immunité innée, qui réagit rapidement mais de manière générique, l'immunité adaptative met plusieurs jours à se déployer pleinement. En effet, seuls quelques lymphocytes spécifiques reconnaissent initialement un antigène donné, et ils doivent d'abord être activés puis proliférer en grand nombre, lors d'un processus d'expansion clonale, pour qu'une réponse efficace puisse se constituer.

La réponse immunitaire adaptative se distingue de l'immunité innée par plusieurs caractéristiques fondamentales. Sa première propriété est sa spécificité : elle repose sur la reconnaissance d'antigènes uniques, portés par les agents pathogènes. Elle met en jeu deux types de lymphocytes, les lymphocytes B et les lymphocytes T, qui portent à leur surface des récepteurs antigéniques identiques entre eux sur un même lymphocyte, mais différents d'un lymphocyte à l'autre, issu d'un processus de recombinaison génétique aléatoire qui confère à l'organisme un répertoire diversifié pour reconnaître les pathogènes. Grâce à cette diversité, l'immunité adaptative est capable de reconnaître des millions d'antigènes distincts, bien au-delà du spectre limité de motifs moléculaires pris en charge par les récepteurs de l'immunité innée.

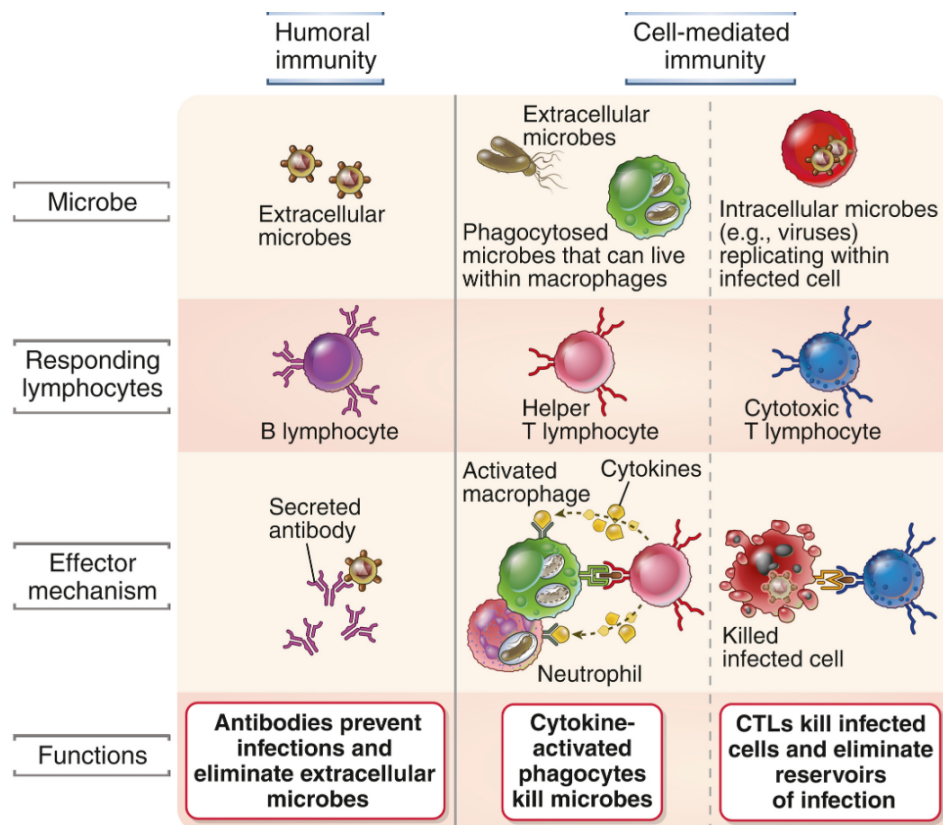


FIGURE 1.2 – Immunité humorale et immunité cellulaire (extrait de Abbas *et al.* (2021)).

L'immunité adaptative comprend deux volets principaux (Figures 1.2 et 1.3), étroitement liés : l'immunité humorale et l'immunité cellulaire. L'immunité humorale est assurée par les lymphocytes B et se manifeste par la production d'anticorps circulants, ou immunoglobulines (Ig), dirigés contre des antigènes spécifiques. Ces anticorps neutralisent les agents pathogènes et facilitent leur élimination, en particulier tant que ceux-ci se trouvent à l'extérieur des cellules.

L'immunité cellulaire, de son côté, est médiée par les lymphocytes T : principalement les T CD4+, dits auxiliaires ou *helpers*, et T CD8+, dits cytotoxiques. Elle se traduit par l'action directe de ces lymphocytes sur les cellules infectées ou anormales et par la libération de cytokines qui activent d'autres cellules immunitaires, permettant de détruire les agents infectieux résidant à l'intérieur des cellules. Ces deux types d'immunité sont décrits plus en détail dans les sous-sections suivantes.

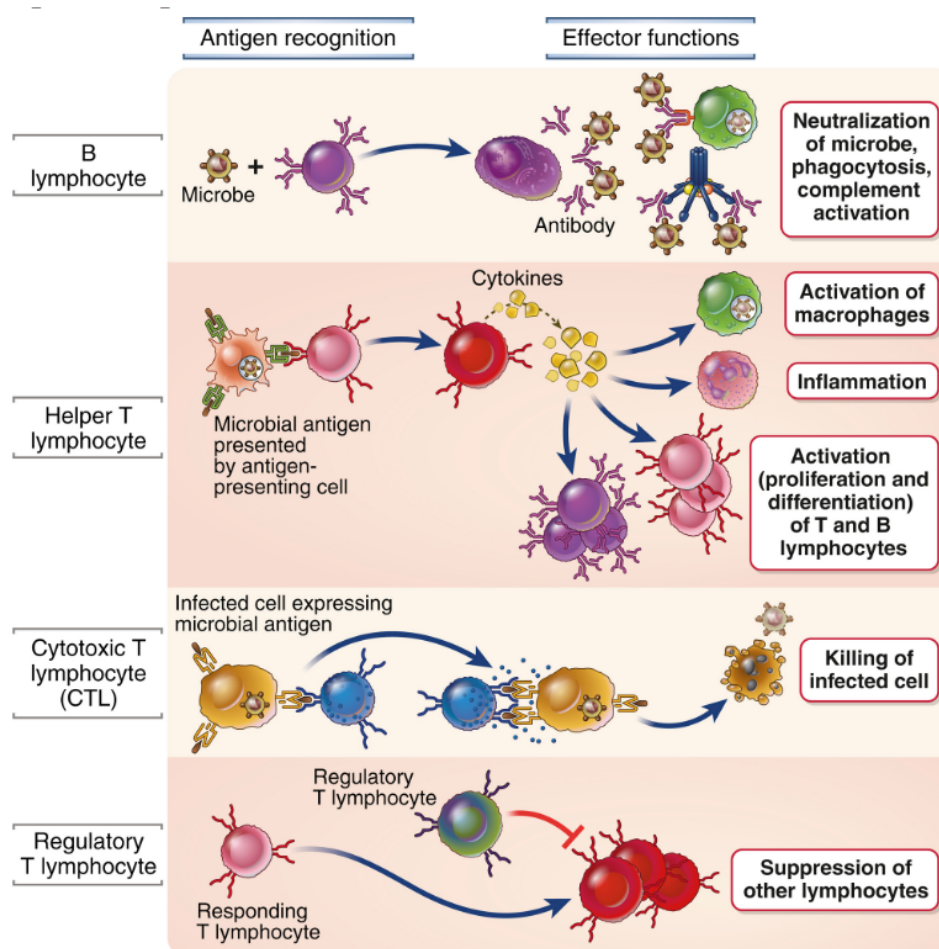


FIGURE 1.3 – **Classes de lymphocytes** (extrait de Abbas *et al.* (2021)).

Les lymphocytes B reconnaissent de nombreux types d'antigènes et se différencient en cellules sécrétrices d'anticorps. Les lymphocytes T auxiliaires reconnaissent les antigènes présents à la surface des CPA et sécrètent des cytokines, qui stimulent divers mécanismes d'immunité et d'inflammation. Les lymphocytes T cytotoxiques reconnaissent les antigènes présents dans les cellules infectées et les détruisent. Les lymphocytes T régulateurs régulent les réponses immunitaires.

Les étapes de la réponse immunitaire adaptative (Figure 1.4) sont : (1) reconnaissance de l'antigène, (2) activation des lymphocytes, (3) phase effectrice d'élimination du pathogène ou des cellules infectées, (4) phase de contraction, où les lymphocytes activés meurent par apoptose, rétablissant ainsi un équilibre cellulaire, (5) mise en place de la mémoire immunitaire. Ces principes s'appliquent à la fois à l'immunité humorale, médiée par les lymphocytes B, et à l'immunité à médiation cellulaire, médiée par les lymphocytes T, et sont détaillés pour chacun de ces cas dans les parties suivantes.

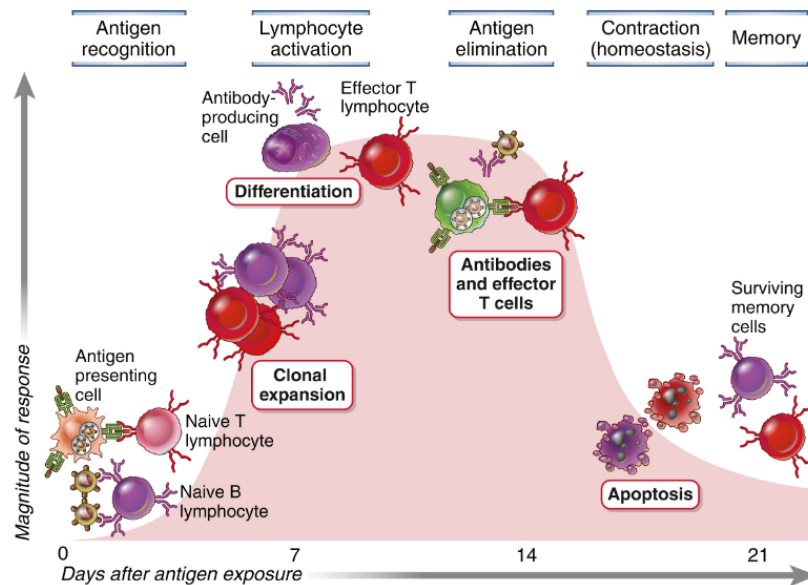


FIGURE 1.4 – **Étapes de la réponse immunitaire adaptative** (extrait de [Abbas et al. \(2021\)](#)).

L'axe des ordonnées représente une mesure relative de l'intensité de la réponse.

Dans le cadre de cette thèse, l'accent sera mis sur la réponse immunitaire humorale, qui constitue un déterminant majeur de l'efficacité vaccinale. En effet, la dynamique des plasmocytes et des anticorps sécrétés sera au cœur des modèles mécanistes développés dans les chapitres suivants. Si l'immunité à médiation cellulaire et certaines voies de l'immunité innée jouent également un rôle important dans la réponse au vaccin, elles ne seront pas modélisées directement mais restent utiles à présenter pour situer la complexité globale de la réponse immunitaire et replacer le travail réalisé dans un cadre biologique plus large.

1.1.3 L'immunité adaptative humorale

L'immunité humorale est assurée par les lymphocytes B, ou cellules B, dont la fonction principale est de produire des anticorps dirigés contre des antigènes spécifiques. Chaque lymphocyte B possède à sa surface des récepteurs spécifiques, appelés *B Cell Receptor* ou BCR, capables de se lier à un antigène particulier. Ainsi, un lymphocyte B donné ne reconnaît qu'un seul déterminant antigénique spécifique. Lorsque ce déterminant est rencontré dans l'organisme, par exemple sur un pathogène extracellulaire, il se fixe sur le BCR d'un lymphocyte B correspondant et y est reconnu. Cette reconnaissance de l'antigène constitue le signal initial de l'activation du lymphocyte B. Le plus souvent, pour qu'un lymphocyte B soit pleinement activé, il doit recevoir aussi des signaux d'activation provenant d'un lymphocyte T CD4+, *helper*. Une fois activé, le lymphocyte B entre en prolifération : il multiplie ses cellules filles, formant un clone de lymphocytes spécifiques identiques : c'est la phase d'expansion clonale. En parallèle, ces cellules filles amorcent un processus de différenciation pour devenir des cellules effectrices de type B ([den Haan et al., 2014](#)).

Après quelques jours d'expansion clonale, une partie des lymphocytes B différenciés se transforme en plasmocytes. Les plasmocytes sont des cellules B effectrices spécialisées dans la sécrétion d'anticorps. Chaque plasmocyte produit de grandes quantités d'anticorps possédant la même spécificité que le BCR initial du clone. Ces anticorps, appelés immunoglobulines, se déclinent en plusieurs classes (IgM, IgG, IgA, IgE, IgD), qui remplissent des rôles complémentaires dans la réponse immunitaire. Lors d'une réponse primaire, les premiers anticorps sécrétés sont majoritairement des IgM, de grande taille et efficaces pour activer le complément, une cascade de protéines plasmatiques décrite précédemment, qui contribue à la destruction des pathogènes et à l'amplification de la réponse inflammatoire. Dans un second temps, les plasmocytes produisent principalement des IgG, qui sont plus petites, persistent plus longtemps dans la circulation, et sont plus efficaces pour neutraliser les pathogènes et faciliter leur élimination. En fonction de la porte d'entrée infectieuse ou du tissu concerné, d'autres classes d'immunoglobulines peuvent également être produites. Par exemple, il existe les IgA présentes dans les muqueuses ou les IgE actives dans les réponses aux allergènes. Néanmoins, dans le cadre de cette thèse, nous nous intéressons particulièrement aux IgG, la classe la plus abondante. Parallèlement, une autre fraction des lymphocytes B activés ne se différencie pas en plasmocytes mais en lymphocytes B mémoires à longue vie. Ces cellules mémoires restent au repos dans l'organisme après la résolution de l'infection initiale, prêtes à se réactiver rapidement en cas de nouvelle exposition au même antigène.

Les anticorps produits par les plasmocytes se répandent dans la circulation sanguine et les autres fluides biologiques pour agir au plus près du site d'infection. Ils peuvent se lier spécifiquement à l'antigène cible et contribuent de multiples façons à l'élimination du pathogène. Par exemple, la liaison d'un anticorps peut neutraliser un micro-organisme ou une toxine en bloquant son interaction avec les cellules hôtes. Les anticorps peuvent aussi opsoniser les pathogènes, c'est-à-dire les marquer pour faciliter leur ingestion et destruction par les phagocytes tels que les macrophages et les neutrophiles. Ils peuvent encore activer le système du complément, un ensemble de protéines sanguines qui participent à la destruction des bactéries et au mécanisme de l'inflammation. L'ensemble de ces mécanismes confère à l'immunité humorale une grande efficacité contre les agents infectieux extracellulaires, avant qu'ils n'infectent les tissus.

La phase initiale de réponse humorale, appelée réponse immunitaire primaire, est relativement lente et de faible amplitude. En effet, lors d'une première infection, l'activation, la prolifération et la différenciation des rares lymphocytes B spécifiques nécessitent du temps : il faut souvent plusieurs jours pour que la production d'anticorps atteigne un niveau significatif. Une fois cette réponse primaire établie et l'infection contrôlée, les plasmocytes générés finissent par mourir, mais les lymphocytes B mémoires persistent, principalement dans les ganglions lymphatiques et la moelle osseuse, où ils peuvent rester en état de veille pendant des années. Si le même antigène est rencontré à nouveau plus tard, ces lymphocytes mémoires l'identifieront immédiatement et se réactiveront en masse pour proliférer et se différencier en nouveaux plasmocytes sécréteurs d'anticorps. La réponse immunitaire secondaire ainsi obtenue est beaucoup plus rapide et plus intense que la primaire, permettant une neutralisation rapide du pathogène. Ce principe de mémoire immunitaire humorale est exploité par les vaccins, qui consistent à provoquer une

première exposition à un antigène, généralement sous une forme inactivée, afin de générer des lymphocytes mémoires protecteurs. C'est également sur ce processus que se concentrent les applications de cette thèse : modéliser la dynamique des cellules sécrétrices d'anticorps et la production d'IgG au cours du temps afin de caractériser quantitativement l'effet d'une injection de vaccin.

1.1.4 L'immunité adaptative à médiation cellulaire

L'immunité cellulaire, ou à médiation cellulaire, repose sur l'action des lymphocytes T, notamment les T CD4+ et T CD8+. À l'inverse des lymphocytes B, les lymphocytes T ne peuvent pas reconnaître un antigène libre directement. Leur récepteur spécifique, le *T Cell Receptor* ou TCR, ne reconnaît des fragments peptidiques d'antigène que si ceux-ci sont présentés à la surface d'une cellule présentatrice. Dans ce processus, une cellule présentatrice d'antigène, typiquement une cellule dendritique, joue un rôle clé en internalisant le pathogène, en dégradant ses antigènes en peptides, puis en exposant ces peptides à sa surface associés à d'autres molécules, ce qui permet leur reconnaissance par le TCR du lymphocyte T spécifique.

Les lymphocytes T CD8+ et les lymphocytes T CD4+ remplissent des rôles distincts grâce à des récepteurs qui leur confèrent une reconnaissance spécifique. Les T CD8+ sont capables de détecter les cellules anormales, par exemple celles qui sont infectées par un virus et qui présentent alors des fragments du pathogène à leur surface. Les lymphocytes T CD4+ reconnaissent les fragments d'antigènes présentés par des cellules présentatrices d'antigène, comme les cellules dendritiques, les macrophages et les lymphocytes B. Grâce à ce système de présentation, chaque lymphocyte T n'est activé que par l'antigène précis pour lequel son TCR est spécifique, de manière analogue aux lymphocytes B, mais via un mécanisme indirect faisant intervenir une autre cellule immunitaire.

Lorsqu'un lymphocyte T naïf reconnaît son antigène présenté sur une cellule présentatrice et reçoit les signaux de co-stimulation nécessaires, indispensables pour éviter une activation inappropriée et prévenir l'auto-immunité, il s'active à son tour. La cellule T activée entre alors en division et prolifère de façon clonale, comme chez les lymphocytes B, produisant de nombreuses cellules filles spécifiques du même antigène. Parallèlement, ces lymphocytes T proliférants se différencient en différents types de cellules T effectrices aux fonctions spécialisées. Les lymphocytes T CD8+ activés se différencient en lymphocytes T cytotoxiques pleinement matures, dont la fonction est de reconnaître et d'éliminer directement les cellules cibles porteuses de l'antigène. Une fois attachés à une cellule infectée ou tumorale qui présente le peptide étranger, ces lymphocytes libèrent des protéines cytotoxiques induisant la destruction de la cellule cible.

Les lymphocytes T CD4+ activés, de leur côté, se différencient en lymphocytes T auxiliaires et en une sous-population de lymphocytes T régulateurs. Ces cellules n'ont pas d'activité cytotoxique directe mais orchestrent la réponse immunitaire par la sécrétion de diverses cytokines et par des interactions de co-stimulation avec d'autres leucocytes. De manière générale, les lymphocytes T auxiliaires fournissent des signaux d'aide indispensables à l'efficacité de la réponse

adaptative globale : ils stimulent les fonctions des phagocytes, optimisent la production d'anticorps par les lymphocytes B et favorisent l'activation des lymphocytes T CD8+ cytotoxiques. Les lymphocytes T régulateurs, quant à eux, ont pour objectif de réguler la réponse immunitaire aux pathogènes et aux auto-antigènes, antigènes de l'organisme qui stimulent le système immunitaire, et peuvent provoquer une maladie auto-immune (Kondělková *et al.*, 2010).

L'immunité cellulaire protège l'organisme par l'élimination des cellules infectées ou dysfonctionnelles et par l'activation des autres maillons de l'immunité. Les lymphocytes T cytotoxiques détruisent la source d'infection en éliminant l'organisme des cellules qui abritent des pathogènes en leur sein. Les lymphocytes T auxiliaires, en libérant des cytokines, stimulent les phagocytes, les macrophages et les neutrophiles, afin qu'ils dégradent au mieux les agents infectieux ingérés, et ils orchestrent les réponses immunitaires en recrutant et en activant d'autres lymphocytes (Laidlaw *et al.*, 2016).

Comme pour l'immunité humorale, la phase effectrice active de l'immunité cellulaire est suivie d'une phase de contraction où la majorité des lymphocytes T activés meurent une fois l'antigène éliminé. Néanmoins, une fraction de ces lymphocytes T activés survit durablement sous forme de lymphocytes T mémoires. Ces cellules mémoires se répandent dans l'organisme, notamment dans les ganglions lymphatiques et la circulation, et assurent une mémoire immunitaire au long terme. En cas de nouvelle infection par le même pathogène, elles pourront détecter immédiatement la présence de l'antigène et déclencher une réaction secondaire rapide et intense, généralement suffisante pour empêcher le développement de la maladie. Ainsi, les lymphocytes T mémoires, auxiliaires et cytotoxiques, contribuent, aux côtés des lymphocytes B mémoires, à une protection immunitaire durable contre les rechutes ou les nouvelles infections par le même agent. Même si l'immunité à médiation cellulaire ne sera pas modélisée directement dans cette thèse, elle reste importante à introduire car les données transcriptomiques utilisées dans les applications reflètent l'activité de différents compartiments immunitaires. En particulier, certains profils d'expression génique peuvent fournir des informations indirectes sur l'activation de voies cellulaires ou de l'immunité innée, venant compléter l'interprétation des dynamiques humorales étudiées.

1.2 Essais cliniques vaccinaux

La vaccination constitue depuis plus de deux siècles un moyen de prévention majeur contre les maladies infectieuses, ayant permis de contrôler, voire d'éradiquer certaines d'entre elles. Historiquement, la première manipulation du système immunitaire à des fins de protection remonte au XVIII^e siècle, avec la découverte du vaccin contre la variole par Edward Jenner. Il observa que les individus ayant contracté une forme bénigne du virus (le cowpox) développaient une réponse atténuée lors d'une seconde exposition à la variole humaine (Riedel, 2005). Cette découverte a initié le développement de nombreux autres vaccins et motivé les recherches ultérieures. Les progrès en immunologie et en biologie moléculaire ont ensuite permis la mise au point de vaccins efficaces contre des pathogènes variés, comme le tétanos, la polio ou la fièvre jaune (Delany *et al.*, 2014; Plotkin, 2014).

Le rôle central de la mémoire adaptative est au cœur du principe vaccinal. L'objectif est de provoquer une première exposition contrôlée du système immunitaire à un antigène d'un pathogène donné, afin d'induire la formation de lymphocytes effecteurs et mémoires spécifiques sans provoquer la maladie. L'injection d'un antigène atténué, inactivé ou recombinant entraîne l'activation des lymphocytes T et B spécifiques ; ceux-ci prolifèrent, se différencient, produisent des anticorps neutralisants, et laissent derrière eux des cellules mémoires capables d'orchestrer une réponse rapide en cas d'exposition ultérieure. Ce déclenchement anticipé et maîtrisé de la mémoire immunitaire a permis de contrôler ou d'éradiquer de nombreuses maladies infectieuses ([Pulendran et Ahmed, 2006](#)). Il existe donc différentes catégories de vaccins selon la nature de l'antigène et la stratégie utilisée : vaccins vivants atténués, contenant un pathogène vivant mais rendu non virulent (comme le vaccin contre la varicelle), vaccins inactivés contenant le pathogène tué entier (comme certains vaccins polio), ou encore vaccins inertes composés uniquement de fragments antigéniques purifiés (par exemple le vaccin contre l'hépatite B).

Dans le cadre de cette thèse, nous nous intéressons aux réponses immunitaires induites par différents vaccins évalués dans des essais cliniques. Ces essais permettent d'étudier la sécurité, l'immunogénicité et l'efficacité des vaccins, mais aussi de mieux comprendre les mécanismes immunologiques sous-jacents à la protection vaccinale. La section ci-dessous présente les grandes étapes du développement vaccinal, les types de données immunologiques générées, ainsi que les essais cliniques utilisés dans ce travail.

1.2.1 Cadre expérimental des essais vaccinaux

Un essai clinique vaccinal est une étude expérimentale conduite chez des participants volontaires pour évaluer la sécurité, l'immunogénicité et l'efficacité d'un vaccin. Ces essais permettent notamment de mieux comprendre les mécanismes immunitaires impliqués, d'identifier les schémas vaccinaux optimaux (nombre de doses, intervalles entre les doses, etc.) et d'établir la durée de la protection immunitaire. Typiquement, les essais cliniques se déroulent en plusieurs phases : la phase I évalue principalement la sécurité, la phase II confirme la sécurité et évalue l'immunogénicité sur un nombre limité de participants, et la phase III vise à démontrer l'efficacité clinique sur une population large, schématisé en [Figure 1.5](#).

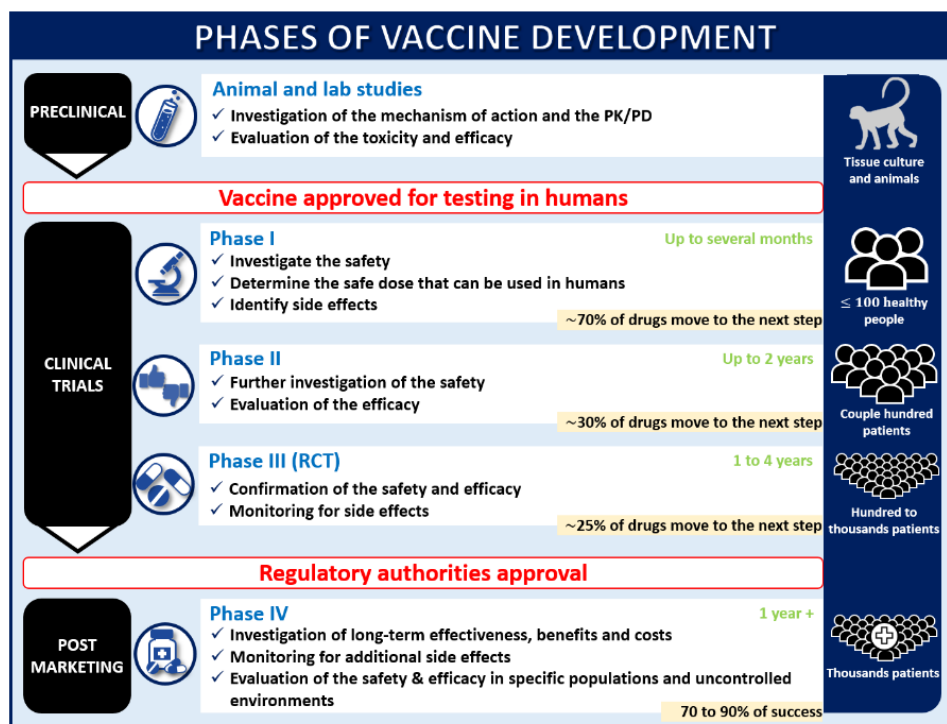


FIGURE 1.5 – Les étapes du développement vaccinal (extrait de Alexandre (2022)).

1.2.2 Données immunologiques d'intérêt dans cette thèse

1.2.2.1 Dosages d'anticorps (Luminex, ELISA)

La méthode la plus classique pour quantifier la réponse humorale à un vaccin est le dosage d'anticorps spécifiques de l'antigène vaccinal, généralement par ELISA (*Enzyme-Linked ImmunoSorbent Assay*). Un test ELISA utilise un antigène immobilisé pour capturer les anticorps présents dans le sérum du participant, puis un révélateur enzymatique pour mesurer la quantité d'anticorps liés. Le résultat est souvent exprimé soit en concentration (en $\mu\text{g/mL}$ ou en unités arbitraires corrélées à une concentration) soit en titre d'anticorps (c'est-à-dire une dilution du sérum au-delà de laquelle les anticorps restent détectables). Les données ELISA sont typiquement des mesures longitudinales : on prélève du sang avant vaccination puis à divers temps post-vaccination (jours, semaines, mois) afin de suivre la cinétique des anticorps. On obtient ainsi des courbes individuelles de concentration d'IgG, ou un autre isotype selon les cas, en fonction du temps, qui traduisent l'augmentation initiale due à la réponse primaire, le pic de réponse, puis la décroissance progressive associée à la mise en mémoire.

En complément des mesures ELISA, des techniques de dosage multiplex comme le Luminex sont également utilisées. La technologie Luminex repose sur des micro-billes couplées à différents antigènes ou réactifs, permettant de mesurer simultanément, dans un même échantillon, plusieurs cibles. Dans le contexte vaccinal, un essai Luminex peut servir par exemple à doser en parallèle les anticorps dirigés contre plusieurs antigènes, ce qui est utile pour les vaccins multivalents ou pour évaluer la spécificité fine des anticorps contre différentes souches ou variants d'un virus.

On peut aussi mesurer différents isotypes d'anticorps (IgG, IgM, IgA) en une seule analyse. Le résultat est similaire à un ELISA en termes de titre ou concentration, mais avec une richesse multidimensionnelle, c'est-à-dire plusieurs mesures simultanées par échantillon.

Ces données sérologiques quantitatives sont cruciales car elles sont souvent considérées comme des critères de jugement immunologiques dans les essais. On peut ainsi comparer les groupes vaccin et placebo en termes de titres moyens d'anticorps, estimer la proportion de sujets atteignant un certain seuil de titre dans chaque groupe, ou modéliser la décroissance des anticorps au cours du temps.

1.2.2.2 Profil d'expression génique (RNA-seq)

Les gènes sont des fragments d'ADN (Acide DésoxyriboNucléique) qui contiennent l'information nécessaire à la synthèse de molécules fonctionnelles, le plus souvent des protéines. L'ADN, support de l'information héréditaire, est une molécule en double hélice composée d'une séquence ordonnée de nucléotides, complémentaire deux à deux - adénine (A) et thymine (T) ; cytosine (C) et guanine (G). Chaque gène occupe une position donnée sur l'ADN et permet la synthèse d'une molécule fonctionnelle, typiquement une protéine, suivant des étapes précises. D'abord, un gène, porté par l'ADN, est transcrit sous forme d'ARN (Acide RiboNucléique) messenger (ARNm), une molécule en simple brin, complémentaire à l'ADN, où la thymine (T) est remplacée par l'uracile (U). Celle-ci est ensuite traduite en protéine par des structures spécialisées appelées complexes ribonucléiques. Ce mécanisme en deux étapes - transcription puis traduction - constitue le fondement de l'expression génique, permettant aux cellules de fabriquer les protéines dont elles ont besoin pour assurer leurs fonctions et répondre aux signaux de leur environnement.

L'expression génique désigne alors l'ensemble de ce processus, permettant la production d'une molécule fonctionnelle à partir d'un gène converti en ARNm, elle est donc mesurable par la quantité d'ARNm transcrite depuis un gène. Il est important de noter que toutes les cellules d'un organisme, à quelques exceptions près, partagent le même génome, mais expriment uniquement une fraction différente de leurs gènes à un moment donné. L'expression génique d'une cellule est régulée par des facteurs de son environnement et reflète donc son activité à un moment donné, et étudier l'expression des gènes est donc crucial pour permettre de comprendre quels gènes sont exprimés ou non selon le contexte (Alberts *et al.*, 2014).

Avec les avancées technologiques de séquençage à haut débit, mesurer simultanément l'expression de milliers de gènes dans un échantillon donné est de moins en moins coûteux et complexe, rendant de plus en plus accessibles ce genre de données. Il est par exemple possible maintenant de mesurer les profils d'expression génique d'un individu à partir d'un simple prélèvement capillaire, réalisé par piqûre au doigt. Cette approche, comparable à celle utilisée pour la mesure de la glycémie, permet d'envisager des suivis longitudinaux fréquents, même en dehors d'un cadre hospitalier (Obermoser *et al.*, 2013; Rinchai *et al.*, 2022). Elle a notamment été mise en œuvre dans l'étude COVAX (Rinchai *et al.*, 2022), où des auto-prélèvements quotidiens ont permis de suivre les profils transcriptomiques après vaccination, révélant des signatures

précoces de l'immunité innée.

L'étude COVERAGE-Immuno (Thiébaud *et al.*, 2025) a renforcé la validité de cette approche dans un contexte clinique. Chez des patients suivis à domicile après infection par le SARS-CoV-2, les profils d'expression génique obtenus à partir de sang capillaire ont été comparés à ceux mesurés par prélèvement sanguin plus classique, avec une concordance satisfaisante. Cette étude a permis un suivi rapproché, avec des prélèvements tous les deux jours, mettant en évidence des dynamiques d'expression génique cohérentes avec les mécanismes immunitaires attendus. Ces résultats confirment la faisabilité de l'usage de ce type de prélèvement pour étudier la réponse immunitaire de manière longitudinale, facilitée et peu invasive, ouvrant des perspectives intéressantes pour les futurs essais vaccinaux.

La technique la plus utilisée aujourd'hui pour quantifier l'expression génique est le séquençage de l'ARN, ou RNA-seq, développé dans les années 2000 (Nagalakshmi *et al.*, 2008; Weber, 2015), technique qui permet de quantifier l'abondance des transcrits d'ARNm, ou, autrement dit, le nombre de chaque ARNm spécifique, au sein d'un échantillon. En *bulk* RNA-seq, que nous utilisons dans cette thèse, l'ARN est extrait d'un échantillon constitué d'un mélange de nombreuses cellules, fournissant ainsi un profil moyen d'expression génique pour cet ensemble cellulaire, par exemple une goutte de sang prélevé au bout d'un doigt comme mentionné précédemment. Ce type de données se distingue du *single-cell* (sc) RNA-seq, qui mesure l'expression génique cellule par cellule et permet d'explorer l'hétérogénéité cellulaire.

La mesure de ces expressions repose sur plusieurs étapes. D'abord, l'ARNm est extrait de l'échantillon étudié. Les ARN extraits sont ensuite rétrotranscrits en ADN complémentaire (ADNc) via une enzyme appelée transcriptase inverse. Nous obtenons ainsi une collection d'ADNc représentant les ARN initiaux, qui sont ensuite amplifiés, afin de constituer ce qu'on appelle une librairie. Cette librairie est ensuite séquencée, permettant d'obtenir des séquences textuelles de nucléotides de ces ADN, aussi appelées *reads* ou *lectures*. Les séquences obtenues sont ensuite alignées sur un génome de référence à l'aide d'algorithmes bioinformatiques tels que Salmon (Patro *et al.*, 2017), de manière à identifier l'origine de chaque séquence, c'est-à-dire de quel gène il provient. On réalise alors le comptage des fragments : le nombre de séquences alignées sur chaque gène est dénombré, fournissant une mesure quantitative de l'expression de chaque gène dans l'échantillon (Mortazavi *et al.*, 2008).

Les données RNA-seq se présentent sous forme de comptages entiers : chaque valeur correspond au nombre de séquences alignées sur un gène donné présent dans l'échantillon. Cependant, ces comptages sont affectés par différents biais techniques, tels que la profondeur de séquençage, la composition des échantillons, etc. (Robinson *et al.*, 2010; Love *et al.*, 2014). Une étape de normalisation est donc essentielle avant toute analyse statistique afin de rendre les échantillons comparables entre eux.

En immunologie et en vaccinologie, les profils d'expression génique permettent d'étudier l'activité des cellules immunitaires en réponse à une infection ou une vaccination (Rinchai *et al.*, 2022). Le RNA-seq constitue ainsi un outil d'investigation précieux pour mieux comprendre les

mécanismes biologiques sous-jacents à la réponse immunitaire. Dans cette thèse, ces données transcriptomiques sont intégrées dans des modèles mécanistes afin d'étudier les dynamiques cellulaires ou d'expliquer la variabilité inter-individuelle observée.

1.2.3 Jeux de données utilisés

Dans cette thèse, nous analysons trois jeux de données principaux issus d'essais vaccinaux : (i) l'essai PREVAC-UP sur le vaccin contre le virus Ebola (Badio *et al.*, 2021), (ii) un essai sur le vaccin ZOSTAVAX contre le virus varicelle-zona dans le cadre du programme HIPC (*Human Immunology Project Consortium*) (Li *et al.*, 2017), et (iii) l'étude COVAX sur la vaccination contre le virus SARS-CoV-2 par vaccin ARNm (Pfizer-BioNTech) (Rinchai *et al.*, 2022).

1.2.3.1 Essai PREVAC-UP : vaccination contre Ebola

Les projets PREVAC (*Partnership for Research on Ebola Vaccination*) et PREVAC-UP ont permis un vaste essai clinique de phase II mené en Afrique de l'Ouest pour évaluer la sécurité et l'immunogénicité de différentes stratégies vaccinales contre la maladie à virus Ebola (Badio *et al.*, 2021). L'essai a recruté un peu plus de 2800 participants, adultes et enfants de plus de 1 an, répartis en plusieurs groupes : (1) un groupe recevant un schéma prime-boost Ad26.ZEBOV suivi 56 jours plus tard d'une dose de MVA-BN-Filo, (2) un groupe recevant le vaccin recombinant à virus VSV (rVSV-ZEBOV, Merck) en une dose unique, (3) un autre recevant le même vaccin rVSV-ZEBOV administré en deux doses espacées de 56 jours. Un groupe placebo était également inclus pour comparaison. Une sous-étude immunologique a été menée sur un sous-échantillon de participants adultes afin d'analyser en détail la réponse immunitaire induite. Pour comparer ces stratégies, un suivi des patients a été mis en place depuis la première vaccination et jusqu'à un an après celle-ci, avec des visites rapprochées durant les premières semaines puis des visites à moyen et long terme. Parmi l'ensemble des paramètres collectés, nous trouvons notamment des titres d'IgG dirigées contre la glycoprotéine d'Ebola, des IgG anti-EBOV-GP, quantifiés par FANG ELISA (*i.e.* un protocole ELISA harmonisé et validé inter-laboratoires pour assurer la comparabilité des mesures entre sites) le jour de la première injection, puis environ une, deux, quatre, huit et neuf semaines plus tard, et enfin à trois, six et douze mois. Les profils d'expression génique générés chez les adultes inclus dans la sous-étude immunologique ont également été mesurés lors de la visite programmée au jour 56, avant l'éventuel rappel, trois heures après cette visite, 7 jours post-vaccination, puis aux mois 3, 6 et 12. D'autres mesures cliniques et biologiques sont également réalisées au fil du suivi (bilans biochimiques, température, taille, poids, etc.) (Badio *et al.*, 2021).

Dans le cadre de cette thèse, nous analysons la réponse humorale mesurée à compter d'une semaine après la visite de rappel, c'est-à-dire à partir du jour 63, sept jours après l'éventuelle seconde injection selon la stratégie vaccinale ; l'analyse se restreint ici au bras rVSV-ZEBOV non boosté parmi les participants de la sous-étude immunologique. Nous intégrons à l'analyse de ces données l'expression génique de ces individus mesurée au jour 56, avant l'éventuel rappel.

L'objectif est d'évaluer dans quelle mesure les profils d'expression génique précédant la visite de rappel sont des prédicteurs de la réponse immunitaire humorale observée ensuite. Les données mesurées utilisées dans le cadre de cette analyse sont représentées dans la Figure 1.6.

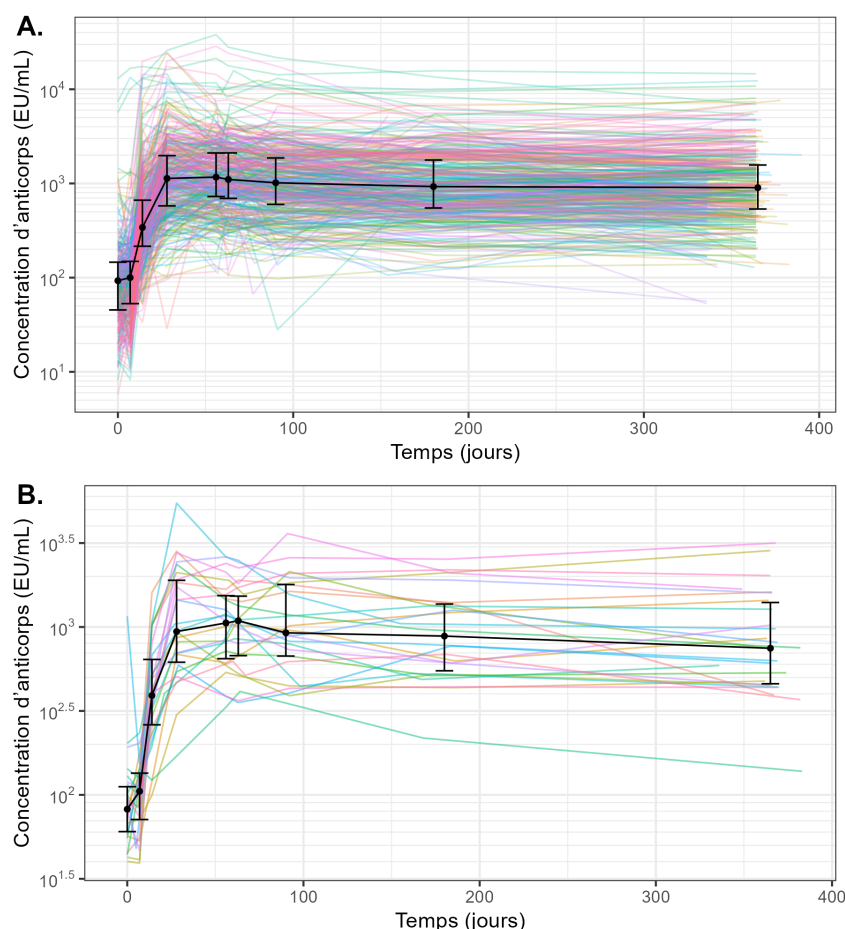


FIGURE 1.6 – **Cinétique des IgG anti-EBOV-GP (EU/mL)** mesurées dans l'essai PREVAC. (A) Participants ayant reçu rVSV-ZEBOV non boosté (n=395) ; (B) Sous-groupe de participants de la sous-étude immunologique ayant reçu rVSV-ZEBOV non boosté (n=26). Les courbes noires indiquent la médiane et l'intervalle interquartile à chaque date de visite.

1.2.3.2 Essai sur le vaccin ZOSTAVAX contre le zona

Un second jeu de données provient d'un essai clinique portant sur le vaccin ZOSTAVAX (Li *et al.*, 2017), un vaccin vivant atténué contre le virus varicelle-zona (VZV), utilisé pour prévenir le zona. L'étude inclut 77 volontaires adultes en bonne santé répartis en deux groupes d'âge (33 sujets jeunes, de moins de 40 ans ; 44 sujets âgés, de plus de 60 ans). Les participants ont été vaccinés par ZOSTAVAX et suivis ensuite les premiers jours après l'injection, puis à moyen et plus long terme. Parmi l'ensemble des données collectées lors de cette étude, nous trouvons, pour un sous-ensemble de 37 participants (16 jeunes, 21 âgés), des titres d'IgG dirigées contre la glycoprotéine du VZV, des IgG anti-VZV-GP, quantifiées par ELISA le jour de la vaccination, environ une et deux semaines plus tard, puis autour d'un, trois et six mois. Nous trouvons également des profils d'expression génique, mesurés le jour de la vaccination, le lendemain, puis

aux troisième et septième jours. D'autres marqueurs immunologiques et cliniques ont également été recueillis au fil du suivi, notamment des cytokines mesurées par Luminex ainsi qu'une quantification de l'ADN du virus varicelle-zona dans le sang (Li *et al.*, 2017). Les données publiques correspondantes sont accessibles sur ImmPort (accession **SDY984**) (Bhattacharya *et al.*, 2018a).

Dans le cadre de cette thèse, nous analysons les titres d'anticorps chez les 35 participants référencés sur ImmPort (SDY984) disposant de profils d'expression génique, et nous les relierons aux profils mesurés le lendemain de la vaccination. L'objectif est d'évaluer si ces signatures très précoces prédisent la dynamique de la réponse immunitaire humorale au cours du suivi à partir de la vaccination. La Figure 1.7 illustre la cinétique des IgG anti-VZV pour ce sous-ensemble.

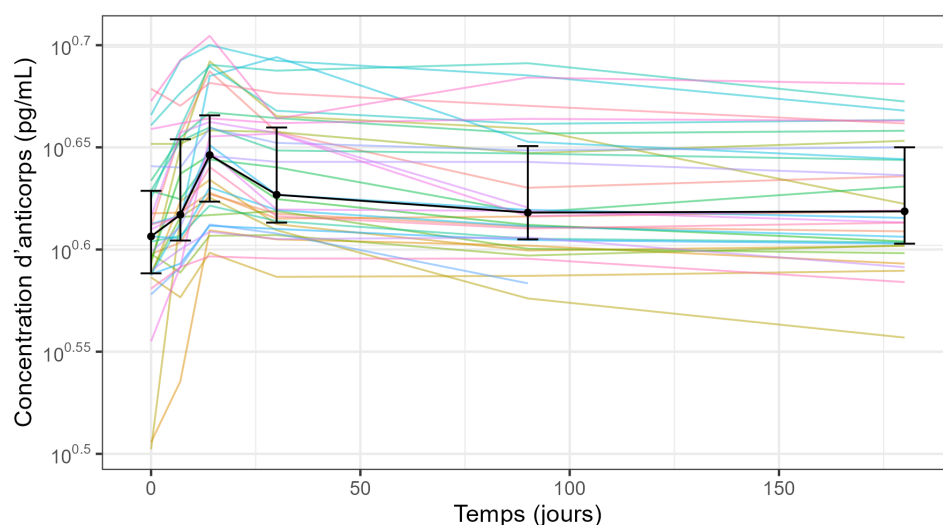


FIGURE 1.7 – **Cinétique des IgG anti-VZV (pg/mL)** mesurées chez les participants de l'étude sur le vaccin ZOSTAVAX disponible sur la plateforme ImmPort (n=35). Les courbes noires montrent la médiane et l'intervalle interquartile de la concentration d'anticorps à chaque visite.

1.2.3.3 Essai sur le vaccin ARNm contre le SARS-CoV-2

L'étude COVAX correspond à une étude de suivi longitudinal de la réponse à un vaccin à ARNm contre le SARS-CoV-2 (Rinchai *et al.*, 2022). Dans cette étude, 23 volontaires ont reçu deux doses de vaccin, espacées de 21 jours, soit Pfizer-BioNTech pour 19 individus, soit Moderna pour 4 individus. Cette étude inclut 17 participants n'ayant jamais été infectés par le virus (15 ayant reçu le vaccin Pfizer-BioNTech et 2 le vaccin Moderna) et 6 ayant déjà été infectés (4 pour le vaccin Pfizer-BioNTech et 2 pour Moderna). Les participants ont été suivis selon un schéma incluant un prélèvement la veille de chaque injection, puis des auto-prélèvements capillaires quotidiens par piqûre au doigt durant les neuf jours qui suivent, ainsi qu'un contrôle autour de deux semaines. Dans cette étude, des profils d'expression génique mesurés par séquençage d'ARN ont été collectés la veille des injections, pendant les neuf jours suivants, puis deux semaines après. En parallèle, des dosages sérologiques multiplex par Luminex des IgM et des IgG dirigées

contre plusieurs antigènes viraux ont été réalisés aux visites la veille des injections, puis une et deux semaines après.

Dans le cadre de cette thèse, nous étudions, pour le sous-ensemble de participants n'ayant jamais été infectés par le virus et ayant reçu le vaccin Pfizer-BioNTech, le lien entre les trajectoires des profils d'expression génique au cours des jours suivant la première dose et les composantes cellulaires de la réponse immunitaire humorale observées ensuite. La Figure 1.8 illustre la dynamique des anticorps chez ces 15 participants.

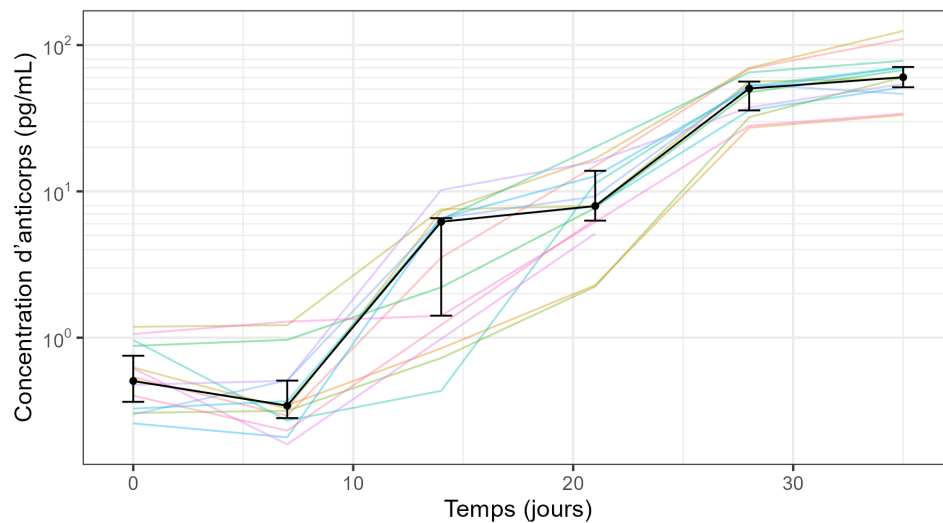


FIGURE 1.8 – **Cinétique des IgG dirigées contre le SARS-CoV-2 (pg/mL)** chez les participants de l'étude ayant reçu le vaccin Pfizer-BioNTech ($n = 15$).
Les lignes noires indiquent la médiane et l'intervalle interquartile à chaque visite.

– Chapitre 2 –

Modèles non-linéaires à effets mixtes et Méthodes de régularisation

Table des matières

2.1	Modèle statistique et estimation	38
2.1.1	Modélisation mécaniste et modèles non-linéaires à effets mixtes	38
2.1.1.1	Modèles linéaires à effets mixtes	38
2.1.1.2	Modèles non-linéaires à effets mixtes	39
2.1.1.3	Modèles mécanistes	40
2.1.2	Méthodes d'estimation dans les modèles non-linéaires à effets mixtes .	42
2.1.2.1	Identifiabilité	43
2.1.2.2	Vraisemblance du modèle	43
2.1.2.3	Approximations de la vraisemblance	44
2.1.2.4	Optimisation de la vraisemblance	47
2.1.2.5	Algorithmes EM et SAEM	48
2.1.2.6	Monolix et l'implémentation de l'algorithme SAEM	50
2.2	Sélection de modèles et régularisation dans les modèles NLEM	51
2.2.1	Critères d'information pour la sélection de modèles	51
2.2.1.1	Critères AIC et BIC	52
2.2.1.2	Critères corrigés : le BICc	53
2.2.2	Pénalisation	53
2.2.2.1	Pénalisation Lasso	55
2.2.2.2	Stratégies de sélection et de calibration	57
2.2.2.3	Algorithme de sélection stable	58
2.2.3	Pénalisation Lasso dans les modèles non-linéaires à effets mixtes	60
2.3	Utilisation des modèles mécanistes pour l'immunologie	62
2.3.1	Modélisation du système immunitaire par Bonin <i>et al.</i>	63
2.3.2	Modélisation de la réponse immunitaire humorale par Dari <i>et al.</i>	66
2.3.3	Modélisation de la réponse humorale post-vaccinale par Balelli <i>et al.</i> et Clairon <i>et al.</i>	68
2.3.4	Modèle à deux compartiments de plasmocytes par Andraud <i>et al.</i>	71
2.3.5	Discussion sur l'identifiabilité et la complexité des modèles	72

Dans ce chapitre, nous présentons le cadre méthodologique utilisé tout au long de cette thèse pour modéliser et analyser des données longitudinales en immunologie, en particulier dans le contexte vaccinal. Ces données, souvent bruitées, de faible dimension par individu et hétérogènes d'un patient à l'autre, requièrent des approches flexibles, capables de rendre compte à la fois des mécanismes biologiques sous-jacents et de la variabilité individuelle. Les modèles NLEM, couplant système d'EDO et variabilité aléatoire, offrent un cadre robuste pour modéliser de telles dynamiques complexes. Nous introduisons d'abord les principes généraux des modèles à effets mixtes, depuis leur version linéaire classique jusqu'aux extensions non-linéaires, et détaillons les méthodes d'estimation associées, dont en particulier l'algorithme SAEM (*Stochastic Approximation Expectation-Maximization*). Nous abordons ensuite les enjeux liés à la sélection de variables dans ces modèles, en particulier dans des contextes de grande dimension, et présentons les méthodes de pénalisation adaptées comme le Lasso et ses variantes. Enfin, nous illustrons l'utilisation des modèles NLEM dans un cadre mécaniste pour modéliser la réponse immunitaire post-vaccinale. Nous passons en revue plusieurs modèles publiés décrivant la cinétique des cellules immunitaires et des anticorps après vaccination. Cette diversité de modèles sert de base à la construction du modèle utilisé dans nos applications, présentées dans les chapitres suivants.

2.1 Modèle statistique et estimation

2.1.1 Modélisation mécaniste et modèles non-linéaires à effets mixtes

En immunologie et dans les essais cliniques, les données collectées sont longitudinales : pour chaque individu suivi au cours de l'étude, des mesures de marqueurs biologiques sont réalisées plusieurs fois au cours du temps. Ces données longitudinales nécessitent des méthodes statistiques adaptées pour les analyser avec précision. Analyser chaque individu séparément en ajustant un modèle distinct à ses données a été reconnu dès les années 1970 comme une approche inefficace pour estimer des paramètres de population (Sheiner *et al.*, 1972, 1977). En particulier, Sheiner *et al.* ont souligné qu'un ajustement patient par patient ne permet pas de distinguer la variabilité inter-individuelle, entre les patients, de la variabilité intra-individuelle, au sein d'un même patient, et oblige à collecter de nombreuses mesures par individu pour obtenir des estimations précises. Ces limitations ont motivé le développement d'approches mixtes qui permettent de modéliser conjointement l'ensemble des trajectoires individuelles dans un échantillon de sujets. Les modèles NLEM répondent à ce besoin en offrant un cadre performant pour l'analyse simultanée de données longitudinales de multiples individus.

2.1.1.1 Modèles linéaires à effets mixtes

Les modèles à effets mixtes ont initialement été formulés dans un cadre linéaire, introduit par Harville (1977) et formalisé par Laird et Ware (1982), pour prendre en compte les variabilités

individuelles. Pour un individu $i = 1, \dots, N$, à la $j^{\text{ème}}$ observation, $j = 1, \dots, n_i$, le modèle linéaire à effets mixtes s'écrit :

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{Z}_{ij}^T \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_{ij},$$

où $\mathbf{Y}_{ij}$ est la réponse observée, éventuellement multivariée pour modéliser conjointement plusieurs marqueurs, $\mathbf{X}_{ij} \in \mathbb{R}^p$ est le vecteur de covariables associé aux effets fixes $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)^T$, égaux pour tous les individus, et $\mathbf{Z}_{ij} \in \mathbb{R}^q$ est le vecteur de covariables associé aux effets aléatoires $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}, \dots, \boldsymbol{\gamma}_{iq})^T$, différents pour chaque individu. Dans le cas d'une réponse observée multivariée, $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}_i$ sont des matrices dont les colonnes correspondent aux différents marqueurs observés et les lignes correspondent aux covariables. Par convention, nous supposons $\mathbb{E}[\boldsymbol{\gamma}_i] = \mathbf{0}$, de sorte que l'effet moyen est capturé par le coefficient $\boldsymbol{\beta}$. Les effets aléatoires $\boldsymbol{\gamma}_i$ sont spécifiques à chaque individu et modélisent les déviations par rapport à la moyenne de la population. Ils sont supposés indépendants entre les individus et identiquement distribués selon une loi normale $\mathcal{N}(0, \boldsymbol{\Omega})$, avec $\boldsymbol{\Omega}$ une matrice de variance-covariance qui capture la variabilité inter-individuelle. L'erreur résiduelle $\boldsymbol{\varepsilon}_{ij}$ représente les bruits de mesure et la variabilité intra-individuelle non expliquée par le modèle ; elle est supposée généralement indépendante et identiquement distribuée (iid) selon une loi normale $\mathcal{N}(0, \boldsymbol{\Sigma})$ et indépendante des effets aléatoires.

Sous forme matricielle, pour chaque individu i , nous avons finalement :

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i,$$

avec $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \dots, \mathbf{Y}_{in_i})^T$ la matrice des observations de l'individu $i = 1, \dots, N$, $\mathbf{X}_i = (\mathbf{X}_{ij})_{j \leq n_i}$ la matrice $n_i \times p$ des covariables individuelles associées au vecteur des effets fixes $\boldsymbol{\beta}$, $\mathbf{Z}_i = (\mathbf{Z}_{ij})_{j \leq n_i}$ la matrice $n_i \times q$ des covariables associées aux effets aléatoires $\boldsymbol{\gamma}_i$ et $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{ij})_{j \leq n_i}$ la matrice des erreurs. Les paramètres à estimer sont finalement $\boldsymbol{\theta} = (\boldsymbol{\Omega}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$.

2.1.1.2 Modèles non-linéaires à effets mixtes

Le cadre linéaire mixte a ensuite été généralisé au cas non-linéaire par [Lindstrom et Bates \(1990\)](#). Dans de nombreux domaines (pharmacocinétique, croissance biologique, neurosciences, etc.), la relation entre la réponse et les paramètres n'est pas linéaire et nécessite une modélisation plus souple. Le modèle NLEM conserve la distinction entre effets fixes et aléatoires, mais permet une dépendance non-linéaire des observations aux paramètres. Formellement, on écrit :

$$\mathbf{Y}_{ij} = \mathbf{f}(\boldsymbol{\psi}_i, t_{ij}) + \boldsymbol{\varepsilon}_{ij}, \quad (2.1)$$

où $\mathbf{f}(\cdot)$ est une fonction non-linéaire décrivant l'évolution temporelle de la dynamique d'intérêt, éventuellement multivariée, pour un individu i donné, t_{ij} est le temps de la mesure, et $\boldsymbol{\psi}_i$ est le vecteur de paramètres individuels pour l'individu i , incluant à la fois l'effet fixe et aléatoire. Ces paramètres $\boldsymbol{\psi}_i$ ne sont pas observés directement et varient d'un individu à l'autre, reflétant la variabilité biologique entre sujets. Le terme d'erreur $\boldsymbol{\varepsilon}_{ij}$ représente toujours les bruits ou erreurs de mesure des différentes composantes du vecteur des observations $\mathbf{Y}_{ij}$ et est généralement supposé indépendant et identiquement distribué selon la loi normale $\mathcal{N}(0, \boldsymbol{\Sigma})$, indépendant des $\boldsymbol{\psi}_i$.

La formulation NLEM permet donc de modéliser des relations potentiellement complexes (par exemple modèles de croissance, exponentielles, etc.) entre la variable d'intérêt et les paramètres à l'aide de la fonction f . Les paramètres individuels ψ_i sont usuellement modélisés via un modèle linéaire, ou via un modèle linéaire généralisé avec une transformation h fixée, de la forme :

$$h(\psi_i) = h(\psi_0) + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i, \quad (2.2)$$

où ψ_0 est la valeur typique du vecteur de paramètres, $\mathbf{X}_i$ le vecteur des covariables de l'individu i , $\boldsymbol{\beta} \in \mathbb{R}^p$ le vecteur d'effets fixes associés aux covariables, et $\boldsymbol{\eta}_i$ sont les effets aléatoires de l'individu i , supposés indépendants et identiquement distribués selon la loi normale $\mathcal{N}(0, \boldsymbol{\Omega})$, où la matrice $\boldsymbol{\Omega}$ possède éventuellement des termes diagonaux nuls pour définir un paramètre sans variabilité inter-individuelle. Dans certains cas, pour garantir des contraintes de positivité ou de bornes sur les paramètres, on applique une transformation h aux paramètres ; par exemple, si un paramètre doit rester positif (taux, concentration initiale, etc.), on pourra utiliser $h(\cdot) = \log(\cdot)$. Les paramètres à estimer de ce modèle sont finalement $\boldsymbol{\theta} = (\psi_0, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$.

Les modèles NLEM se sont imposés comme un outil standard pour analyser des données longitudinales non-linéaires. Ils sont largement utilisés dans des domaines variés tels que la pharmacologie (modélisation pharmacocinétique/pharmacodynamique) (Ibarra *et al.*, 2021), l'écologie (croissance d'animaux ou de végétaux) (Baey *et al.*, 2018), la sociologie (O'Brien *et al.*, 2008), etc. L'approche NLEM, en empruntant de l'information à travers la population, permet d'obtenir des estimations robustes même lorsque chaque patient n'a que peu de mesures individuelles.

2.1.1.3 Modèles mécanistes

Dans le cadre de cette thèse, on s'intéresse particulièrement aux modèles NLEM dits mécanistes, c'est-à-dire fondés sur la description de processus sous-jacents à l'aide d'équations différentielles. Un modèle mécaniste est un modèle qui incorpore explicitement les mécanismes ou lois physiques, biologiques, gouvernant le phénomène étudié. Autrement dit, c'est un modèle basé sur les processus connus du système. Par exemple, en pharmacologie, on modélisera la concentration d'un médicament par des équations différentielles traduisant l'absorption, la distribution et l'élimination du médicament dans l'organisme (Ibarra *et al.*, 2021). De même, en épidémiologie, on utilisera des équations différentielles pour représenter les transitions d'individus entre états susceptible, infecté et rétabli (modèles SIR et dérivés) (Ganser *et al.*, 2024). Un atout majeur des modèles mécanistes est leur interprétabilité : les paramètres ψ_i ont souvent une signification biologique précise (taux de dissipation, capacité maximale, constante de réaction, etc.), et les équations décrivent la dynamique causale des composantes du système. Ces modèles s'appuient sur des connaissances préalables, les connaissances *a priori*, ou des résultats expérimentaux par exemple, pour formuler mathématiquement les relations de cause à effet dans le système.

Dans la pratique, les modèles mécanistes en biostatistique prennent souvent la forme de systèmes d'EDO. Soit $\mathbf{S}(t) = (S_1(t), \dots, S_K(t))$ le vecteur représentant l'état du système à

l'instant t , dont les composantes $S_k(t)$, $k = 1, \dots, K$, correspondent aux différentes variables ou compartiments du modèle à l'instant t , comme par exemple les concentrations de différentes populations de cellules, virus, médicaments, etc., à cet instant. Un modèle mécaniste défini par des EDOs s'écrit alors sous la forme :

$$\frac{dS_k}{dt}(t) = u_k(\mathbf{S}(t), \boldsymbol{\psi}_i, t), \quad k = 1, \dots, K,$$

où chaque fonction u_k décrit la dynamique de la $k^{\text{ème}}$ composante en fonction de l'état global $\mathbf{S}(t)$ et des paramètres individuels $\boldsymbol{\psi}_i$ de l'individu considéré. Une condition initiale $S_{k0}(\boldsymbol{\psi}_i)$ pour l'individu i complète la spécification du modèle dynamique. La solution de ce système d'EDO s fournit la trajectoire $\mathbf{S}_i(t)$ pour un individu i et des paramètres $\boldsymbol{\psi}_i$ donnés. Dans de nombreux cas complexes, il n'existe pas de solution analytique simple pour $\mathbf{S}(t)$, et il faut recourir à des méthodes de résolution numérique pour simuler les trajectoires (Curtiss et Hirschfelder, 1952). De nouveau, on peut définir les paramètres individuels $\boldsymbol{\psi}_i$ par le modèle (2.2).

Au cours de cette thèse, nous présentons les modèles NLEM en les décomposant en trois parties distinctes : **le modèle structurel**, **le modèle statistique** et **le modèle d'observation**.

Le **modèle structurel** correspond à la description formelle du processus sous-jacent, en l'absence d'erreur de mesure. Il s'agit du modèle mécaniste déterministe décrivant la dynamique intra-individuelle du phénomène d'intérêt. Dans notre contexte, il s'agit généralement d'un système d'EDO tel que présenté plus haut. Pour un individu i , de paramètres $\boldsymbol{\psi}_i$, on note $\mathbf{S}_i(t) = (S_{i1}(t), \dots, S_{iK}(t))$ la solution du système :

$$\begin{cases} \frac{dS_{ik}}{dt}(t) &= u_k(\mathbf{S}_i(t), \boldsymbol{\psi}_i, t), \quad k = 1, \dots, K, \\ S_{ik}(0) &= S_{k0}(\boldsymbol{\psi}_i). \end{cases} \quad (2.3)$$

Le modèle structurel fournit ainsi, pour un individu donné, une prédiction théorique de l'évolution de ses variables d'intérêt $\mathbf{S}_i(t)$ au cours du temps, en l'absence de bruit.

Le **modèle statistique** introduit la variabilité entre individus et l'incertitude sur les paramètres du modèle structurel. Autrement dit, il permet de décrire comment les paramètres $\boldsymbol{\psi}_i$ varient d'un individu à l'autre au sein de la population, et comment ils se lient à des covariables observées éventuellement. Dans un modèle NLEM, on suppose que chaque vecteur de paramètres individuels $\boldsymbol{\psi}_i$ est une réalisation aléatoire d'une certaine distribution de population. On modélise usuellement $\boldsymbol{\psi}_i$ comme distribué suivant une loi normale centrée autour d'une valeur typique ou moyenne. Formellement, on peut écrire pour chaque individu, à une transformation h près, le modèle linéaire généralisé (2.2).

Le **modèle d'observation**, ou d'erreur, relie le modèle structurel aux données réellement observées $\mathbf{Y}_{ij}$. Cette composante est indispensable pour prendre en compte les erreurs de mesure inhérente aux données, et plus généralement la différence entre la réalité simulée par le modèle et les observations empiriques. Elle permet en outre de considérer dans la modélisation des données de nature variée, comme des données de comptage ou des données censurées à gauches (en présence de limite de détection).

En pratique, même si le modèle structurel décrivait parfaitement la dynamique intra-individuelle moyenne, les données mesurées Y_{ij} s'écartent de la trajectoire idéale prédite par le modèle, à cause du bruit expérimental et de facteurs non modélisés. De plus, toutes les composantes du modèle $S_i(t)$ ne sont généralement pas observées : seules certaines variables peuvent parfois être mesurées, comme par exemple les concentrations d'anticorps dans le cadre de la réponse immunitaire humorale, alors que d'autres compartiments du système, comme dans notre exemple les populations de cellules sécrétrices d'anticorps, ne le sont pas directement.

On formalise généralement le modèle d'observation par une équation du type :

$$Y_{ij} = f(S_i(t_{ij}), \psi_i) + \varepsilon_{ij},$$

où $f(\cdot)$ est une fonction qui sélectionne ou transforme certaines composantes de l'état latent $S_i(t_{ij})$, éventuellement en fonction des paramètres ψ_i . Le terme ε_{ij} représente l'erreur résiduelle, supposée en général indépendante et identiquement distribuée suivant $\mathcal{N}(0, \Sigma)$.

Il est également possible de considérer des modèles d'erreur plus complexes, dans lesquels l'écart-type dépend lui-même de l'état latent ou des paramètres. On écrit alors :

$$Y_{ij} = f(S_i(t_{ij}), \psi_i) + f_{\text{err}}(S_i(t_{ij}), \psi_i, \xi) \varepsilon_{ij}, \quad (2.4)$$

où $\varepsilon_{ij} \sim \mathcal{N}(0, I)$, ξ correspond aux paramètres du modèle d'erreur et $f_{\text{err}}(\cdot)$ décrit la structure de variance-covariance (par exemple constante, proportionnelle, ou combinée). Ainsi, dans un modèle d'erreur constant, on a $f_{\text{err}} = \Sigma$; dans un modèle proportionnel, $f_{\text{err}}(\cdot) = a f(S_i(t), \psi_i)$; et dans un modèle combiné, $f_{\text{err}}(\cdot) = a f(S_i(t), \psi_i) + b$. Pour chacun de ces cas, le vecteur ξ correspond alors respectivement aux paramètres Σ , a et (a, b) .

Les paramètres à estimer de ce modèle mécaniste sont finalement $\theta = (\psi_0, \beta, \Omega, \xi)$.

2.1.2 Méthodes d'estimation dans les modèles non-linéaires à effets mixtes

Dans cette section, nous présentons les approches numériques et algorithmiques utilisées pour estimer les paramètres des modèles NLEM. L'objectif est d'obtenir, à partir des données longitudinales observées, des estimateurs des paramètres de population θ , mais aussi des paramètres individuels. L'estimation par maximum de vraisemblance, rappelée ci-dessous, constitue une approche classique. Toutefois, dans le cas des modèles NLEM, la présence d'effets aléatoires latents et la structure non-linéaire du modèle, en particulier lorsque celui-ci repose sur des équations différentielles, font que la vraisemblance marginale ne peut généralement pas être exprimée sous forme analytique. Son optimisation directe devient alors difficile, ce qui impose de recourir à des méthodes numériques ou approchées pour procéder à l'estimation.

2.1.2.1 Identifiabilité

Une condition préalable essentielle à la bonne estimation des paramètres θ est l'identifiabilité du modèle. Cela désigne la possibilité de retrouver de manière unique les paramètres à partir des données disponibles. Si cette condition n'est pas remplie, plusieurs vecteurs de paramètres différents peuvent expliquer les mêmes observations, ce qui rend l'inférence impossible ou ambiguë.

On distingue généralement deux niveaux d'identifiabilité : l'identifiabilité théorique ou structurelle et l'identifiabilité pratique. L'**identifiabilité structurelle** est une propriété théorique qui dépend uniquement de la formulation mathématique du modèle. Elle correspond à la capacité à déterminer les paramètres de manière unique dans un cadre idéal où l'on aurait accès aux trajectoires exactes du système au cours du temps, sans bruit ni incertitude de mesure et en tout temps. Cette analyse peut être conduite par des outils d'algèbre différentielle, comme DAISY (Bellu *et al.*, 2007), ou par des approches intégrant l'étude des dérivées de paramètres (Dong *et al.*, 2023). L'**identifiabilité pratique**, quant à elle, concerne la situation réelle où les données sont limitées et bruitées. Même lorsqu'un modèle est structurellement identifiable, il peut devenir difficile, voire impossible, d'estimer correctement certains paramètres si les observations ne sont pas suffisantes ou suffisamment informatives. On évalue alors cela au moyen d'analyses de sensibilité, de simulations de type Monte Carlo ou encore l'analyse du profil de vraisemblance (Lavielle et Aarons, 2016; Raue *et al.*, 2009).

Dans le cadre de cette thèse, cette problématique est particulièrement cruciale. Les modèles mécanistes basés sur des équations différentielles incluent souvent des compartiments ou des paramètres non observés directement, comme certaines sous-populations cellulaires. L'absence d'identifiabilité complète impose alors de fixer certains paramètres, de simplifier la structure du modèle ou de combiner plusieurs sources de données afin d'améliorer l'information disponible.

2.1.2.2 Vraisemblance du modèle

Dans les modèles NLEM, les effets aléatoires individuels η_i sont des variables latentes. La vraisemblance marginale des données s'écrit alors comme un produit d'intégrales selon la distribution de ces effets aléatoires. Pour $i = 1, \dots, N$, en notant $\mathbf{Y}_i = (Y_{ij})_{j \leq n_i}$ le vecteur des observations de l'individu i , on a :

$$L(\theta) = p(\mathbf{Y} | \theta) = \prod_{i=1}^N p(\mathbf{Y}_i | \theta) = \prod_{i=1}^N \int p(\mathbf{Y}_i | \eta_i, \theta) p(\eta_i | \theta) d\eta_i, \quad (2.5)$$

où θ regroupe tous les paramètres du modèle, $p(\mathbf{Y}_i | \theta)$ correspond à la distribution marginale des observations pour l'individu i et $p(\mathbf{Y}_i | \eta_i, \theta)$ correspond à la distribution conditionnelle des observations sachant les effets aléatoires η_i pour l'individu i . En général, cette intégrale n'admet pas de solution analytique, et l'optimisation de la vraisemblance observée devient alors difficile et contraint d'employer des méthodes approchées ou numériques pour calculer ou maximiser $L(\theta)$.

À noter que, dans le cas particulier des modèles linéaires à effets mixtes, pour une réponse

univariée $\mathbf{Y}_i \in \mathbb{R}^{n_i}$, la distribution marginale du vecteur des observations de l'individu i , est alors une gaussienne $\mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\boldsymbol{\Omega}\mathbf{Z}_i^\top + \sigma^2\mathbf{I}_{n_i})$, où $\boldsymbol{\gamma}_i \sim \mathcal{N}(0, \boldsymbol{\Omega})$ et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Dès lors, la vraisemblance s'obtient analytiquement en utilisant cette loi gaussienne, permettant directement son optimisation.

2.1.2.3 Approximations de la vraisemblance

L'évaluation exacte de la vraisemblance marginale observée dans les modèles NLEM est en général impossible, en particulier lorsqu'ils reposent sur une modélisation mécaniste définie par des EDOs. De fait, diverses méthodes ont été proposées pour approcher cette intégrale, et ainsi permettre l'application d'algorithmes d'optimisation. Dans la suite, nous supposons que, pour tout vecteur de paramètres individuels $\boldsymbol{\psi}_i$ et pour tout temps t , la fonction de prédiction $\mathbf{f}(\boldsymbol{\psi}_i, t)$ peut être évaluée numériquement, même en l'absence de solution analytique du système différentiel sous-jacent. Cette hypothèse repose sur l'utilisation de solveurs numériques adaptés aux systèmes d'EDOs, même dits raides, c'est-à-dire dont la sensibilité aux paramètres fait que certains évoluent beaucoup plus rapidement que d'autres, tels que les méthodes implicites de type *Backward Differentiation Formulas* (BDF) ou de Runge-Kutta implicites (Curtiss et Hirschfelder, 1952). Ces méthodes permettent une approximation stable et précise des trajectoires modélisées, en particulier dans les modèles mécanistes complexes rencontrés en pharmacométrie ou en immunologie.

Une première classe d'approches repose sur la linéarisation locale du modèle. La méthode FOA (*First Order Approximation*) (Sheiner et Beal, 1980a; Beal et Sheiner, 1982) consiste à approcher à l'ordre 1 la fonction de réponse $\mathbf{f}(\cdot)$ autour de la moyenne des effets aléatoires, notée $\bar{\boldsymbol{\psi}}$, typiquement $\bar{\boldsymbol{\psi}} = \mathbb{E}[\boldsymbol{\psi}_i] = \mathbb{E}[h^{-1}(h(\boldsymbol{\psi}_0) + \mathbf{X}_i\boldsymbol{\beta})]$ d'après (2.2). Cela revient alors à écrire :

$$\mathbf{f}(\boldsymbol{\psi}_i, t) \approx \mathbf{f}(\hat{\boldsymbol{\psi}}_i, t) + \mathbf{Z}_i(\bar{\boldsymbol{\psi}}) \boldsymbol{\eta}_i,$$

où $\boldsymbol{\eta}_i = h(\boldsymbol{\psi}_i) - h(\bar{\boldsymbol{\psi}})$ et $\mathbf{Z}_i$ est la jacobienne de $\mathbf{f}$ évaluée en $\bar{\boldsymbol{\psi}}$. Cette approximation ramène le modèle à un modèle linéaire mixte gaussien, pour lequel la vraisemblance marginale admet une forme explicite. Toutefois, cette approximation peut introduire des biais significatifs, notamment lorsque la non-linéarité est forte ou que la variance des effets aléatoires est élevée (Davidian et Gallant, 1992).

La méthode FOCE (*First Order Conditional Estimation*) (Lindstrom et Bates, 1990) améliore la méthode FOA en réalisant la linéarisation autour des estimateurs *a posteriori* individuels $\hat{\boldsymbol{\psi}}_i$, calculés par maximisation de la densité $p(\boldsymbol{\psi}_i | \mathbf{Y}_i, \boldsymbol{\theta})$ *a posteriori* des effets aléatoires individuels $\boldsymbol{\psi}_i$ conditionnellement aux données observées $\mathbf{Y}_i$ et aux paramètres de population $\boldsymbol{\theta}$. L'approximation devient alors :

$$\mathbf{f}(\boldsymbol{\psi}_i, t) \approx \mathbf{f}(\hat{\boldsymbol{\psi}}_i, t) + \mathbf{Z}_i(\hat{\boldsymbol{\psi}}_i) (\boldsymbol{\psi}_i - \hat{\boldsymbol{\psi}}_i),$$

ce qui améliore la précision tout en augmentant le coût de calcul. Ces méthodes, bien que rapides, présentent des limites pratiques : elles ne convergent pas vers le véritable maximum de

vraisemblance et peuvent biaiser l'estimation des variances, en particulier dans les situations de forte variabilité inter-individuelle (Comets *et al.*, 2017).

Une alternative robuste aux méthodes linéaires consiste à utiliser des techniques de Monte Carlo pour approcher la vraisemblance marginale (Friel et Pettitt, 2008), pour un individu i :

$$p(Y_i | \theta) = \int p(Y_i | \psi_i, \theta) p(\psi_i | \theta) d\psi_i = \mathbb{E}_{p(\psi_i | \theta)} [p(Y_i | \psi_i, \theta)].$$

On peut ainsi estimer $p(Y_i | \theta)$ par la moyenne empirique de la densité $p(Y_i | \psi_i, \theta)$, calculable analytiquement, pour des réalisations indépendantes $(\psi_i^{(m)})_{1 \leq m \leq M}$ échantillonnées selon $p(\psi_i | \theta)$:

$$\hat{p}_{i,M} = \frac{1}{M} \sum_{m=1}^M p(Y_i | \psi_i^{(m)}, \theta).$$

Cet estimateur présente de bonnes propriétés : il est non biaisé et converge au taux $O(1/\sqrt{M})$.

Il est possible d'améliorer cette estimation par un échantillonnage préférentiel, en utilisant une distribution de proposition $\tilde{p}(\psi_i | \theta)$ supposée absolument continue par rapport à $p(\psi_i | \theta)$ (Kloek et van Dijk, 1978; Robert et Casella, 2004). On peut alors écrire :

$$p(Y_i | \theta) = \int p(Y_i | \psi_i, \theta) \frac{p(\psi_i | \theta)}{\tilde{p}(\psi_i | \theta)} \tilde{p}(\psi_i | \theta) d\psi_i = \mathbb{E}_{\tilde{p}(\psi_i | \theta)} \left[p(Y_i | \psi_i, \theta) \frac{p(\psi_i | \theta)}{\tilde{p}(\psi_i | \theta)} \right].$$

Ce qui permet de construire l'estimateur :

$$\hat{p}_{i,M} = \frac{1}{M} \sum_{m=1}^M p(Y_i | \psi_i^{(m)}, \theta) \frac{p(\psi_i^{(m)} | \theta)}{\tilde{p}(\psi_i^{(m)} | \theta)},$$

où $(\psi_i^{(m)})_{1 \leq m \leq M}$ sont indépendamment échantillonnés selon $\tilde{p}(\psi_i | \theta)$. Selon la distribution de proposition utilisée, cette approche peut conduire à un meilleur taux de convergence. La solution optimale serait d'utiliser comme loi de proposition la distribution conditionnelle $p(\psi_i | Y_i, \theta)$, ce qui est en pratique impossible directement.

Une approche classique pour alors approcher cette loi cible repose sur les méthodes de Monte Carlo par chaîne de Markov (MCMC), qui consistent à construire une procédure stochastique qui, à l'état stationnaire, permet d'échantillonner à partir de la distribution de probabilité qui nous intéresse, ici $p(\psi_i | Y_i, \theta)$. Cela permet de générer des échantillons représentatifs de cette loi, même lorsqu'il est impossible de les simuler directement.

Parmi ces méthodes, l'algorithme de Metropolis-Hastings est l'un des plus utilisés (Hastings, 1970). Cette méthode utilise la formule :

$$p(\psi_i | Y_i, \theta) = \frac{p(Y_i | \psi_i, \theta) p(\psi_i | \theta)}{p(Y_i | \theta)},$$

dans laquelle les quantités $p(Y_i | \psi_i, \theta)$ et $p(\psi_i | \theta)$ peuvent être calculées analytiquement, tandis que $p(Y_i | \theta)$ est une constante inobservable. L'algorithme de Metropolis-Hastings permet alors

d'échantillonner suivant $p(\boldsymbol{\psi}_i | \mathbf{Y}_i, \boldsymbol{\theta})$ en procédant de la manière suivante : à chaque itération k , on propose une nouvelle valeur $\boldsymbol{\psi}'_i$ selon une loi de proposition $q(\cdot | \boldsymbol{\psi}_i^{(k)})$, où $\boldsymbol{\psi}_i^{(k)}$ est la valeur des effets aléatoires individuels à l'itération k , puis on l'accepte avec une probabilité

$$\alpha = \min \left(1, \frac{p(\mathbf{Y}_i | \boldsymbol{\psi}'_i, \boldsymbol{\theta}) p(\boldsymbol{\psi}'_i | \boldsymbol{\theta}) q(\boldsymbol{\psi}_i^{(k)} | \boldsymbol{\psi}'_i)}{p(\mathbf{Y}_i | \boldsymbol{\psi}_i^{(k)}, \boldsymbol{\theta}) p(\boldsymbol{\psi}_i^{(k)} | \boldsymbol{\theta}) q(\boldsymbol{\psi}'_i | \boldsymbol{\psi}_i^{(k)})} \right).$$

Après un certain nombre d'itérations, la chaîne atteint un régime stationnaire, et les valeurs acceptées suivent la loi cible $p(\boldsymbol{\psi}_i | \mathbf{Y}_i, \boldsymbol{\theta})$, permettant ainsi d'échantillonner efficacement selon cette distribution. Il est alors également possible d'estimer les paramètres individuels $(\boldsymbol{\psi}_i)_{i \leq N}$.

Enfin, une dernière stratégie consiste à recourir à l'intégration numérique via des quadratures de Gauss, comme celle de Gauss-Hermite (Davidian et Gallant, 1993). Celle-ci approche l'intégrale comme une somme pondérée :

$$\int_{\mathbb{R}^q} g(\boldsymbol{\eta}) d\boldsymbol{\eta} \approx \sum_{k=1}^K w_k g(\boldsymbol{\eta}_k),$$

où les noeuds $\boldsymbol{\eta}_k$ sont choisis selon les racines des polynômes d'Hermite, et w_k sont les poids associés (Golub et Welsch, 1969). Cette méthode est adaptée pour un petit nombre d'effets aléatoires, mais devient rapidement impraticable lorsque ce nombre augmente en raison de la croissance exponentielle du nombre de points par dimension construits.

Des variantes dites adaptatives de la quadrature gaussienne ont été proposées afin d'améliorer l'efficacité (Pinheiro et Bates, 1995). Contrairement à la quadrature standard, qui place les points d'intégration autour de zéro, l'approche adaptative déplace ces points pour les centrer autour de la région où la fonction à intégrer est la plus concentrée, autrement dit autour de la masse de la densité des effets aléatoires.

Dans le cadre de notre vraisemblance intégrée, la distribution *a posteriori* des effets aléatoires $\boldsymbol{\mu}_i$ pour chaque individu i , obtenue par exemple via des méthodes MCMC présentées précédemment, est utilisée pour guider le placement des points de quadrature. Ceux-ci sont recentrés autour des estimateurs *a posteriori* individuels $\hat{\boldsymbol{\psi}}_i$, avec une dispersion ajustée en fonction de l'écart-type de la distribution. Cette adaptation permet de concentrer les points dans la région de plus forte densité, ce qui améliore l'efficacité numérique de l'intégration tout en réduisant le nombre de points nécessaires.

En résumé, les méthodes par linéarisation, quadrature ou Monte Carlo visent toutes à fournir une approximation exploitable de la vraisemblance marginale observée, nécessaire à l'optimisation par algorithme du type Newton-Raphson, présenté dans la Section 2.1.2.4 suivante. Le choix de la méthode dépend du compromis souhaité entre précision et coût computationnel.

2.1.2.4 Optimisation de la vraisemblance

L'estimation du maximum de vraisemblance dans les modèles NLEM s'appuie classiquement sur un algorithme itératif de Newton-Raphson (Lindstrom et Bates, 1988; Verbeke et Cools, 1995). Cet algorithme vise à maximiser une fonction deux fois continûment différentiable en utilisant ses dérivées premières, via son gradient, et secondes, via sa Hessienne. Il est ainsi utilisé pour maximiser la log-vraisemblance observée $LL(\theta)$ définie en (2.5), et donc résoudre :

$$\hat{\theta} = \arg \min_{\theta} -LL(\theta).$$

À chaque itération k , la valeur du paramètre $\theta^{(k)}$ est mise à jour selon la règle :

$$\theta^{(k+1)} = \theta^{(k)} - \mathbf{H}^{-1}(\theta^{(k)}) \mathbf{J}(\theta^{(k)}),$$

où $\mathbf{J}(\theta) = (\partial_{\theta_p} LL(\theta))_{p \leq P}$ est le gradient de la log-vraisemblance observée, et $\mathbf{H}(\theta) = (\partial_{\theta_p, \theta_q}^2 LL(\theta))_{p, q \leq P}$ sa matrice hessienne, avec P la dimension du vecteur de paramètres θ .

Pour assurer la convergence de l'algorithme, il est nécessaire que la hessienne $\mathbf{H}$ soit définie positive. Or, cette condition peut manquer lorsque le modèle est mal spécifié ou que l'approximation de la vraisemblance est instable. Des variantes robustes ont donc été proposées, comme la méthode de Marquardt (Marquardt, 1963; Philipps *et al.*, 2021), dans laquelle la hessienne est régularisée par un terme proportionnel à la matrice identité :

$$\mathbf{G}(\theta^{(k)}) = \mathbf{H}(\theta^{(k)}) + \eta_k \mathbf{I}_P, \quad (2.6)$$

où $\eta_k > 0$ est un coefficient d'amortissement ajusté à chaque itération (Philipps *et al.*, 2021). Cette stratégie permet de garantir que la matrice utilisée soit définie positive à l'étape de mise à jour. L'itération devient alors :

$$\theta^{(k+1)} = \theta^{(k)} - \lambda_k \mathbf{G}^{-1}(\theta^{(k)}) \mathbf{J}(\theta^{(k)}), \quad (2.7)$$

avec λ_k un taux d'apprentissage assurant une décroissance suffisante de la fonction objectif, c'est-à-dire de la log-vraisemblance.

Comme expliqué précédemment, dans les modèles NLEM, et en particulier dans ceux basés sur des EDOs, la log-vraisemblance observée est rarement disponible sous forme analytique. Par conséquent, les dérivées nécessaires à l'algorithme de Newton-Raphson doivent souvent être approximées numériquement, en utilisant par exemple les approximations obtenues par l'une des méthodes présentées dans la Section 2.1.2.3 afin de les dériver numériquement (Burden et Faires, 2000). Il est également possible de relier les quantités $\mathbf{J}$ et $\mathbf{H}$ à la log-vraisemblance complète

$LL_c(\boldsymbol{\theta}) = \log p(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta})$ à l'aide des formules de Fisher et de Louis (Fisher, 1922; Louis, 1982) :

$$\begin{aligned} \mathbf{J}(\boldsymbol{\theta}) &= \mathbb{E} [\partial_{\boldsymbol{\theta}} LL_c(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}], \\ \mathbf{H}(\boldsymbol{\theta}) &= \mathbb{E} [\partial_{\boldsymbol{\theta}}^2 LL_c(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}] + \text{Var} (\partial_{\boldsymbol{\theta}} LL_c(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}), \end{aligned}$$

permettant ainsi une approximation par des schémas numériques (McCulloch, 1997; Gu et Li, 1998).

2.1.2.5 Algorithmes EM et SAEM

L'algorithme EM (*Expectation-Maximization*) a été introduit par Dempster *et al.* (1977) pour l'estimation des paramètres $\boldsymbol{\theta}$ maximisant la vraisemblance $p(\mathbf{Y} \mid \boldsymbol{\theta})$ dans des modèles à variables latentes, comme dans le cadre des modèles à effets mixtes, où les paramètres individuels $\boldsymbol{\psi}_i$ sont des variables latentes non observées. L'algorithme EM procède de manière itérative selon deux étapes. A l'itération k et étant donné $\boldsymbol{\theta}^{(k-1)}$ l'estimation précédente des paramètres :

- **E-step (*Expectation*)** : calcul de l'espérance conditionnelle de la log-vraisemblance complète, définie par

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k-1)}) = \mathbb{E}_{\boldsymbol{\psi} \mid \mathbf{Y}, \boldsymbol{\theta}^{(k-1)}} [\log p(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta})].$$

- **M-step (*Maximization*)** : mise à jour du paramètre par maximisation :

$$\boldsymbol{\theta}^{(k)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k-1)}).$$

Dans les modèles non-linéaires à effets mixtes, la densité conditionnelle $p(\boldsymbol{\psi} \mid \mathbf{Y}, \boldsymbol{\theta})$ n'est généralement pas disponible sous forme analytique, et l'étape E ne peut pas être calculée analytiquement. Pour pallier cette difficulté, Delyon *et al.* (1999) ont proposé une version stochastique de l'algorithme EM, appelée SAEM (*Stochastic Approximation EM*), qui repose sur une simulation de l'étape E suivie d'une mise à jour stochastique de l'espérance Q . L'algorithme SAEM procède selon trois étapes à chaque itération k :

- **S-step (*Simulation*)** : génération d'un échantillon $(\boldsymbol{\psi}^{(k,l)})_{l \leq m_k}$ suivant la loi conditionnelle $p(\boldsymbol{\psi} \mid \mathbf{Y}, \boldsymbol{\theta}^{(k-1)})$.
- **SA-step (*Stochastic Approximation*)** : mise à jour de l'espérance via la formule :

$$Q_k(\boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k-1)}) = Q_{k-1}(\boldsymbol{\theta}) + \gamma_k \left[\frac{1}{m_k} \sum_{l=1}^{m_k} \log p(\mathbf{Y}, \boldsymbol{\psi}^{(k,l)} \mid \boldsymbol{\theta}) - Q_{k-1}(\boldsymbol{\theta}) \right],$$

où γ_k est une suite décroissante de pas d'apprentissage.

- **M-step (*Maximization*)** : mise à jour des paramètres :

$$\boldsymbol{\theta}^{(k)} = \arg \max_{\boldsymbol{\theta}} Q_k(\boldsymbol{\theta}).$$

À l'itération k , étant donné l'estimateur précédemment obtenu $\boldsymbol{\theta}^{(k-1)}$, la log-vraisemblance complète $\log p(\mathbf{Y}, \boldsymbol{\psi} \mid \boldsymbol{\theta}^{(k-1)})$ peut être calculée sous forme explicite. De plus, en considérant un modèle d'erreur constant, le modèle complet appartient à la famille exponentielle. Cela permet de ramener la mise à jour stochastique de la fonction Q à la mise à jour de trois statistiques suffisantes plus simples à calculer (Lavielle et Mentré, 2007) :

$$\mathbb{E}[\boldsymbol{\psi}_i \mid \mathbf{Y}, \boldsymbol{\theta}^{(k-1)}], \quad \mathbb{E}\left[\sum_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T \mid \mathbf{Y}, \boldsymbol{\theta}^{(k-1)}\right], \quad \text{et} \quad \mathbb{E}\left[\sum_{i,j} \left(Y_{ij} - f\left(S_i(t_{ij}), \tilde{\boldsymbol{\psi}}_i^{(k-1)}\right)\right)^2 \mid \mathbf{Y}, \boldsymbol{\theta}^{(k-1)}\right],$$

où $\tilde{\boldsymbol{\psi}}_i^{(k-1)}$ désigne les paramètres individuels prédits par $\boldsymbol{\theta}^{(k-1)}$ en l'absence d'effets aléatoires.

Les approximations stochastiques de ces quantités, notées respectivement $\mathbf{T}_{1,i}^{(k)}$, $\mathbf{T}_2^{(k)}$ et $\mathbf{T}_3^{(k)}$, sont mises à jour selon :

$$\begin{aligned} \mathbf{T}_{1,i}^{(k)} &= \mathbf{T}_{1,i}^{(k-1)} + \gamma_k \left(\frac{1}{m_k} \sum_{l=1}^{m_k} \boldsymbol{\psi}_i^{(k,l)} - \mathbf{T}_{1,i}^{(k-1)} \right), \\ \mathbf{T}_2^{(k)} &= \mathbf{T}_2^{(k-1)} + \gamma_k \left(\frac{1}{m_k} \sum_{l=1}^{m_k} \sum_{i=1}^N \boldsymbol{\psi}_i^{(k,l)} \boldsymbol{\psi}_i^{(k,l)T} - \mathbf{T}_2^{(k-1)} \right), \\ \mathbf{T}_3^{(k)} &= \mathbf{T}_3^{(k-1)} + \gamma_k \left(\frac{1}{m_k} \sum_{l=1}^{m_k} \sum_{i=1}^N \sum_{j=1}^{n_i} \left[Y_{ij} - f\left(S_i(t_{ij}), \tilde{\boldsymbol{\psi}}_i^{(k-1)}\right) \right]^2 - \mathbf{T}_3^{(k-1)} \right). \end{aligned}$$

À partir de ces approximations, la valeur des nouveaux paramètres $\boldsymbol{\theta}^{(k)}$ obtenue lors de l'étape de maximisation peut alors être calculée par des formules explicites, dans le cas d'un modèle d'erreur constant de variance σ^2 :

$$\begin{aligned} \boldsymbol{\mu}^{(k)} &= (\boldsymbol{\psi}_0^{(k)}, \boldsymbol{\beta}^{(k)}) = \left(\sum_{i=1}^N \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i \right)^{-1} \sum_{i=1}^N \tilde{\mathbf{X}}_i^T \mathbf{T}_{1,i}^{(k)}, \\ \boldsymbol{\Omega}^{(k+1)} &= \text{diag} \left(\frac{1}{N} \left[\mathbf{T}_2^{(k)} - \sum_{i=1}^N (\tilde{\mathbf{X}}_i \boldsymbol{\mu}^{(k)}) \mathbf{T}_{1,i}^{(k)T} - \sum_{i=1}^N \mathbf{T}_{1,i}^{(k)} (\tilde{\mathbf{X}}_i \boldsymbol{\mu}^{(k)})^T + \sum_{i=1}^N (\tilde{\mathbf{X}}_i \boldsymbol{\mu}^{(k)}) (\tilde{\mathbf{X}}_i \boldsymbol{\mu}^{(k)})^T \right] \right), \\ \sigma^{(k)} &= \sqrt{\frac{\mathbf{T}_3^{(k)}}{\sum_{i=1}^N n_i}}. \end{aligned}$$

où $\tilde{\mathbf{X}}_i$ correspond à la matrice des covariables individuelles et de l'intercept $\tilde{\mathbf{X}}_i = (\mathbf{1}_N, \mathbf{X}_i)$.

En revanche, lorsque le modèle n'appartient pas à la famille exponentielle, comme c'est le cas par exemple en présence d'un modèle d'erreur proportionnelle ou combinée, toutes les statistiques suffisantes ne peuvent pas être mises à jour par approche stochastique. L'étape de maximisation ne peut alors pas être réalisée à l'aide de formules fermées comme ci-dessus, et repose sur des algorithmes numériques d'optimisation, tels que l'algorithme de Newton-Raphson, comme présenté en Section 2.1.2.4. Enfin, Delyon *et al.* (1999) ont démontré la convergence du SAEM vers un maximum, au moins local, de la vraisemblance observée sous des conditions générales.

L'approche a été adaptée par Kuhn et Lavielle (2004), qui ont proposé un couplage SAEM-

MCMC dans lequel les trajectoires des effets aléatoires sont simulées à l'aide de l'algorithme de Metropolis-Hastings.

2.1.2.6 Monolix et l'implémentation de l'algorithme SAEM

Monolix (MODèles NON-Linéaires à effets miXtes) est un logiciel développé par Lixoft spécifiquement conçu pour l'estimation, la simulation et l'analyse de modèles non-linéaires à effets mixtes (Lixoft, 2024b). Il est largement utilisé en pharmacométrie, en immunologie, et plus généralement dans les sciences biomédicales nécessitant la modélisation de données longitudinales complexes. Monolix propose une interface utilisateur ainsi que des connecteurs R (Lixoft, 2024a) permettant d'effectuer toutes les étapes de l'analyse statistique, depuis l'importation des données jusqu'aux diagnostics post-estimation.

L'**algorithme SAEM** est implémenté dans Monolix pour estimer les paramètres de population θ dans les modèles NLEM (Kuhn et Lavielle, 2004). Son implémentation repose sur une exécution en plusieurs phases. D'abord, une *phase de burn-in* initialise la chaîne de Markov utilisée pour simuler les effets aléatoires individuels ψ_i , sans encore rentrer dans l'algorithme SAEM. Cette stabilisation est cruciale pour obtenir un échantillonnage fiable lors de l'étape S de l'algorithme dans la phase suivante. Une fois la chaîne de Markov initialisée, l'algorithme SAEM commence réellement avec d'abord une *phase d'exploration*, où l'espérance stochastique de la log-vraisemblance complète est mise à jour dans l'étape SA avec un pas constant $\gamma_k = 1$. Une fois les critères de convergence atteints, Monolix entre dans une *phase de stabilisation*, où le pas utilisé dans l'étape SA décroît progressivement. Cela permet de lisser les estimations et de stabiliser la suite $(\theta^{(k)})$ autour d'un maximum local. L'échantillonnage des effets aléatoires lors de l'étape S est assuré par un algorithme Metropolis-Hastings adaptatif, similaire à celui introduit précédemment, basé sur des lois de proposition ajustées en cours d'itération. L'ensemble des étapes est géré de manière automatique et robuste, permettant l'estimation efficace de ces modèles complexes. Le critère d'arrêt du SAEM implémenté dans Monolix est basé sur la stabilisation des paramètres de population θ estimés au cours des itérations.

D'autres méthodes statistiques sont disponibles sur Monolix pour l'analyse des modèles NLEM. D'abord, l'**estimation des paramètres individuels** après convergence est réalisée par échantillonnage par MCMC conditionnellement aux paramètres $\hat{\theta}$ estimés. Les réalisations obtenues dans la dernière phase de l'algorithme SAEM sont ainsi réutilisées pour approcher la distribution $p(\psi_i | Y_i, \hat{\theta})$. Il est également possible d'estimer la **matrice d'information de Fisher** afin d'évaluer la précision des estimateurs. Elle est obtenue soit via l'approximation de Louis (Louis, 1982), soit par des méthodes numériques. Elle permet notamment d'obtenir des intervalles de confiance pour les paramètres estimés et d'évaluer leur corrélation *a posteriori*. Il est également possible d'obtenir des **prédictions individuelles** par l'estimateur MAP (*Maximum a posteriori*), qui consiste à calculer, pour chaque individu i , la valeur $\hat{\psi}_i$ qui maximise la densité conditionnelle $p(\psi_i | Y_i, \hat{\theta})$. Ces prédictions permettent de générer des trajectoires individuelles et d'évaluer la qualité de l'ajustement au niveau du sujet. Enfin, les **critères d'information**, définis dans la Section 2.2.1, sont également calculés à partir de l'estimation de

la log-vraisemblance marginale. Leur rôle pour la sélection de modèles est également discuté en Section 2.2.1.

Monolix propose de nombreuses sorties et diagnostics automatiques, incluant les résidus conditionnels, les prédictions marginales ou conditionnelles, les distributions empiriques des effets aléatoires, ainsi que des outils graphiques pour évaluer la qualité de l’ajustement ou détecter d’éventuelles inadéquations du modèle, tels que les VPC (*Visual Predictive Checks*), qui comparent les trajectoires prédites aux distributions empiriques, ou les graphiques d’ajustement individuels. Il est également possible de générer des jeux de données simulés à partir de paramètres donnés, afin d’explorer le comportement du modèle dans divers scénarios (Lixoft, 2024c).

2.2 Sélection de modèles et régularisation dans les modèles NLEM

Dans les applications modernes (génomique, imagerie, épidémiologie, etc.), le nombre de variables explicatives p peut être très grand par rapport au nombre d’individus N . Typiquement, dans les applications en vaccinologie avec des données d’expression génique, des milliers de gènes sont mesurés pour seulement une vingtaine d’individus (Rinchai *et al.*, 2022; Badio *et al.*, 2021; Li *et al.*, 2017). Or, les outils de statistique traditionnels sont basés sur des hypothèses avec p petit et N grand, où les résultats asymptotiques (pour $N \rightarrow \infty$ à p fixé) sont fiables (Giraud, 2021). En grande dimension, ce cadre classique s’effondre : les estimateurs deviennent instables et le risque de surajustement (ou *overfitting*) explose. Il est donc essentiel de restreindre la complexité des modèles pour obtenir des estimations fiables. D’une part, on cherche à sélectionner un sous-ensemble de variables pertinentes, afin d’obtenir un modèle parcimonieux, plutôt que d’estimer tous les paramètres. D’autre part, on introduit des méthodes de régularisation, ou pénalisation, qui limitent la taille effective des paramètres. Ces approches permettent un compromis entre qualité d’ajustement, ou vraisemblance, et complexité du modèle, souvent formalisé par des critères pénalisés. Une perspective classique est de considérer des critères d’information, comme l’AIC, le BIC, ou des méthodes pénalisées, comme le *Ridge* et le Lasso. Ces principes guident la sélection de modèles en grande dimension pour les modèles NLEM.

2.2.1 Critères d’information pour la sélection de modèles

Historiquement, le premier critère de sélection de modèles fondé sur une pénalisation explicite a été proposé par Mallows (1973) dans le cadre du modèle linéaire. Ce critère, appelé C_p , vise à améliorer l’estimation de l’erreur de prédiction en corrigeant l’erreur d’ajustement observée, ou risque quadratique empirique, par un terme de biais associé à la complexité du modèle. Ce biais, parfois désigné comme *optimisme de l’estimateur*, peut être interprété comme une approximation de la composante de variance dans la décomposition biais-variance du risque prédictif.

Dans le cas de la régression linéaire avec une fonction de perte quadratique, c’est-à-dire

lorsque l'on cherche à minimiser l'erreur quadratique moyenne entre les prédictions $\hat{y}$ et les observations y , le critère de Mallows s'écrit :

$$C_p = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \frac{d}{n} \hat{\sigma}^2,$$

où d désigne le nombre de paramètres estimés, n le nombre d'observations, et $\hat{\sigma}^2$ une estimation de la variance de l'erreur dans le modèle saturé, c'est-à-dire le modèle intégrant toutes les variables explicatives. Ce terme correctif $2 \frac{d}{n} \hat{\sigma}^2$ pénalise la complexité du modèle pour éviter le surajustement.

De manière plus générale, un critère d'information est une règle de sélection de modèles qui cherche à équilibrer la qualité d'ajustement aux données et la complexité du modèle, en vue d'optimiser la capacité prédictive. Ces critères, tels que l'*Akaike Information Criterion* (AIC) ou le *Bayesian Information Criterion* (BIC), permettent de comparer plusieurs modèles candidats, le meilleur étant celui qui minimise la valeur du critère. Bien qu'ils reposent sur des fondements théoriques distincts, ils partagent l'idée commune d'ajouter à la log-vraisemblance un terme pénalisant la complexité du modèle.

2.2.1.1 Critères AIC et BIC

Les critères d'information AIC et BIC sont largement utilisés pour la sélection de modèles en statistique.

Tout d'abord, l'*Akaike Information Criterion* (AIC), introduit par Akaike (1973), se définit ainsi :

$$\text{AIC} = -2LL(\hat{\theta}) + 2d,$$

où $LL(\hat{\theta})$ est la log-vraisemblance évaluée au maximum obtenu, et d est le nombre de paramètres estimés dans le modèle, typiquement la dimension de θ . Ce critère s'applique au modèle estimé par maximum de vraisemblance et suppose que la densité réelle des observations Y appartient à la famille de densités considérée. L'AIC vise à sélectionner le modèle selon un critère de qualité basé sur la divergence de Kullback-Leibler avec la réalité, favorisant souvent la prédiction, mais il a tendance à surajuster en présence de nombreuses variables explicatives.

Le *Bayesian Information Criterion* (BIC) de Schwarz (1978) est quant à lui défini par :

$$\text{BIC} = -2LL(\hat{\theta}) + d \log(N),$$

introduisant une pénalité plus lourde que l'AIC pour des échantillons de grande taille, puisqu'elle croît avec $\log(N)$. Il consiste à chercher, asymptotiquement, le modèle associé à la plus grande probabilité *a posteriori*. Le BIC a l'avantage d'être consistant en termes de sélection de modèles, c'est-à-dire qu'il sélectionne avec probabilité tendant vers 1 le bon modèle si celui-ci figure parmi les candidats. Cependant, en grande dimension, il peut aussi être sujet au surajustement. En

effet, lorsque le nombre de variables explicatives p croît avec la taille de l'échantillon N , par exemple $p = N^\alpha$, $\alpha \geq 1/2$, la taille attendue du modèle sélectionné par BIC peut diverger, ce qui invalide l'usage classique du critère dans ce contexte (Giraud, 2021).

2.2.1.2 Critères corrigés : le BICc

Les critères AIC et BIC présentent ainsi des biais notables lorsqu'ils sont appliqués sur de petits échantillons ou dans des contextes de grande dimension, comme c'est souvent le cas dans les applications mettant en jeu des données transcriptomiques. Pour pallier ces limites, le BICc (*corrected BIC*) a été proposé par Delattre *et al.* (2014) et est particulièrement adapté pour les modèles non-linéaires à effets mixtes. Il constitue ainsi un critère central pour la sélection de modèles dans le cadre de cette thèse.

Le BICc ajuste le terme de pénalisation du BIC afin de mieux refléter la complexité des modèles mixtes et corriger les biais du BIC cités précédemment. Il s'écrit comme :

$$\text{BICc} = -2LL(\hat{\theta}) + \log(n) \dim(\theta_F) + \log(N) \dim(\theta_R),$$

où d'une part, $\dim(\theta_F)$ correspond au nombre de paramètres dits fixes dans θ , comme les valeurs typiques des paramètres ψ_0 , les coefficients β des paramètres sans effets aléatoires ou les paramètres du modèle d'erreur ξ ; d'autre part, $\dim(\theta_R)$ correspond au nombre de paramètres dits aléatoires dans θ , comme les variances des effets aléatoires Ω ou les coefficients β des paramètres individuels avec effets aléatoires.

2.2.2 Pénalisation

Une stratégie classique pour limiter la complexité d'un modèle et prévenir le surajustement consiste à modifier le critère d'optimisation en y ajoutant un terme de pénalisation dépendant des paramètres. Ces approches, dites pénalisées, consistent à résoudre un problème de la forme :

$$\hat{\theta} = \arg \min_{\theta \in \Theta} C(\theta) + \text{pen}(\theta, \Lambda),$$

où $C(\theta)$ est une fonction de coût mesurant l'adéquation du modèle aux données, par exemple la perte quadratique, l'opposé de log-vraisemblance ou la déviance, et $\text{pen}(\theta, \Lambda)$ est un terme de régularisation contraignant la structure du modèle, généralement par une norme. Le vecteur Λ regroupe les paramètres de régularisation ou de pénalité, modulant l'intensité de la pénalisation.

Dans le cas particulier de l'estimation par maximum de vraisemblance, le critère $C(\theta)$ est pris comme l'opposée de la log-vraisemblance, de sorte que le problème d'optimisation revient à maximiser la vraisemblance pénalisée :

$$LL_{\text{pen}}(\theta) = LL(\theta) - \text{pen}(\theta, \Lambda).$$

Ces méthodes sont particulièrement utiles pour la sélection de covariables, notamment dans les modèles NLEM, où les paramètres individuels sont reliés à des covariables via un modèle linéaire généralisé (2.2). On note $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ la matrice des p covariables, avec $\mathbf{X}_k = (X_{ik})_{i \leq N}$, et $\boldsymbol{\beta} = (\beta_k)_{k \leq p}$ le vecteur des coefficients associés. L'objectif est d'identifier le support :

$$S = \{ k \leq p \mid \beta_k \neq 0 \},$$

c'est-à-dire les covariables expliquant significativement la variabilité inter-individuelle. On cherche ainsi à estimer un vecteur $\boldsymbol{\beta}$ parcimonieux tout en contrôlant le risque de sélection incorrecte.

La qualité d'une méthode de sélection peut être évaluée à l'aide de critères dérivés des notions classiques d'erreurs de type I (faux positifs (FP), correspondant aux covariables sélectionnées à tort) et de type II (faux négatifs (FN), correspondant à celles non sélectionnées à tort). Ces erreurs se traduisent, dans le cadre de la sélection de variables, par les indicateurs suivants :

- le *False Positive Rate* (FPR), mesurant la proportion de covariables non pertinentes incorrectement sélectionnées :

$$FPR = \frac{FP}{FP + TN},$$

- le *False Negative Rate* (FNR), correspondant à la proportion de covariables pertinentes omises :

$$FNR = \frac{FN}{TP + FN},$$

où TN (vrais négatifs) désigne les covariables justement non sélectionnées, et TP (vrais positifs) celles justement retenues. Le FPR est directement relié à la probabilité de commettre une erreur de type I, tandis que le FNR correspond à une erreur de type II. Ces indicateurs sont donc particulièrement informatifs pour quantifier les erreurs de sélection et guider le choix d'une méthode selon les priorités souhaitées de l'analyse.

D'autres mesures synthétiques sont également utilisées :

- la *False Discovery Proportion* (FDP), définie comme la proportion d'erreurs parmi les covariables sélectionnées :

$$FDP = \frac{FP}{FP + TP},$$

- le *False Discovery Rate* (FDR), qui correspond à l'espérance du FDP :

$$FDR = \mathbb{E}[FDP],$$

- et enfin, le score F_1 , qui combine précision et rappel dans une mesure harmonique :

$$F_1 = \frac{2TP}{2TP + FP + FN}.$$

Le FDR est particulièrement utilisé dans les contextes de grande dimension, où l'on cherche à contrôler le taux d'erreurs parmi les variables déclarées significatives, par exemple en trans-

criptomique. Le score F_1 , quant à lui, fournit une mesure unique de la performance globale de la sélection, équilibrant les contributions du rappel, la capacité à ne pas manquer de vraies variables explicatives, et de la précision, la fiabilité des sélections effectuées.

Il est important de souligner qu'il n'existe en général aucune garantie qu'une méthode de sélection atteigne simultanément un FDR et un FNR nuls et un score F_1 maximal de 1. Le contrôle de ces indicateurs est délicat et impose souvent des compromis : réduire le FDR entraîne généralement une hausse du FNR, et inversement. Dans la pratique, la stratégie adoptée dépend des objectifs de l'analyse et de la tolérance relative aux faux positifs ou aux faux négatifs (Benjamini et Hochberg, 1995; Genovese et Wasserman, 2002; Bogdan *et al.*, 2015).

Dans la suite, on se place dans un cadre simplifié de régression linéaire, où les observations $\mathbf{Y} \in \mathbb{R}^N$ suivent le modèle :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

avec $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ un bruit gaussien centré. Ce modèle constitue un cadre de référence pour introduire les méthodes de pénalisation. Il est, par ailleurs, structurellement proche du modèle linéaire utilisé pour modéliser les paramètres individuels dans (2.2).

Traditionnellement, le vecteur $\boldsymbol{\beta}$ est estimé par les moindres carrés :

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

où $\|\cdot\|_2$ désigne la norme euclidienne. Si la matrice des covariables $\mathbf{X}$ est de rang plein, c'est-à-dire de rang égal à p , alors $\mathbf{X}^T \mathbf{X}$ est inversible, et la solution s'écrit :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

ce qui assure unicité et identifiabilité du modèle. En revanche, lorsque $p > n$ ou que les colonnes de $\mathbf{X}$ sont colinéaires, cette condition n'est plus vérifiée, et la solution n'est plus unique. Pour permettre l'estimation, on impose alors des hypothèses structurelles supplémentaires, en particulier que le vecteur $\boldsymbol{\beta}$ soit *sparse*, c'est-à-dire contient de nombreuses composantes nulles. Ces hypothèses justifient le recours à des techniques de régularisation, notamment les pénalisations de type L_1 (Lasso), permettant de retrouver une solution unique dans des contextes surparamétrés.

2.2.2.1 Pénalisation Lasso

Le Lasso (Tibshirani, 1996), *Least Absolute Shrinkage and Selection Operator*, est une méthode de régression pénalisée reposant sur une pénalité L_1 . Il résout le problème d'optimisation suivant :

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

Cette formulation favorise des estimateurs parcimonieux, c'est-à-dire comportant de nombreux coefficients nuls. Ainsi, le Lasso réalise simultanément l'estimation des paramètres et la sélection des variables explicatives. Lorsque $\lambda = 0$, on retrouve l'estimateur des moindres carrés classiques. À l'inverse, lorsque λ tend vers l'infini, tous les coefficients s'annulent. Cette régularisation s'appuie sur une contrainte de type L_1 , qui impose une limite sur la somme des valeurs absolues des coefficients. Géométriquement, cela revient à forcer les solutions à rester à l'intérieur de la boule L_1 . Or, cette forme particulière favorise naturellement des solutions avec des coefficients exactement nuls (Tibshirani, 1996). C'est ce qui permet au Lasso de faire une sélection des covariables : seules celles ayant un effet suffisamment marqué sont conservées dans le modèle.

Cette méthode est particulièrement adaptée aux contextes de grande dimension ($p > n$), bien qu'il n'existe alors pas de solution analytique. Dans ce cas, des méthodes d'optimisation numérique, telles que la descente de coordonnées cycliques, sont employées (Friedman *et al.*, 2010b). Celle-ci repose sur l'idée de mettre à jour séquentiellement chaque coefficient β_k , $k \leq p$, en gardant les autres fixés. Plus précisément, à l'itération t , on fixe les valeurs actuelles des autres coefficients $\boldsymbol{\beta}_{-k}^{(t)} = (\beta_1^{(t)}, \dots, \beta_{k-1}^{(t)}, \beta_{k+1}^{(t-1)}, \dots, \beta_p^{(t-1)})$, et on cherche la valeur de $\beta_k^{(t)}$ qui minimise la fonction objectif conditionnelle :

$$\beta_k^{(t)} = \arg \min_{\beta_k \in \mathbb{R}} \left\{ \|Y - \mathbf{X}_{-k} \boldsymbol{\beta}_{-k}^{(t)} - \mathbf{X}_k \beta_k\|_2^2 + \lambda |\beta_k| \right\}.$$

Cette minimisation unidimensionnelle s'effectue efficacement par une opération de “seuil doux” (*soft-thresholding*). En notant $\mathbf{R}_k = \mathbf{Y} - \mathbf{X}_{-k} \boldsymbol{\beta}_{-k}$ le résidu partiel, c'est-à-dire la réponse ajustée en enlevant la contribution de $\mathbf{X}_k$, on peut exprimer la mise à jour de β_k comme :

$$\beta_k^{(t)} = S \left(\frac{1}{n} \mathbf{X}_k^T \mathbf{R}_k, \frac{\lambda}{2n} \right),$$

où $S(z, \gamma)$ est l'opérateur de *soft-thresholding*, défini par :

$$S(z, \gamma) = \text{sign}(z) \cdot (|z| - \gamma)_+,$$

avec $(\cdot)_+ = \max(0, \cdot)$. Cet opérateur effectue une contraction vers zéro : si $|z| \leq \gamma$, alors le coefficient est mis à zéro. L'algorithme alterne ainsi les mises à jour coordonnées par coordonnées, jusqu'à convergence des coefficients.

D'autres formes de pénalisation ont été proposées. La régression *Ridge*, introduite par Hoerl et Kennard (1970), repose sur une pénalité L_2 :

$$\text{pen}(\boldsymbol{\beta}, \lambda) = \lambda \|\boldsymbol{\beta}\|_2^2 = \lambda \sum_{k=1}^p \beta_k^2.$$

Cette approche conserve toutes les variables mais limite la variance des prédictions en contraignant la norme des coefficients. Elle est donc utile en présence de multicollinéarité. Cependant, sa principale limite réside dans l'absence de sélection de variables, rendant l'interprétation plus difficile.

La méthode *elastic net*, introduite par Zou et Hastie (2005), combine les pénalités L_1 et L_2 du Lasso et du *Ridge* :

$$\text{pen}(\boldsymbol{\beta}, \lambda, \alpha) = \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 \right),$$

avec $\alpha \in [0, 1]$. Pour $\alpha = 1$, on retrouve le Lasso ; pour $\alpha = 0$, la régression *Ridge*. Cette formulation offre un compromis : la sélection de variables via le Lasso, et la stabilisation des coefficients grâce à la pénalité *Ridge*. Elle favorise notamment les effets de groupe : les variables corrélées ont tendance à être sélectionnées ou exclues conjointement.

Plusieurs extensions du Lasso ont été proposées pour des cas particuliers. Le *group Lasso* (Yuan et Lin, 2006) généralise le Lasso classique en appliquant la pénalité sur des groupes de variables prédéfinis. Ainsi, un groupe entier de covariables peut être inclus ou exclu en bloc. Le *fused Lasso* (Tibshirani et al., 2005) ajoute, en plus de la pénalité L_1 habituelle, une pénalité sur les différences entre coefficients successifs. Cette technique est adaptée lorsque les variables peuvent être ordonnées : elle favorise alors des coefficients proches pour des covariables voisines, produisant des solutions par morceaux constantes.

D'autres méthodes reposent sur des pénalités non convexes. Par exemple, la pénalité SCAD (*Smoothly Clipped Absolute Deviation*) (Fan et Li, 2001) est une fonction continue par morceaux qui favorise aussi la parcimonie tout en limitant le biais sur les grands coefficients. De manière analogue, la pénalité MCP (*Minimax Concave Penalty*) (Zhang, 2010) est une autre fonction concave visant à réduire le biais du Lasso tout en réalisant la sélection.

2.2.2.2 Stratégies de sélection et de calibration

Les méthodes pénalisées telles que la régression Lasso ou l'*elastic net* dépendent de paramètres de régularisation $\boldsymbol{\Lambda}$, comme λ ou α , dont la calibration est essentielle pour garantir un bon compromis entre biais, variance et parcimonie du modèle. Plusieurs stratégies existent pour permettre la calibration de ces hyperparamètres.

Une première approche consiste à utiliser des critères d'information tels que l'AIC (Akaike, 1973) ou le BIC (Schwarz, 1978) présentés précédemment. Concrètement, pour chaque valeur candidate de $\boldsymbol{\Lambda}$, le modèle est ajusté. La valeur de l'hyperparamètre retenue est finalement celle qui minimise le critère.

Une autre stratégie courante est la validation croisée (*Cross-Validation*). Elle consiste à diviser l'échantillon en K sous-groupes de taille la plus égale possible, appelés plis, à ajuster le modèle sur $K - 1$ de ces plis, puis à calculer l'erreur de prédiction sur le pli restant. Cette opération est répétée K fois, chaque pli servant successivement de pli de test. L'erreur de validation croisée est obtenue en moyennant les erreurs de prédiction sur les K partitions, et le paramètre $\boldsymbol{\Lambda}^*$ optimal est celui qui minimise cette erreur. Cette méthode est souvent utilisée car elle ne repose sur aucune hypothèse de distribution particulière, et elle vise directement à optimiser la performance prédictive du modèle (Stone, 1974; Hastie et al., 2009). Cependant, la validation croisée présente plusieurs limites. Elle peut induire une sélection de variables instable :

la variabilité induite par la répartition des plis peut conduire à des sélections très différentes d'un échantillonnage à l'autre (Lim et Yu, 2016).

2.2.2.3 Algorithme de sélection stable

Pour répondre à ces problèmes, l'algorithme de sélection stable (*stability selection*) a été introduit par Meinshausen et Bühlmann (2010), afin d'améliorer la robustesse des méthodes de sélection de variables, notamment en contexte de grande dimension. Il repose sur le principe de répéter la procédure de sélection, comme par exemple, une régression Lasso avec pénalisation L_1 , sur un grand nombre de sous-échantillons aléatoires, et de ne retenir que les variables sélectionnées un nombre suffisant de fois. Ce procédé permet de limiter l'instabilité observée dans les méthodes de sélection classiques basées sur un seul échantillon, comme la validation croisée, en introduisant une forme d'agrégation des résultats.

Formellement, pour N la taille de notre échantillon, on génère B sous-échantillons $(I_b)_{b \leq B}$, chacun constitué d'environ $\lfloor N/2 \rfloor$ individus tirés aléatoirement. Sur chaque sous-échantillon I_b , on applique une méthode de sélection donnée, dépendant d'un vecteur d'hyperparamètres Λ supposé ici quelconque. Dans le cas du Lasso par exemple, Λ correspond au paramètre de pénalité λ et dans le cas de l'*elastic net* $\Lambda = (\lambda, \alpha)$. On note ainsi $\hat{S}_b^\Lambda = \{k \leq p \mid \hat{\beta}_k^{\Lambda, b} \neq 0\}$ l'ensemble des variables sélectionnées par la méthode dans le sous-échantillon b .

On calcule ensuite, pour chaque covariable $k \in \{1, \dots, p\}$, sa fréquence de sélection empirique :

$$\hat{\pi}^\Lambda(k) = \frac{1}{B} \sum_{b=1}^B \mathbb{1} \left\{ \hat{\beta}_k^{\Lambda, b} \neq 0 \right\},$$

où $\mathbb{1}\{\cdot\}$ désigne la fonction indicatrice. Cette quantité représente la proportion de fois où la variable k a été sélectionnée parmi les B sous-échantillons, pour une valeur de Λ donnée.

La sélection finale est obtenue en fixant un seuil de stabilité $\pi \in [0, 1]$, et en conservant uniquement les variables dont la fréquence de sélection dépasse ce seuil. L'ensemble sélectionné est alors défini par :

$$\hat{S}^{\Lambda, \pi} = \{k \leq p \mid \hat{\pi}^\Lambda(k) \geq \pi\}.$$

Ainsi, une variable n'est retenue que si elle est sélectionnée dans une proportion suffisamment élevée de sous-échantillons, ce qui traduit une forme de robustesse de sa sélection.

Au-delà de son intérêt empirique, la méthode de sélection stable présente des garanties théoriques, notamment en ce qui concerne le contrôle du nombre de faux positifs. En effet, Meinshausen et Bühlmann (2010) ont démontré que, sous certaines hypothèses, notamment d'échangeabilité des covariables non pertinentes et d'indépendance conditionnelle, le nombre espéré de faux positifs, appelé *Per-Family Error Rate* (PFER), peut être borné explicitement. Si l'on note $q_\Lambda = |\hat{S}_b^\Lambda|$ la taille moyenne des ensembles sélectionnés sur les sous-échantillons, alors pour tout

seuil $\pi > 0.5$, on a la borne suivante :

$$\mathbb{E}[V] \leq \frac{q_{\Lambda}^2}{(2\pi - 1)p},$$

où V désigne le nombre de faux positifs et p le nombre total de covariables considérées. Ce résultat fournit un outil pratique de calibration des hyperparamètres Λ et π : en fixant une valeur cible du PFER, on peut rechercher les couples (Λ, π) vérifiant cette inégalité. Le PFER étant l'espérance du nombre de faux positifs, il est directement lié au FDR, défini comme l'espérance de la proportion de faux positifs parmi les éléments sélectionnés. Ainsi, les bornes obtenues sur le PFER fournissent également une borne sur le FDR présenté précédemment.

Ce cadre a ensuite été étendu par [Shah et Samworth \(2013\)](#), qui proposent une version généralisée de la sélection stable fondée sur des sous-échantillons complémentaires appariés, et relâchent certaines hypothèses, notamment celle d'échangeabilité. Ces travaux fournissent également des bornes améliorées pour le PFER et introduisent des critères de sélection moins conservateurs, tout en conservant la garantie théorique :

$$\mathbb{E}[V] \leq \begin{cases} \frac{1}{2(2\pi - 1 - \frac{1}{B})} \times \frac{q^2}{p} & , \text{ si } \pi \leq 0,75, \\ \frac{4(1 - \pi + \frac{1}{B})}{1 + \frac{2}{B}} \times \frac{q^2}{p} & , \text{ sinon.} \end{cases}$$

Ces bornes théoriques fournissent des outils précieux pour calibrer les hyperparamètres (Λ, π) en imposant une contrainte explicite sur le nombre moyen de fausses sélections. Toutefois, cette calibration repose généralement sur un choix arbitraire d'au moins l'un de ces hyperparamètres.

Pour pallier cette difficulté, [Bodinier et al. \(2023\)](#) ont proposé une méthode de calibration automatique fondée sur un score de stabilité. Leur approche repose sur l'hypothèse selon laquelle, sous un modèle instable, chaque covariable aurait la même probabilité de sélection $\gamma_{\Lambda} = q_{\Lambda}/p$. Sous cette hypothèse nulle, le nombre de sélections d'une variable sur les B sous-échantillons suit une loi binomiale de paramètres (B, γ_{Λ}) . À partir de cette hypothèse, une log-vraisemblance sous l'hypothèse d'instabilité, notée $LL_{\Lambda, \pi}$, est définie. Le score de stabilité correspondant est alors donné par :

$$S_{\Lambda, \pi} = -LL_{\Lambda, \pi}.$$

Plus ce score est élevé, plus la sélection observée est improbable sous l'hypothèse d'instabilité, et donc plus la sélection est considérée comme stable. Enfin, [Bodinier et al. \(2023\)](#) proposent de calibrer automatiquement les hyperparamètres en optimisant ce score de stabilité sous contrainte que le PFER soit inférieur à un seuil fixé *a priori*, contrainte obtenue à l'aide des bornes de [Meinshausen et Bühlmann \(2010\)](#) ou de [Shah et Samworth \(2013\)](#) présentées ci-dessus, permettant ainsi d'obtenir une sélection stable et contrôlée sans choix arbitraire initial.

2.2.3 Pénalisation Lasso dans les modèles non-linéaires à effets mixtes

Dans le contexte de l'inférence statistique et de la sélection de modèles, les méthodes de régularisation, et en particulier le Lasso, se sont imposées comme des outils privilégiés pour la sélection parcimonieuse de covariables dans les modèles à effets mixtes. Plusieurs approches ont été développées récemment pour intégrer la pénalisation Lasso au sein des modèles NLEM, répondant aux enjeux posés par les données de grande dimension ou les petits échantillons.

Ribbing *et al.* (2007) proposent d'abord d'intégrer la sélection Lasso dans les modèles NLEM en estimant le modèle complet, c'est-à-dire incluant toutes les relations covariable-paramètre possibles, sous la contrainte $\sum_k |\beta_k| \leq t$, où t est un paramètre de pénalisation contrôlé par validation croisée. Cette contrainte est implémentée dans le logiciel NONMEM à l'aide d'un facteur de correction sur les paramètres, qui adapte dynamiquement leur amplitude pour respecter la borne t . NONMEM (*NONlinear Mixed-Effect Modeling*) (Sheiner et Beal, 1980b) est un logiciel dédié à l'estimation des modèles NLEM en pharmacométrie, beaucoup utilisé dans l'industrie. L'approche de Ribbing *et al.* (2007) a montré des performances prédictives supérieures à celles de la méthode de sélection pas-à-pas plus traditionnellement utilisée (Zhang, 2016) dans le cas de petits échantillons. Cette dernière consiste en une procédure séquentielle d'inclusion et d'exclusion de covariables selon des seuils de p-valeur ou des critères d'information (AIC, BIC) ; bien qu'implémentée dans de nombreux outils comme NONMEM et Monolix, elle peut se révéler chronophage et instable en grande dimension, particulièrement sur des jeux de données de petite taille. En comparaison, la méthode proposée effectue une sélection conjointe à l'estimation, ce qui favorise une meilleure robustesse prédictive lorsque la puissance statistique est limitée. Toutefois, ses performances se dégradent dans des échantillons plus larges ou en présence de colinéarité entre covariables. Pour pallier ces limites, Haem *et al.* (2017) proposent d'utiliser des variantes adaptatives du Lasso, telles que l'*Adaptive Lasso* (ALasso) et l'*Adjusted Adaptive Lasso* (AALasso), qu'ils ont intégrées dans NONMEM. Le ALasso, introduit initialement par Zou (2006), utilise les estimateurs des moindres carrés ordinaires comme poids initiaux afin d'ajuster la pénalisation de manière différenciée selon la magnitude des coefficients. Cette approche permet une pénalisation plus forte des coefficients proches de zéro, tout en allégeant la contrainte sur ceux significativement non nuls. Néanmoins, en présence de forte colinéarité entre les covariables, le ALasso peut s'avérer sous-optimal. Pour cette raison, le AALasso (Qian et Yang, 2013; Algamil et Lee, 2015) a été proposé : il tient compte de l'incertitude des estimateurs en utilisant comme poids initiaux le ratio entre l'erreur standard de l'estimateur du maximum de vraisemblance (EMV) et la valeur de cet EMV. Cette pondération rend la méthode plus robuste en situation de multicolinéarité. Haem *et al.* (2017) ont ainsi comparé les performances de l'estimation et de la sélection conjointe par Lasso, du ALasso et du AALasso dans NONMEM, montrant l'intérêt du AALasso dans les contextes complexes des modèles NLEM.

Bertrand et Balding (2013) ont introduit une approche itérative dans laquelle, après avoir estimé les paramètres du modèle via un algorithme SAEM, une régression pénalisée de type Lasso est effectuée sur les estimateurs empiriques *a posteriori* (EBEs). Toutefois, l'utilisation des EBEs comme réponse dans la régression pénalisée induit un biais de rétraction (ou *shrinkage*) :

lorsque les données individuelles sont limitées, les EBEs tendent à être excessivement rétractés vers la moyenne de population, ce qui peut biaiser les associations covariable-effet aléatoires et affaiblir la puissance de sélection (Nguyen *et al.*, 2017). Une alternative pour pallier ces limites consiste à intégrer une régression pénalisée directement dans les modèles NLEM, à chaque itération du SAEM, permettant une sélection simultanée des covariables et une estimation des paramètres (Bertrand *et al.*, 2015). Plutôt que de modifier l'algorithme de sélection, une autre approche est d'agir directement sur les EBEs afin de corriger leur biais avant la régression pénalisée. Récemment, Yuan *et al.* (2021) ont proposé une méthode de correction analytique du *shrinkage* des EBEs en amont de l'application du Lasso. En pratique, ils ajustent les EBEs calculés à partir du modèle sans covariables par une formule visant à enlever le biais de rétraction, obtenant des estimateurs dé-biaisés. L'utilisation de ces "EBEs corrigés" dans la régression Lasso permet théoriquement de restaurer une puissance de détection proche de celle qu'on aurait avec les véritables effets individuels non biaisés.

Parallèlement, plusieurs travaux ont proposé l'utilisation de méthodes de sélection pénalisée dans les modèles linéaires généralisés à effets mixtes ainsi que dans les modèles NLEM. En particulier, Pan et Huang (2014) ont introduit une approche visant à sélectionner directement les effets aléatoires pertinents en imposant une pénalité sur les composantes de la matrice de variance-covariance. Pour cela, ils utilisent une reparamétrisation basée sur la décomposition de Cholesky permettant d'annuler efficacement les composantes inutiles par une pénalisation adaptée. De leur côté, Li *et al.* (2018) ont développé une méthode de double régularisation pour la sélection simultanée des effets fixes et aléatoires, en combinant une pénalité L_1 sur les coefficients fixes et une pénalisation de type *group Lasso* sur les effets aléatoires. Cette stratégie permet de sélectionner simultanément les covariables fixes pertinentes et les effets aléatoires nécessaires tout en assurant la validité mathématique de la matrice de variance-covariance par une paramétrisation appropriée. L'estimation est réalisée par un algorithme efficace basé sur une descente de coordonnées. Plus récemment, Yu et Xiang (2024) ont développé une procédure itérative alternant la sélection pénalisée des effets fixes et aléatoires. Leur approche repose sur l'utilisation d'approximations variationnelles, permettant ainsi de réduire considérablement le coût computationnel associé aux modèles à effets mixtes pénalisés tout en conservant une sélection robuste et cohérente des paramètres. Ensuite, Fort *et al.* (2019) ont proposé deux algorithmes de gradient proximal pour intégrer des pénalités convexes, telles que le Lasso, dans les modèles à effets mixtes non-linéaires. Le premier, appelé *Monte Carlo Proximal-Gradient* (MCPG), approxime à chaque itération le gradient de la log-vraisemblance à l'aide d'un échantillon Monte Carlo des effets aléatoires. Le second, *Stochastic Approximation Proximal-Gradient* (SAPG), utilise la moyenne cumulée des points simulés au cours des itérations précédentes pour stabiliser l'approximation. Ces méthodes permettent une estimation conjointe des paramètres et une sélection automatique des covariables, avec des garanties théoriques établies dans le cas linéaire. Ollier (2022) ont ensuite adapté cette approche au cadre NLEM pour permettre la sélection simultanée des covariables et des paramètres de corrélation entre effets aléatoires.

Par ailleurs, dans le cadre de modèles NLEM conjoints à un modèle de survie, Caillebotte *et al.* (2024) ont permis l'intégration efficace des données transcriptomiques, en tant que cova-

riables, en employant une pénalisation de type Lasso. Leur méthode repose sur une optimisation par gradient stochastique couplée à un opérateur proximal, permettant de sélectionner les biomarqueurs influents parmi une multitude de candidats tout en estimant efficacement les paramètres du modèle conjoint. Enfin, [Sun et Basu \(2024\)](#) ont confirmé l'intérêt de combiner plusieurs biomarqueurs longitudinaux dans des modèles de survie, sélectionnés à l'aide d'une approche par Lasso. Cette intégration simultanée offre des modèles plus parcimonieux et interprétables biologiquement, en plus d'améliorer la performance prédictive, illustrant ainsi clairement l'intérêt de l'approche pénalisée pour la compréhension des mécanismes biologiques sous-jacents.

L'ensemble de ces travaux témoigne de l'intérêt croissant, en statistique, pour l'intégration de pénalités de type Lasso dans les approches d'estimation et de sélection de variables au sein des modèles NLEM, en particulier dans des contextes de grande dimension, dû à un accès à des données de plus en plus riches. Toutefois, la majorité de ces contributions ne s'appliquent pas directement aux modèles mécanistes fondés sur des EDOs. Dans cette thèse, nous nous inscrivons dans cette dynamique en s'appuyant sur certaines de ces approches de la littérature, nous proposons des approches originales pour l'estimation et la sélection des données de grande dimension dans les modèles NLEM basés sur des EDOs.

2.3 Utilisation des modèles mécanistes pour l'immunologie

Comme introduit au Chapitre 1, l'être humain est confronté à de nombreux pathogènes, tels que les virus, les bactéries, les parasites ou les champignons, susceptibles de provoquer des maladies. Lors de l'exposition à ces pathogènes, le système immunitaire orchestre une réponse d'abord innée puis adaptative, mobilisant de nombreux types de cellules et d'entités biologiques. La modélisation mathématique joue un rôle crucial pour mieux comprendre les mécanismes sous-jacents à ces réponses immunitaires et les caractéristiques d'une infection, afin de mieux s'en protéger, de vacciner plus efficacement et d'identifier les facteurs individuels pouvant moduler cette réponse. Ces modèles permettent de décrire les interactions dynamiques entre les différentes populations cellulaires impliquées dans la réponse immunitaire, notamment dans le contexte de la vaccination, qui est un intérêt de cette thèse. En particulier, les modèles NLEM se sont imposés pour représenter la variabilité inter- et intra-individuelle des réponses vaccinales.

Cependant, malgré leurs atouts, les modèles mécanistes les plus détaillés, comme ceux de [Bonin et al. \(2017\)](#) ou [Dari et al. \(2023\)](#) présentés dans les Sections 2.3.1 et 2.3.2 respectivement, s'avèrent souvent non identifiables à partir des seules données expérimentales disponibles. Ils contraignent à fixer *a priori* un grand nombre, voire la majorité, de leurs paramètres, ce qui représente en soit une tâche délicate pouvant induire des biais dans l'estimation finale. Cette complexité limite en outre leur application pratique, notamment dans le cadre de développements méthodologiques, où l'utilisation de modèles comportant un trop grand nombre de paramètres et de compartiments est impossible à cause de problèmes d'identifiabilités. C'est le cas par exemple quand seule une partie des compartiments est observée, comme pour la réponse immunitaire humorale où seules des observations du compartiment des anticorps sont généralement

disponibles (Pasin *et al.*, 2019; Alexandre, 2022). C'est pourquoi des modèles plus simples et parcimonieux sont généralement privilégiés comme point de départ pour le développement de nouvelles méthodes d'analyse. Par exemple, Andraud *et al.* (2012) ont proposé un modèle mécaniste humoral simplifié, présenté en Section 2.3.4, ne décrivant que deux compartiments de plasmocytes, c'est-à-dire les cellules sécrétrices d'anticorps (*Antibody Secreting Cells*, ASCs), à courte et longue durée de vie, et les anticorps produits correspondants. Ce type de modèle, moins paramétré, offre une base initiale propice à l'élaboration de méthodes tout en garantissant une meilleure identifiabilité et une interprétation biologique claire.

2.3.1 Modélisation du système immunitaire par Bonin *et al.*

Des modèles mathématiques complets ont ainsi été proposés pour modéliser la réponse immunitaire faisant suite à l'injection de vaccins contre différentes maladies. Par exemple, Bonin *et al.* (2017) ont développé un modèle de la réponse immunitaire à la vaccination contre la fièvre jaune, sous la forme d'un système d'équations différentielles ordinaires représentant les principales populations du système immunitaire (cellules B, cellules T CD4+ et CD8+, cellules présentatrices d'antigène, anticorps, etc.), et décrit par la Figure 2.1. Un tel modèle permet de simuler *in silico* la réponse à des vaccins en cours de développement ou d'étude, et d'évaluer différents scénarios en faisant varier les paramètres du modèle. En effet, Bonin *et al.* (2017) ont paramétré ce modèle pour la fièvre jaune et montré que les simulations obtenues étaient cohérentes avec les données expérimentales disponibles. Ce modèle se veut par ailleurs suffisamment générique pour s'appliquer à d'autres maladies ou vaccins viraux. Par la suite, Bonin *et al.* (2020) ont validé ce modèle à l'aide de données expérimentales issues de plusieurs études. Ce travail a démontré que les outils de modélisation et de simulation computationnelle peuvent être utilisés pour reproduire différents scénarios de vaccination contre la fièvre jaune et la réponse immunitaire correspondante, apportant ainsi des éléments de réponse à certaines questions en suspens.

Le modèle de Bonin *et al.* repose sur un système d'équations différentielles ordinaires décrivant l'évolution de chaque population importante dans la réponse immunitaire post-vaccinale, ainsi que du virus vaccinal lui-même. Ces populations incluent le virus vivant atténué de la fièvre jaune inoculé par le vaccin, les cellules présentatrices d'antigène immatures et matures, les lymphocytes T CD4+ naïfs et effecteurs, les lymphocytes T CD8+ naïfs et effecteurs, les lymphocytes B naïfs et activés, les plasmocytes à courte vie et à longue vie, les cellules B mémoires, et enfin les anticorps produits. Dans la suite, les paramètres du modèle sont mis en rouge afin de bien différencier les variables du modèle de ses paramètres.

Le système d'équations modélisant les composantes du modèle est donné ci-dessous :

$$\begin{aligned}
 (B) = \left\{ \begin{array}{l}
 \frac{dV}{dt} = \pi_V V - \frac{c_{v1} V}{c_{v2} + V} - k_{v1} VA - k_{v2} VT_{ke}, \quad (2.8) \\
 \frac{dA_p}{dt} = \alpha_{ap}(A_{p0} - A_p) - \beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2} + V}, \quad (2.9) \\
 \frac{dA_{pm}}{dt} = \beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2} + V} - \delta_{apm} A_{pm}, \quad (2.10) \\
 \frac{dT_{hn}}{dt} = \alpha_{th}(T_{hn0} - T_{hn}) - \beta_{th} A_{pm} T_{hn}, \quad (2.11) \\
 \frac{dT_{he}}{dt} = \beta_{th} A_{pm} T_{hn} + \pi_{th} A_{pm} T_{he} - \delta_{th} T_{he}, \quad (2.12) \\
 \frac{dT_{kn}}{dt} = \alpha_{tk}(T_{kn0} - T_{kn}) - \beta_{tk} A_{pm} T_{kn}, \quad (2.13) \\
 \frac{dT_{ke}}{dt} = \beta_{tk} A_{pm} T_{kn} + \pi_{tk} A_{pm} T_{ke} - \delta_{tk} T_{ke}, \quad (2.14) \\
 \frac{dB}{dt} = \alpha_b(B_0 - B) + \pi_{b1} VB + \pi_{b2} T_{he} B - \beta_{ps} A_{pm} B - \beta_{pl} T_{he} B - \beta_{bm} T_{he} B, \quad (2.15) \\
 \frac{dP_s}{dt} = \beta_{ps} A_{pm} B - \delta_{ps} P_s, \quad (2.16) \\
 \frac{dP_l}{dt} = \beta_{pl} T_{he} B - \delta_{pl} P_l + \gamma_{bm} B_m, \quad (2.17) \\
 \frac{dB_m}{dt} = \beta_{bm} T_{he} B + \pi_{bm1} B_m \left(1 - \frac{B_m}{\pi_{bm2}}\right) - \gamma_{bm} B_m, \quad (2.18) \\
 \frac{dA_b}{dt} = \pi_{ps} P_s + \pi_{pl} P_l - \delta_{ab} A_b. \quad (2.19)
 \end{array} \right.
 \end{aligned}$$

Dans ce système, chaque équation représente la dynamique d'une population donnée au cours du temps. L'Équation (2.8) décrit le virus vaccinal V . Celui-ci se multiplie à un taux π_V . Le terme $\frac{c_{v1} V}{c_{v2} + V}$ correspond à une élimination du virus lors de l'immunité innée, c'est-à-dire sans reconnaissance spécifique de celui-ci (Section 1.1.1). Le terme $k_{v1} VA$ représente l'élimination spécifique du virus due aux anticorps (par opsonisation et élimination médiée par le système immunitaire, Section 1.1.3), k_{v1} étant le taux d'élimination associé. Enfin, le terme $k_{v2} VT_{ke}$ modélise l'élimination spécifique du virus induite par l'apoptose des cellules infectées sous l'action des lymphocytes T CD8+ effecteurs T_{ke} , où k_{v2} est le taux d'élimination correspondant (Section 1.1.4).

Les CPAs regroupent des cellules telles que les dendritiques et les macrophages, capables d'internaliser des antigènes et de les présenter. Deux stades sont modélisés : les CPAs immatures A_p et les CPAs matures A_{pm} . L'Équation (2.9) décrit l'évolution des CPAs immatures. Le terme $\alpha_{ap}(A_{p0} - A_p)$ représente leur homéostasie, c'est-à-dire leur renouvellement à partir d'une valeur de base A_{p0} , avec un taux α_{ap} . Le terme $\beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2} + V}$ modélise l'activation des CPAs immatures en CPAs matures sous l'effet de la présence de virus. Ce même terme apparaît positivement dans l'Équation (2.10), qui décrit la population des CPAs matures. Dans l'Équation (2.10), le second terme $-\delta_{apm} A_{pm}$ traduit la disparition naturelle des CPAs matures, avec un taux de décroissance δ_{apm} .

Ensuite, les lymphocytes T CD4+ sont modélisés en distinguant les cellules naïves T_{hn} et les cellules effectrices auxiliaires T_{he} (Section 1.1.4). L'Équation (2.11) représente la dynamique des T CD4+ naïfs, avec un terme d'homéostasie $\alpha_{th}(T_{hn0} - T_{hn})$, avec α_{th} le taux maintenant la population à un niveau de base T_{hn0} , et un terme d'activation $-\beta_{th} A_{pm} T_{hn}$ correspondant à l'activation des T naïfs par les CPAs matures au taux β_{th} . L'Équation (2.12) décrit la population des T CD4+ effecteurs : elle s'accroît lorsque les T naïfs sont activés, décrit par le terme $\beta_{th} A_{pm} T_{hn}$, avec le même taux β_{th} , et par prolifération des effecteurs eux-mêmes sous stimulation des CPAs matures, décrit par le terme $\pi_{th} A_{pm} T_{he}$, avec π_{th} le taux de prolifération. Les T CD4+ effecteurs décroissent naturellement à un taux δ_{th} .

De façon analogue, les lymphocytes T CD8+ sont divisés en lymphocytes naïfs T_{kn} et lymphocytes effecteurs cytotoxiques T_{ke} . L'Équation (2.13) modélise les T CD8+ naïfs avec une homéostasie $\alpha_{tk}(T_{kn0} - T_{kn})$, au taux α_{tk} , et une activation par les CPAs matures $-\beta_{tk} A_{pm} T_{kn}$, au taux d'activation β_{tk} . L'Équation (2.14) représente les T CD8+ effecteurs, générés par l'activation des T naïfs, décrit par le terme $\beta_{tk} A_{pm} T_{kn}$, et proliférant sous l'action des CPAs matures, avec un taux π_{tk} , décrit par le terme $\pi_{tk} A_{pm} T_{ke}$. Ces cellules effectrices T CD8+ disparaissent avec un taux δ_{tk} et migrent vers le site d'infection vaccinal pour induire l'apoptose des cellules infectées.

L'Équation (2.15) modélise la population des lymphocytes B, en ne distinguant pas explicitement les B naïfs des B activés afin de simplifier le système. Le terme $\alpha_b(B_0 - B)$ représente l'homéostasie des cellules B, avec B_0 leur niveau de base et α_b le taux de renouvellement. Les termes $\pi_{b1} VB$ et $\pi_{b2} T_{he} B$ traduisent la prolifération des B stimulés soit par l'antigène, interaction directe avec le virus V , soit médiés par des T CD4+ effecteurs T_{he} . Par ailleurs, les termes $-\beta_{ps} A_{pm} B$, $-\beta_{pl} T_{he} B$ et $-\beta_{bm} T_{he} B$ correspondent à la différenciation des cellules B activées en plasmocytes à courte vie P_s , en plasmocytes à longue vie P_l et en cellules B mémoires B_m respectivement, les taux de différenciation étant respectivement β_{ps} , β_{pl} et β_{bm} (Section 1.1.3).

Les plasmocytes à courte durée de vie P_s , issus de la différenciation des B, sont décrits par l'Équation (2.16). Leur production est donnée par le terme $\beta_{ps} A_{pm} B$, issu des B activés sous stimulation des CPAs matures, et ils disparaissent naturellement à un taux δ_{ps} . Les plasmocytes à longue durée de vie P_l sont modélisés par l'Équation (2.17). Ils sont générés à partir des B activés avec l'aide des T auxiliaires, comme décrit par le terme $\beta_{pl} T_{he} B$, et subissent une décroissance naturelle $-\delta_{pl} P_l$. Un terme supplémentaire $\gamma_{bm} B_m$ traduit le maintien de ces plasmocytes à longue vie par les cellules B mémoires, avec γ_{bm} le taux de production de nouveaux plasmocytes long-vivants à partir des B_m .

L'Équation (2.18) représente les cellules B mémoires B_m . Celles-ci sont issues des lymphocytes B activés, via le terme $\beta_{bm} T_{he} B$, et leur dynamique suit une croissance logistique $\pi_{bm1} B_m(1 - B_m/\pi_{bm2})$, traduisant le fait que leur nombre tend vers une capacité maximale π_{bm2} , avec π_{bm1} le taux de croissance initial. Par ailleurs, le terme $-\gamma_{bm} B_m$ dans (2.18) représente la différenciation de ces cellules mémoires en plasmocytes à longue vie, au taux γ_{bm} .

Enfin, l'Équation (2.19) décrit la production des anticorps A_b par les plasmocytes : les

plasmocytes à courte vie produisent des anticorps à un taux π_{ps} , et ceux à longue vie à un taux π_{pl} ; ces anticorps sont éliminés progressivement selon un taux δ_{ab} .

Le schéma de la Figure 2.1 illustre de manière synthétique le modèle (B) issu des interactions entre toutes ces populations cellulaires, moléculaires et virales considérées dans le modèle. Ce type de modèle de réponse immunitaire vaccinale est très complet et détaillé, mais il présente des limitations computationnelles notables. En pratique, son utilisation s'avère difficile car il nécessite de fixer ou d'estimer un grand nombre de paramètres, ce qui alourdit le calcul et peut conduire à des incertitudes importantes. Bien que ces modèles mécanistes permettraient une description précise de la réponse vaccinale, leur complexité peut limiter leur applicabilité dans la pratique courante.

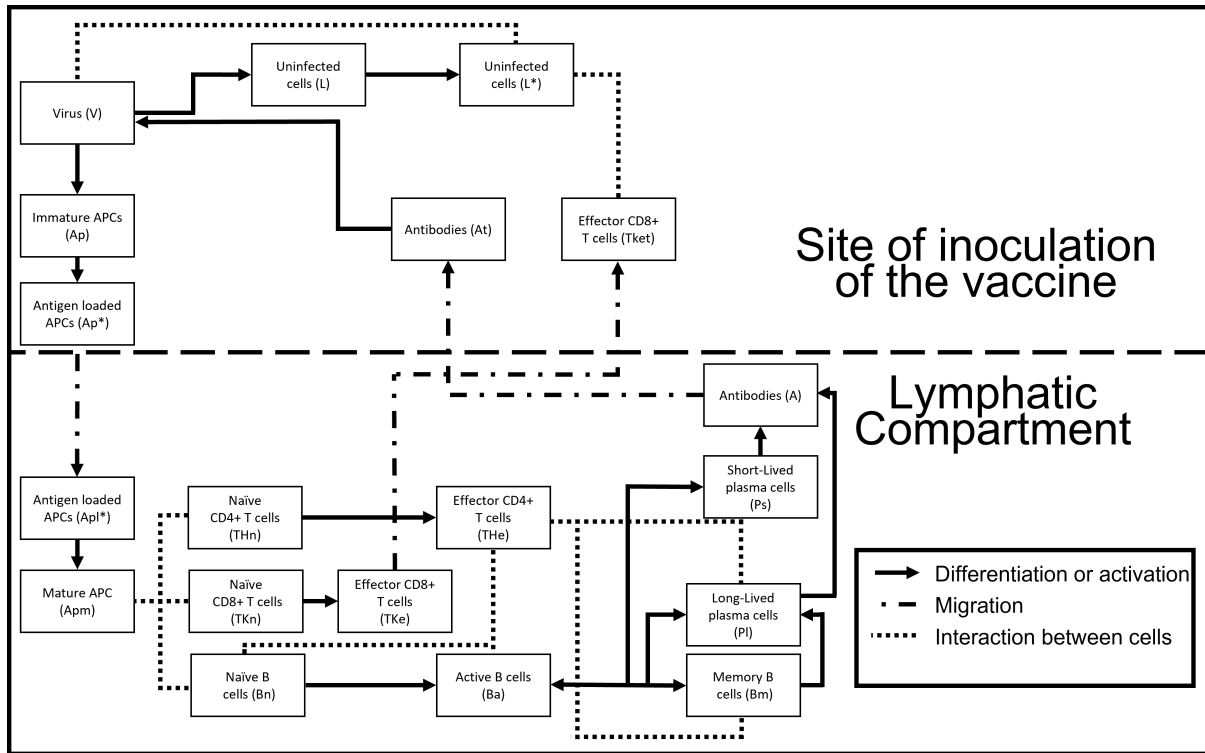


FIGURE 2.1 – Schéma de l'interaction entre les populations de cellules, de molécules et de virus prises en compte dans le modèle pour la fièvre jaune (Extrait de Bonin *et al.* (2017)).

2.3.2 Modélisation de la réponse immunitaire humorale par Dari *et al.*

Un second modèle, plus simple, a été proposé par Dari *et al.* (2023), avec pour objectif de modéliser la persistance des anticorps après vaccination par Ad26.COV2.S, à partir des données combinées de cinq essais cliniques. Dans cette étude, un modèle mécaniste est utilisé pour caractériser l'évolution temporelle de la réponse humorale sur une période de quinze mois, chez des participants initialement séronégatifs au SARS-CoV-2 et vaccinés avec une ou deux doses d'Ad26.COV2.S. Des simulations basées sur ce modèle ont ensuite permis de projeter la persistance des anticorps, définie comme le pourcentage de participants présentant des anticorps quantifiables, après une dose unique ou une dose de rappel homologue administrée à deux, trois

ou six mois d'intervalle.

La principale particularité de ce modèle est d'intégrer deux voies de prolifération distinctes pour la réponse des cellules B, afin de représenter deux mécanismes de différenciation conduisant à la production de plasmocytes. La première voie correspond aux cellules B mémoires, notées M_1 , qui aboutissent à la formation de plasmocytes à courte durée de vie (notés S pour *Short-Lived Plasma Cells*, SLPCs). La seconde représente la formation de centres germinatifs, notés M_2 , menant à des plasmocytes à longue durée de vie (notés L pour *Long-Lived Plasma Cells*, LLPCs). La Figure 2.2 illustre les composantes clés de la réponse humorale incluses dans ce modèle, défini par le système d'équations différentielles suivant :

$$(D) = \begin{cases} \frac{dA_g}{dt} = -k_{ag}A_g, & (2.20) \\ \frac{dM_1}{dt} = (k_{pAg}A_g + k_{pM}M_1) \left(1 - \frac{M_1}{N_1}\right) - (k_{pS}A_g + k_{eM_1})M_1, & (2.21) \\ \frac{dM_2}{dt} = (k_{pAg}A_g) \left(1 - \frac{M_2}{N_2}\right) - (k_{pL}A_g + k_{eM_2})M_2, & (2.22) \\ \frac{dS}{dt} = k_{pS}A_gM_1 - k_{eS}S, & (2.23) \\ \frac{dL}{dt} = k_{pL}A_gM_2 - k_{eL}L, & (2.24) \\ \frac{dA_b}{dt} = k_{AbS}S + k_{AbL}L - k_{eAb}A_b - k_{AbP}A_b + k_{PAb}P, & (2.25) \\ \frac{dP}{dt} = k_{AbP}A_b - k_{PAb}P. & (2.26) \end{cases}$$

L'Équation (2.20) décrit la dynamique du compartiment antigène A_g inoculé. Celui-ci est modélisé de manière simplifiée par un compartiment virtuel décrivant la disparition progressive de l'antigène dans l'organisme après l'injection, selon une décroissance exponentielle contrôlée par la constante k_{ag} .

Les Équations (2.21) et (2.22) régissent la prolifération des cellules B dans les deux compartiments M_1 et M_2 , correspondant respectivement aux cellules B mémoires et aux centres germinatifs. Dans ce modèle, les cellules B mémoires sont supposées responsables de la production de plasmocytes à courte durée de vie S , tandis que celles issues des centres germinatifs produisent les plasmocytes à longue durée de vie L . Chacune de ces équations comprend un terme de production stimulée par l'antigène A_g et modélisée par $k_{pAg}A_g$, modulé par un facteur de saturation logistique $\left(1 - \frac{M_i}{N_i}\right)$ limitant le nombre maximal de cellules, avec N_1 pour M_1 et N_2 pour M_2 . Elles comportent également un terme de disparition proportionnel à M_i , traduisant la perte de ces cellules par différenciation ($-k_{pS}A_gM_1$ et $-k_{pL}A_gM_2$ respectivement) ou par élimination naturelle ($-k_{eM_1}M_1$ et $-k_{eM_2}M_2$ respectivement). Pour les cellules B mémoires M_1 , Le terme $-(k_{pS}A_g + k_{eM_1})M_1$ reflète à la fois leur conversion en plasmocytes à courte vie S , selon un taux k_{pS} dépendant de l'antigène, et leur élimination naturelle au taux k_{eM_1} . Pour les centres germinatifs M_2 , le terme $-(k_{pL}A_g + k_{eM_2})M_2$ représente de façon analogue la différenciation en plasmocytes à longue vie L et leur décroissance naturelle. Pour les cellules B mémoires M_1 , un terme additionnel $k_{pM}M_1$ traduit leur capacité d'auto-renouvellement, en l'absence de

stimulation antigénique.

La production des plasmocytes à courte vie S et à longue vie L est décrite respectivement par les Équations (2.23) et (2.24). Ces processus de différenciation dépendent à la fois de la présence de l'antigène A_g et des cellules précurseurs, M_1 pour S et M_2 pour L , avec des taux de production k_{pS} et k_{pL} . Les cellules S sont éphémères, éliminées en quelques jours, au taux k_{eS} , tandis que les cellules L persistent plusieurs mois ou années, avec un taux de décroissance k_{eL} .

L'Équation (2.25) modélise la concentration des anticorps circulants A_b comme la somme des productions issues des deux types de plasmocytes. Les anticorps sont éliminés au taux k_{eAb} , mais peuvent également être transférés vers un compartiment périphérique P , comme par exemple les tissus, au taux k_{AbP} , et revenir au compartiment central au taux k_{PAb} . Ces échanges sont décrits dans les Équations (2.25) et (2.26).

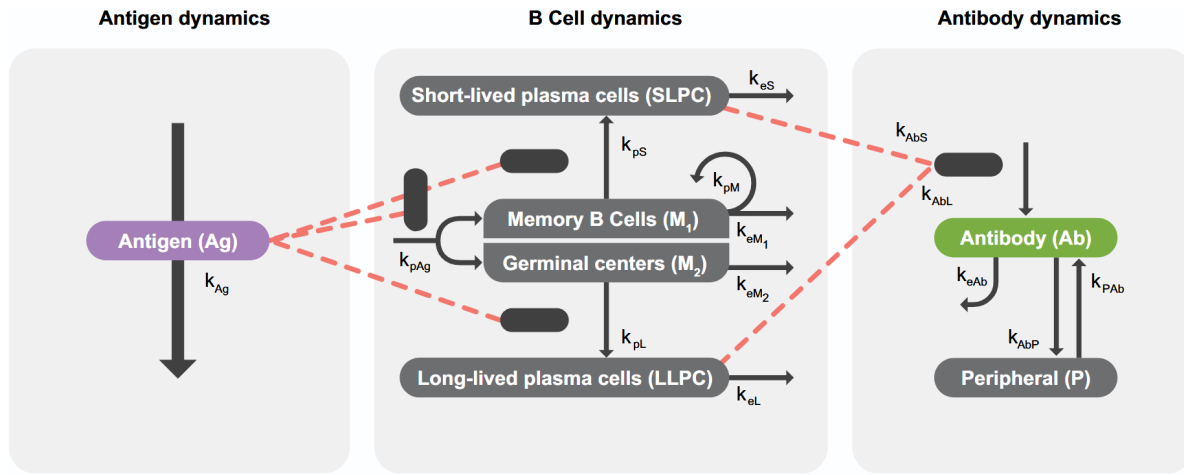


FIGURE 2.2 – Schéma du modèle mécaniste de la réponse immunitaire humorale dans le cadre de la vaccination contre le SARS-CoV-2 (Extrait de Dari *et al.* (2023)).

Malgré une structure plus simple que celle du modèle (B) proposé par Bonin *et al.* (2017), ce modèle repose aussi sur un grand nombre de paramètres fixés afin d'éviter les problèmes d'identifiabilité, ce qui en limite l'extension à des cadres d'analyse plus complexes.

2.3.3 Modélisation de la réponse humorale post-vaccinale par Balelli *et al.* et Clairon *et al.*

Balelli *et al.* (2020) ont proposé un modèle mécaniste décrivant l'établissement, le maintien et la réactivation de la réponse immunitaire humorale après une vaccination en deux doses contre le virus Ebola. Une représentation schématique en est donnée à la Figure 2.3. Le système s'écrit :

$$(Ba) = \begin{cases} \frac{dA}{dt} = -\delta_A A, & (2.27) \\ \frac{dM}{dt} = \tilde{\rho} A - (\tilde{\mu}_S + \tilde{\mu}_L) AM - \delta_M M, & (2.28) \\ \frac{dS}{dt} = \tilde{\mu}_S AM - \delta_S S, & (2.29) \\ \frac{dL}{dt} = \tilde{\mu}_L AM - \delta_L L, & (2.30) \\ \frac{dA_b}{dt} = \theta_S S + \theta_L L - \delta_{Ab} A_b. & (2.31) \end{cases}$$

La réponse immunitaire est initiée par l'injection d'un antigène A au temps $t = 0$, qui est ensuite éliminé selon un taux δ_A , comme modélisé dans l'Équation (2.27).

L'Équation (2.28) décrit la dynamique des cellules B mémoire M . L'antigène A présent conduit à la production de lymphocytes B mémoire M au taux $\tilde{\rho} A$. Ces cellules se différencient ensuite à la fois en plasmocytes à courte durée de vie S et à longue durée de vie L aux taux respectifs $\tilde{\mu}_S$ et $\tilde{\mu}_L$. Ce compartiment joue ainsi un rôle hybride, regroupant les fonctions des compartiments M_1 et M_2 du modèle (D). Ces cellules sont ensuite éliminées au taux δ_M .

Les Équations (2.29) et (2.30) modélisent les deux populations de plasmocytes S et L . Les cellules M se différencient en plasmocytes S au taux $\tilde{\mu}_S$ et en plasmocytes L au taux $\tilde{\mu}_L$. Ces plasmocytes disparaissent ensuite aux taux δ_S et δ_L respectivement. Enfin, les anticorps A_b , modélisés dans l'Équation (2.31), sont produits par les plasmocytes S avec un taux de sécrétion θ_S et par les plasmocytes L au taux θ_L , puis éliminés au taux δ_{Ab} .

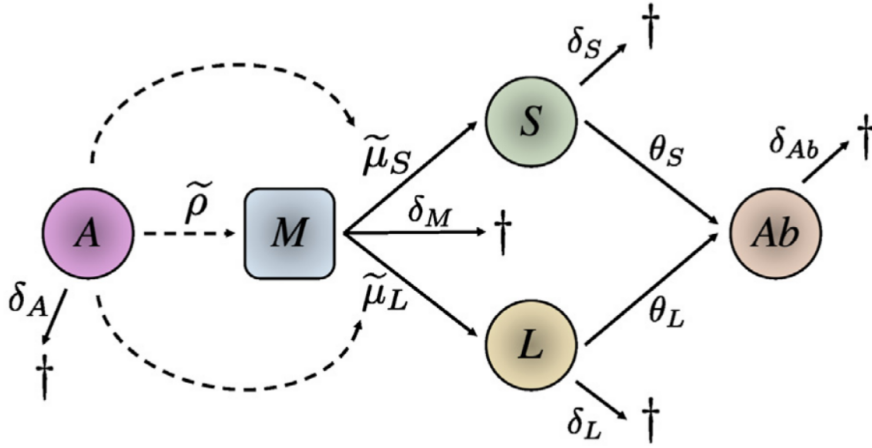


FIGURE 2.3 – Schéma du modèle mécaniste de la réponse immunitaire humorale post-vaccinale (extrait de Balelli *et al.* (2020)).

Ce modèle reproduit fidèlement les dynamiques d'anticorps IgG observées après vaccination. Il met en évidence l'importance du compartiment mémoire pour assurer la réponse à une dose de rappel et la persistance de l'immunité à long terme. Toutefois, certains paramètres, notamment $\tilde{\mu}_L$, $\tilde{\mu}_S$, $\tilde{\rho}$, θ_L et θ_S , ne peuvent pas être identifiés de manière unique à partir des

seules courbes d'anticorps, ce qui pose un problème d'identifiabilité structurelle. Pour lever cette indétermination, il est nécessaire soit de fixer *a priori* certains de ces paramètres, soit d'intégrer des données supplémentaires sur d'autres composantes de la réponse. Par exemple, disposer de la cinétique de $(S + L)(t)$ ainsi que de $A_b(0)$ permettrait d'assurer l'identifiabilité du modèle.

Ce modèle a ensuite été simplifié par [Clairon et al. \(2023b\)](#) pour modéliser la réponse humorale à un schéma multi-doses du vaccin BNT162b2 contre le SARS-CoV-2. Les auteurs proposent une version réduite du modèle ([Ba](#)), suffisamment flexible pour capturer l'effet de rappels successifs tout en évitant les problèmes d'identifiabilité.

Premièrement, les compartiments S et L sont fusionnés en un unique compartiment effecteur S , les analyses préliminaires n'ayant pas montré d'amélioration statistique à distinguer les deux populations. Le système devient alors :

$$\begin{cases} \frac{dM}{dt} = \rho A - \mu AM - \delta_M M, \\ \frac{dS}{dt} = \mu AM - \delta_S S, \\ \frac{dA_b}{dt} = \theta_S S - \delta_{Ab} A_b. \end{cases}$$

Deuxièmement, entre deux doses, le compartiment M n'est plus modélisé dynamiquement mais fixé à une valeur constante $\overline{M}_k$ immédiatement après l'injection k , par analogie avec [Balelli et al. \(2020\)](#). Troisièmement, la dynamique de l'antigène est remplacée par sa solution analytique entre deux doses : $A(t) = A_0 e^{-\delta_A(t-t_k)}$ pour $t \in [t_k, t_{k+1}]$, où t_k est le temps de la $k^{\text{ème}}$ injection et A_0 supposé identique pour chaque dose. Le système devient :

$$\begin{cases} \frac{dS}{dt} = \mu \overline{M}_k A_0 e^{-\delta_A(t-t_k)} - \delta_S S, \\ \frac{dA_b}{dt} = \theta_S S - \delta_{Ab} A_b. \end{cases}$$

Enfin, une invariance par changement d'échelle permet de normaliser S : pour tout $a > 0$,

$$\begin{cases} \frac{dS_a}{dt} = a \mu \overline{M}_k A_0 e^{-\delta_A(t-t_k)} - \delta_S S_a, \\ \frac{dA_b}{dt} = \frac{\theta_S}{a} S_a - \delta_{Ab} A_b, \end{cases}$$

la dynamique de A_b reste inchangée. En prenant $a = (\mu \overline{M}_k A_0)^{-1}$, on obtient le système final :

$$(C) = \begin{cases} \frac{dS}{dt} = \overline{f_{M_k}} e^{-\delta_A(t-t_k)} - \delta_S S, & (2.32) \\ \frac{dA_b}{dt} = \vartheta S - \delta_{Ab} A_b, & (2.33) \end{cases}$$

où $\overline{f_{M_k}} = \frac{\overline{M}_k}{M_1}$ est le facteur d'accroissement du compartiment mémoire après la $k^{\text{ème}}$ injection.

Le modèle (C) permet ainsi de tenir compte de l'effet des rappels vaccinaux sur la cinétique des anticorps, tout en conservant une structure simple et identifiable. Il a été appliqué à l'analyse de la réponse humorale induite par une, deux ou trois doses de vaccin BNT162b2, et a permis d'estimer l'impact du nombre de doses et de l'âge sur la rapidité et l'intensité de la réponse.

2.3.4 Modèle à deux compartiments de plasmocytes par Andraud *et al.*

Andraud *et al.* (2012) ont proposé un modèle mécaniste simple de la réponse humorale, reposant sur deux sous-populations de plasmocytes : les plasmocytes à courte durée de vie S et à longue durée de vie L . Ce modèle, développé initialement pour étudier la cinétique des anticorps après vaccination contre le virus de l'hépatite A, repose sur l'idée que ces deux types de plasmocytes produisent des anticorps circulants A_b qui sont ensuite éliminés progressivement selon leurs propres dynamiques. La cinétique conjointe des plasmocytes et des anticorps est décrite par le système suivant :

$$(A) = \begin{cases} \frac{dS}{dt} = -\delta_S S, & (2.34) \\ \frac{dL}{dt} = -\delta_L L, & (2.35) \\ \frac{dA_b}{dt} = \varphi_S S + \varphi_L L - \delta_{Ab} A_b, & (2.36) \end{cases}$$

où δ_S , δ_L et δ_{Ab} sont les taux de décroissance des plasmocytes S , L et des anticorps A_b respectivement, tandis que φ_S et φ_L désignent les taux de production d'anticorps par les cellules S et L .

Le système (A) peut être simplifié en résolvant explicitement les équations de S et L , qui suivent des lois exponentielles : $S(t) = S_0 e^{-\delta_S t}$ et $L(t) = L_0 e^{-\delta_L t}$, avec S_0 et L_0 les effectifs initiaux de plasmocytes après stimulation antigénique. En injectant ces solutions dans l'équation de A_b , on obtient une équation différentielle réduite pour la dynamique des anticorps :

$$\frac{dA_b}{dt} = \theta_S e^{-\delta_S t} + \theta_L e^{-\delta_L t} - \delta_{Ab} A_b, \quad (2.37)$$

où $\theta_S = \varphi_S S_0$ et $\theta_L = \varphi_L L_0$ représentent les taux initiaux de production d'anticorps induits par chaque sous-population.

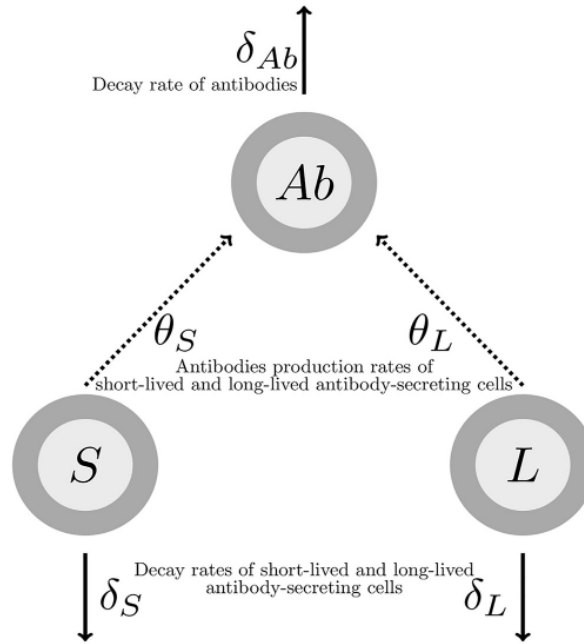


FIGURE 2.4 – Modèle de la réponse immunitaire humorale distinguant deux sous-populations de plasmocytes (extrait de [Pasin *et al.* \(2019\)](#)).

Ce type de modèle mécaniste, à la fois simple et biologiquement interprétable, s'est révélé efficace pour décrire et prédire la dynamique des anticorps post-vaccinaux dans des contextes variés. Son faible nombre de paramètres, sa structure analytique explicite et sa capacité à s'adapter à des jeux de données restreints en font un outil particulièrement adapté aux applications translationnelles. Il permet notamment de comparer différents schémas vaccinaux ou populations, et d'identifier les déterminants de la persistance humorale, là où des modèles plus complexes se heurteraient à des limitations d'estimation ou d'interprétation.

Par exemple, [Pasin *et al.* \(2019\)](#) ont utilisé ce modèle pour quantifier les différences de réponse humorale à long terme entre populations européennes et africaines 7 jours après une seconde dose de vaccin contre le virus Ebola, mettant en évidence une persistance significativement plus élevée dans la population européenne. Plus récemment, [Alexandre *et al.* \(2023\)](#) ont montré que ce même modèle permettait d'anticiper la durée de protection induite par différents schémas vaccinaux, en prédisant la proportion d'individus restant au-dessus d'un seuil d'anticorps neutralisants plusieurs années après injection.

2.3.5 Discussion sur l'identifiabilité et la complexité des modèles

Les modèles présentés ci-dessus illustrent le compromis entre le degré de détail d'un modèle immunologique et la possibilité d'en estimer les paramètres à partir des données disponibles. Le modèle complet de [Bonin *et al.* \(2017\)](#) offre une représentation exhaustive de la réponse immunitaire à une injection de vaccin, en intégrant de nombreuses composantes immunitaires, mais au prix d'un très grand nombre de paramètres à estimer. Dans la pratique, la calibration

d'un tel modèle requiert souvent de fixer de nombreux, voire la majorité, des paramètres sur la base de connaissances externes ou de la littérature existante, sans quoi le modèle n'est pas identifiable, les données expérimentales disponibles ne permettant en général d'observer qu'un nombre restreint de compartiments parmi ceux décrits par le modèle. À l'opposé, le modèle plus simple de [Andraud et al. \(2012\)](#), réduit à deux compartiments de plasmocytes et aux anticorps, comporte peu de paramètres et se révèle utilisable en pratique.

D'autres approches intermédiaires existent, [Dari et al. \(2023\)](#), par exemple, ont restreint leur modèle à la seule réponse humorale (antigène, plasmocytes à courte et longue vie, anticorps), limitant ainsi la complexité tout en reproduisant les mécanismes clés de la vaccination. De même, [Balelli et al. \(2020\)](#) ont introduit explicitement la différenciation des plasmocytes à longue vie afin de modéliser la persistance des anticorps, en ajoutant un compartiment de mémoire immunologique. Cependant, le modèle de Balelli et al. n'est pas identifiable à partir des cinétiques d'anticorps seules : estimer séparément la contribution des plasmocytes à courte et à longue vie nécessiterait soit de fixer certains paramètres *a priori*, soit de disposer de mesures expérimentales additionnelles. La simplification proposée par [Clairon et al. \(2023b\)](#) a justement permis de lever cette indétermination structurelle et d'obtenir un modèle plus simple, sans sacrifier la capacité à reproduire l'effet de rappels vaccinaux.

Ces exemples soulignent l'importance d'adapter la complexité du modèle à l'information apportée par les données. À terme, l'objectif est de pouvoir exploiter pleinement le potentiel des modèles détaillés tout en garantissant leur identifiabilité.

Dans ce contexte une piste prometteuse réside dans l'intégration de données de transcriptomiques, par exemple les profils d'expression génique, afin de fournir des observations supplémentaires pour la calibration des modèles. Certains marqueurs moléculaires, tels que des gènes dont l'expression reflète l'activité de cellules immunitaires spécifiques, pourraient alors servir de proxies pour suivre indirectement la dynamique de compartiments difficilement mesurables jusqu'alors. L'utilisation conjointe de ces données de grande dimension et de marqueurs plus facilement accessibles, pouvant être mis en correspondance avec des compartiments latents jusque-là inobservés, et de modèles mécanistes complexes, permettrait ainsi de mieux estimer les paramètres et d'atténuer une partie des problèmes d'identifiabilité dont souffrent ces modèles, ouvrant la voie à des analyses plus précises de la réponse immunitaire.

– Chapitre 3 –

Sélection de covariables par une approche Lasso et de sélection stable dans un algorithme de construction de modèles non-linéaires à effets mixtes

Publication :

Ce chapitre présente le travail de l'article suivant :

Gabaut, A., Thiébaud, R., Proust-Lima, C. & Prague, M. (2025). **A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effects Model**. *En révision dans Statistics in Medicine*. <https://doi.org/10.1101/2025.05.15.25327667>

Un paquet R LSAMBA est disponible sur le CRAN.

Table des matières

3.1	Positionnement et contribution	75
3.2	Article : A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effects Model	80
3.3	Travaux additionnels : Analyse des données de l'essai clinique PREVAC-UP	96
3.4	Perspectives	100

3.1 Positionnement et contribution

Ce chapitre s'intéresse à l'intégration de données de grande dimension, comme les profils d'expression génique, présentés en Section 1.2.2.2, dans les modèles mécanistes, introduits en Section 2.1. Plus précisément, nous cherchons à incorporer ces données, mesurées à des instants fixes en variables explicatives $\mathbf{X}_i$, dans le modèle statistique d'un modèle NLEM mécaniste, défini en Section 2.1.1.3, de la manière suivante :

$$\begin{aligned} Y_{ij} &= f(S_i(t_{ij}), \psi_i) + f_{\text{err}}(S_i(t_{ij}), \psi_i, \xi) \varepsilon_{ij}, \\ h(\psi_i) &= h(\psi_0) + \mathbf{X}_i \beta + \eta_i. \end{aligned}$$

Sauf mention explicite, le modèle d'erreur f_{err} considéré est celui constant $f_{err} = \sigma > 0$. Les notations présentées dans les sections précédentes sont conservées dans le chapitre suivant.

L'objectif est ici d'identifier quelles composantes du vecteur de coefficients β sont non nulles, c'est-à-dire quelles covariables ont un effet significatif sur les paramètres du modèle.

Plusieurs stratégies ont été développées ces dernières années pour répondre à ce défi. Une approche classique consiste en un processus en deux étapes. Dans un premier temps, on estime les paramètres du modèle NLEM sans inclure de covariables. Dans un second temps, on ajuste un modèle de régression liant ces paramètres estimés aux covariables candidates, en utilisant par exemple une méthode de sélection pas à pas (*stepwise*) (Lehr et al., 2010). En particulier, Ribbing et al. (2007) ont introduit l'utilisation du Lasso pour la construction de modèles de covariables en pharmacocinétique populationnelle. Le Lasso réalise une sélection automatique des variables en contraignant certains coefficients β à zéro via une pénalisation L_1 . Haem et al. (2017) ont aussi proposé un Lasso adaptatif ajusté pour pallier les problèmes de multicolinéarité entre covariables. Bertrand et Balding (2013) ont comparé la stratégie *stepwise* classique à plusieurs méthodes pénalisées (*Ridge*, *Lasso* et *hyper-Lasso*) pour la sélection de covariables génétiques, concluant à une meilleure efficacité des approches pénalisées. Bertrand et al. (2015) ont ensuite proposé une version pénalisée de l'algorithme SAEM intégrant directement la pénalisation dans l'étape de maximisation, et Fort et al. (2019) et Ollier (2022) ont proposé des estimateurs avec des pénalités plus générales dans le cas des modèles non-linéaires à effets mixtes utilisant des algorithmes de gradient proximal. Ces approches sont mentionnées précédemment dans la Section 2.2.3.

Une alternative est d'intégrer la sélection de covariables directement au sein de l'estimation du modèle NLEM. La méthode SCM (*Stepwise Covariate Modeling*), initialement popularisée en pharmacométrie à la fin des années 1990, est devenue la procédure standard pour identifier les relations covariables-paramètres significatives. Il s'agit d'une approche itérative pas à pas qui ajoute puis retire systématiquement des covariables du modèle en fonction de tests statistiques fondés sur la vraisemblance (Jonsson et Karlsson, 1998). Concrètement, on commence typiquement avec un modèle de base sans covariables, puis, à chaque étape de la phase d'inclusion, on ajoute la covariable qui améliore le plus le critère d'ajustement, par exemple, la log-vraisemblance de manière significative ou un critère d'information. Une fois qu'aucune covariable additionnelle n'apporte d'amélioration significative, une phase d'élimination peut être réalisée pour retirer les covariables dont l'effet n'est plus jugé significatif dans le modèle final. Le processus s'arrête lorsqu'aucune inclusion ou exclusion ne conduit à une amélioration significative du modèle. La méthode SCM requiert généralement l'estimation d'un très grand nombre de modèles intermédiaires et peut donc être très coûteuse en temps de calcul. De plus, son efficacité à identifier les covariables vraiment pertinentes n'est pas optimale, en particulier dans des contextes complexes avec de nombreuses covariables candidates.

Afin d'améliorer l'efficacité de la méthode, des variantes de SCM ont été proposées. Par exemple, SCM+ (Svensson et Jonsson, 2022) améliore le SCM standard en introduisant deux mécanismes. Le premier est une réduction adaptative du champ de recherche : seules les covariables ayant montré un effet potentiel dans des itérations précédentes ou dans des modèles

plus simples sont réévaluées, ce qui permet de réduire le nombre de modèles testés. Le second est un filtrage hiérarchique des covariables selon leur degré de justification biologique ou empirique. Les covariables sont ainsi classées en trois catégories : les *mécanistes*, qui reposent sur une relation biologique bien établie avec les paramètres du modèle ; les *structurelles*, qui sont généralement incluses sur la base d’observations empiriques ou de conventions, comme pour des covariables démographiques telles que l’âge ou le sexe ; et enfin les *exploratoires*, qui ne reposent sur aucune hypothèse forte *a priori*, comme c’est souvent le cas des données de transcriptomiques. La sélection commence par les covariables mécanistes, puis passe aux structurelles, et enfin aux exploratoires, en interrompant le processus dès qu’aucune covariable d’une catégorie donnée n’améliore significativement le modèle. Cette stratégie hiérarchisée permet d’accélérer considérablement la construction du modèle tout en favorisant l’identification de relations pertinentes.

Dans la même optique, [Ayrat et al. \(2021\)](#) ont proposé la méthode COSSAC (*COnditional Sampling use for Stepwise Approach based on Correlation tests*) pour optimiser la procédure pas à pas. COSSAC est une méthode itérative qui exploite l’information issue du modèle courant en réalisant, à chaque itération, des tests de corrélation entre chaque covariable candidate et les paramètres individuels estimés, provenant de la distribution conditionnelle des effets aléatoires sous le modèle actuel. Seules les covariables montrant une corrélation suffisamment marquée avec un paramètre sont alors testées dans le modèle lors de l’étape d’inclusion, ce qui réduit considérablement le nombre de modèles évalués tout en conservant les améliorations de vraisemblance potentielles. De même, lors des étapes d’élimination, des tests de corrélation permettent de détecter les covariables dont l’effet pourrait être négligeable. Grâce à cette stratégie de présélection, COSSAC aboutit, dans la grande majorité des cas, au même modèle final que SCM, tout en nécessitant entre 2 et 20 fois moins d’estimations de modèles. Toutefois, dans les contextes de très grande dimension, les méthodes *stepwise* classiques comme SCM ou COSSAC deviennent pratiquement inutilisables en temps raisonnable, en raison du nombre exponentiel de modèles à estimer, rendant l’exploration complète de l’espace des covariables irréaliste.

Une autre approche récente est l’algorithme SAMBA (*Stochastic Approximation for Model Building Algorithm*), introduit par [Prague et Lavielle \(2022\)](#), qui vise à automatiser la construction de modèles NLEM complets. Il permet la sélection conjointe de trois composantes du modèle mécaniste :

- (1) **le modèle de covariables** $\mathcal{M}^{\text{COV}}$, en identifiant quelles covariables influencent les paramètres du modèle, via la sélection des composantes non nulles de β ;
- (2) **le modèle d’erreur** $\mathcal{M}^{\text{ERR}}$, en choisissant, pour chaque variable observée, la forme fonctionnelle de l’erreur résiduelle la plus appropriée parmi les modèles standards (constante, proportionnelle ou combinée, définis en Section 2.1.1.3) ;
- (3) **le modèle de corrélation** $\mathcal{M}^{\text{CORR}}$, en inférant la structure de la matrice de variance-covariance Ω des effets aléatoires, c’est-à-dire en déterminant quels coefficients de corrélation doivent être considérés comme nuls ou non.

Le fonctionnement de l'algorithme repose sur une procédure itérative de construction de modèle. À l'itération k , à partir du modèle courant $\mathcal{M}^{(k-1)}$ estimé précédemment, plusieurs modifications ponctuelles sont envisagées sur chacun des sous-modèles $\mathcal{M}^{\text{COV}}$, $\mathcal{M}^{\text{ERR}}$ et $\mathcal{M}^{\text{CORR}}$. Ces modifications sont définies à partir des paramètres du modèle courant $\theta^{(k-1)}$ estimés via SAEM, et des paramètres individuels $(\psi_i^{(k-1)})_{i \leq N}$ simulés par MCMC. Les modifications testées incluent typiquement l'ajout ou le retrait d'une covariable, le changement de modèle d'erreur, ou encore l'introduction d'un terme de corrélation entre deux effets aléatoires. Chaque version modifiée du modèle est ensuite évaluée à l'aide d'un critère d'information, comme le BICc, et la modification apportant la plus grande amélioration est retenue. Le modèle $\mathcal{M}^{(k)}$ est ainsi obtenu, et le processus est répété jusqu'à ce qu'aucune modification ne permette d'améliorer le critère. L'intérêt principal de SAMBA réside dans sa capacité à réduire considérablement le nombre total de modèles testés, en évitant une exploration exhaustive de l'espace des modèles, tout en assurant une descente rigoureuse du critère d'information, ce qui garantit la convergence vers un modèle optimal local au sens du critère choisi. Le schéma de l'algorithme SAMBA, tel que présenté dans [Prague et Lavielle \(2022\)](#), est rappelé à la Figure 3.1 pour illustrer le déroulement général de cette procédure.

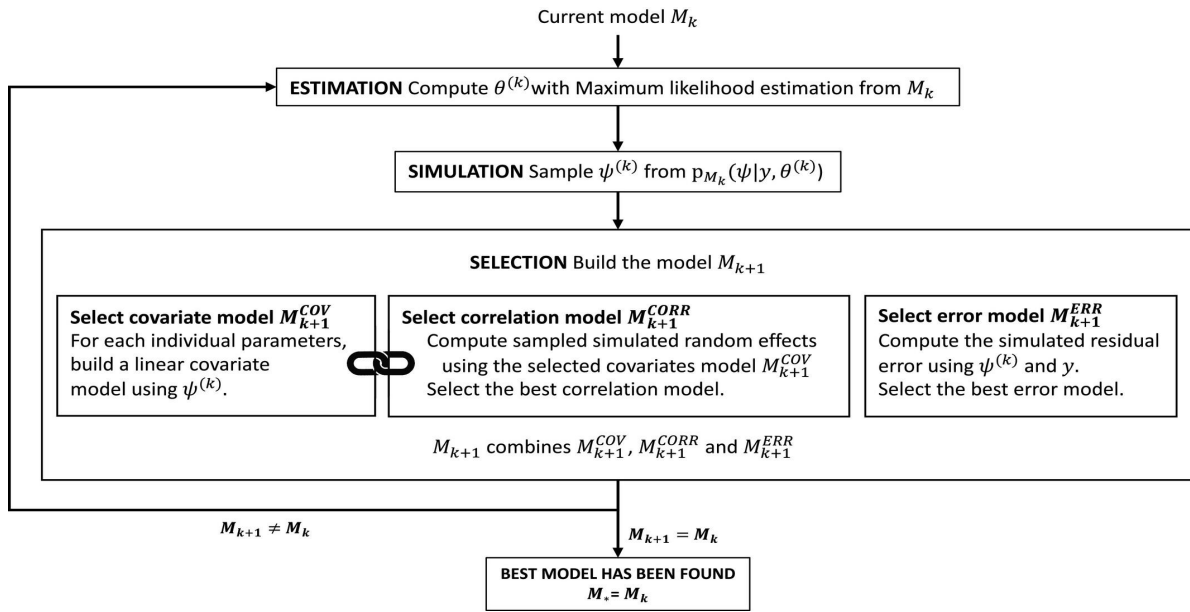


FIGURE 3.1 – Schéma de l'algorithme SAMBA (extrait de [Prague et Lavielle \(2022\)](#)).

Dans sa version originale, la sélection de covariables dans SAMBA repose sur un algorithme de sélection pas à pas ([Zhang, 2016](#)). L'ajout séquentiel des covariables, puis leur retrait, sont testés dans le cadre du modèle de régression suivant :

$$h(\psi_i^{(k)}) = \mu + X_i \beta + \eta_i, \quad i \leq N,$$

où η_i regroupe des erreurs gaussiennes représentant les effets aléatoires de l'individu i sur les différents paramètres, μ est le vecteur des ordonnées à l'origine du modèle de régression, et X_i

désigne les covariables individuelles. Ce modèle est défini à partir des paramètres individuels $(\psi_i^{(k)})_{i \leq N}$ simulés à l'aide de l'étape MCMC sous le modèle courant $\mathcal{M}^{(k)}$. Contrairement à SCM ou COSSAC, cette sélection ne s'effectue donc pas directement sur le modèle complet, mais sur le modèle de régression des effets individuels, ce qui permet une réduction substantielle du coût computationnel à chaque itération. Néanmoins, dans des contextes de grande dimension, la sélection de covariables par algorithme *stepwise* se comporte mal et devient sujette à de l'*overfitting*.

Enfin, des approches bayésiennes récentes ont également été développées, comme celle de Naveau *et al.* (2023) appelé SAMEVS (*SAEM Variable Selection*), combinant un *prior spike-and-slab* multivarié gaussien sur β et une estimation par SAEM-MCMC. Le principe du *prior spike-and-slab* est d'introduire un mécanisme de sélection via un mélange de deux lois gaussiennes : une première composante, dite *spike*, très concentrée autour de zéro, représentant les coefficients supposés nuls ou négligeables, caractérisée par une variance ν_0 , et une seconde composante, dite *slab*, plus étalée, permettant aux coefficients d'avoir des valeurs potentiellement non nulles et caractérisée par une variance ν_1 . Dans ce *prior*, c'est la variance de la distribution *spike*, ν_0 , qui permet la sélection des covariables et constitue donc le paramètre de pénalisation de la méthode. Celui-ci est choisi par minimisation d'un critère d'information. Après estimation par SAEM-MCMC du MAP des paramètres du modèle, la probabilité *a posteriori* d'inclusion de la covariable β_c est calculée, permettant d'en dériver un seuil d'inclusion dépendant des variances ν_0 , ν_1 et de cet estimateur MAP. Si le coefficient β_c est en dessous de ce seuil à l'issue de l'optimisation, alors la covariable c n'est pas sélectionnée dans le modèle final. Ce seuil *a posteriori* permet ainsi de discriminer les covariables sélectionnées.

Dans ce travail, nous proposons une méthode de construction de modèles NLEM en grande dimension, basée sur l'algorithme SAMBA (Prague et Lavielle, 2022), dans laquelle la sélection *stepwise* est remplacée par une sélection par Lasso (Tibshirani, 1996). L'objectif est d'identifier le support de β , défini par

$$\mathcal{S} = \{c \leq p \mid \beta_c \neq 0\}.$$

Pour cela, nous utilisons une procédure de sélection par Lasso couplée à un algorithme de sélection stable (Meinshausen et Bühlmann, 2010; Shah et Samworth, 2013; Bodinier *et al.*, 2023), présenté plus en détail en Section 2.2.2.3. Pour un paramètre de pénalisation λ donné, on note $\hat{\pi}^\lambda(c)$ la fréquence de sélection de la covariable c parmi les sous-échantillons générés. L'ensemble des covariables sélectionnées est alors défini par

$$\hat{\mathcal{S}}^{\lambda, \pi} = \{c \leq p \mid \hat{\pi}^\lambda(c) \geq \pi\}.$$

Cette méthode repose donc sur deux hyperparamètres : $\lambda > 0$ et $\pi \in [0, 1]$. Nous calibrons ces paramètres de façon conjointe, en combinant deux critères : (1) la minimisation du BICc, présenté en Section 2.2.1, et (2) le respect d'une contrainte sur le nombre attendu de faux positifs, qui doit être inférieur à un seuil α fixé par l'utilisateur, présenté en Section 2.2.2.3. Autrement dit, parmi les couples (λ, π) satisfaisant la contrainte sur les faux positifs, nous sélectionnons celui qui minimise le BICc afin d'assurer la convergence de l'algorithme.

A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effects Model

Auriane Gabaut^{1,2,3}, Rodolphe Thiébaud^{1,2,3,4}, Cécile Proust-Lima¹, Mélanie Prague^{1,2,3*}

¹ : Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France

² : Univ. Bordeaux, Inria, Bordeaux, France

³ : Vaccine Research Institute, Paris, France

⁴ : CHU de Bordeaux, Service d'Information Médicale, INSERM, U1219, F-33000 Bordeaux, France

* : melanie.prague@inria.fr

Abstract : Non-linear mixed effect models (NLMEMs) defined by ordinary differential equations (ODEs) are central to modeling complex biological systems over time, particularly in pharmacometrics, virus dynamics and immunology. However, selecting relevant covariates associated with a dynamics in high-dimensional settings remains a major challenge. This study introduces a novel model-building approach called Lasso-SAMBA that integrates Lasso regression with a stability selection algorithm for robust covariate selection within ODE-based NLMEMs.

The method iteratively constructs models by coupling penalized regression with mechanistic model estimation using SAEM algorithm. It extends a prior strategy (SAMBA : Stochastic Approximation for Model Building Algorithm) originally based on stepwise inclusion, by replacing this step with a penalized, stability-driven approach that reduces false discoveries and improves selection robustness. By maintaining the monotonic decrease of the information criterion through a calibrated exploration of penalization parameters, the proposed method outperforms conventional stepwise and Bayesian variable selection alternatives.

Extensive simulation studies, spanning pharmacokinetic and immunological models, demonstrate the superiority of Lasso-SAMBA in variable selection fidelity, FDR (False Discovery Rate) control, and computational efficiency. The Lasso-SAMBA method is implemented in an R package. Applied to a Varicella-Zoster virus vaccination study, the method reveals robust, biologically plausible associations between parameters of the mechanistic model of the humoral immune response with early transcriptomic expressions. These results underscore the practical utility of our method for high-dimensional model building in systems vaccinology and beyond.

Keywords : non-linear mixed effect model - model selection - Lasso regression - high-dimension - mechanistic models

Acknowledgment

This study was carried out in the framework of the University of Bordeaux’s France 2030 program RRI PHDS. We thank the Human Immunology Project Consortium for providing the ImmPort platform for open-access immunology data. We also thank the investigators who contributed data for study SDY984, supported by the Human Immunology Project Consortium (HIPC) RFA-AI-15-041. We thank Simulations Plus, Lixoft division for the free academic use of the MonolixSuite. Computer time for this study was provided by the computing facilities of the PlaFRIM (Plateforme Fédérative pour la Recherche en Informatique et Mathématiques). We disclose that generative AI tools were used to polish English language.

1 Introduction

ODE-based models, referred to as mechanistic models, are powerful tools for understanding complex biological processes. They are used in many fields such as virus dynamics (Perelson et Ribeiro, 2018), epidemic modeling (Ganser et al., 2024), and pharmacometrics (Ibarra et al., 2021). They were introduced to analyze observations repeatedly collected from multiple individuals in a population of interest (Lavielle, 2014). These models are defined by systems of differential equations whose parameters often carry clear biological meaning—such as the half-life of a drug, the infection rate of a virus, or the turnover of immune cells. As such, they enable a causal interpretation of the relationships between compartments and offer a predictive framework that can simulate system dynamics under varying conditions (Aalen et al., 2012; Prague et al., 2017). However, to ensure accurate predictions, it is essential to carefully identify all exogenous relationships with covariates that may influence the parameters and behavior of the model. The availability of high-dimensional data has increased significantly due to advances in data acquisition and analysis methods (Lowe et al., 2017), complicating the problem of mechanistic model building. High-dimensional variable selection, where the number of covariates exceeds the number of observations, is well-documented in standard regression models. However, methodological developments remain limited for its application in mechanistic models.

In a broad sense, the mechanistic model of interest belongs to the family of non-linear mixed effect models (NLME), where the non-linearity does not necessarily allow for a closed-form solution and is instead defined by an ordinary differential equation (ODE). The variability inherent in the data, captured by the model parameters, can be attributed to different sources (intraindividual, interindividual, and residual). Mixed-effect models on parameters of the ODE allow the study of N individuals exhibiting the same overall behavior. Inter-individual variability in the biological parameters can be partly explained by observed covariates through fixed effects, while residual unexplained variability is captured by individual-specific random effects. This decomposition enables the model to incorporate both structured variation, due to measurable factors (e.g., demographic covariates, biomarkers), and residual variation, reflecting subject-specific deviations. Model inference and selection are essential steps in identifying the most

appropriate model to best represent the data and make accurate predictions. In this article, inference methods are not considered a major problem since many inference methods exist (Clairon *et al.*, 2023a; Wang, 2007; Lavielle et Mentr , 2007). We mainly focus on the Stochastic Approximation version of the EM algorithm (Dempster *et al.*, 1977; Delyon *et al.*, 1999; Kuhn et Lavielle, 2005), with the selection of the “best” model among the available models performed using an extension of the Bayesian information criterion (BICc) that is well suited for NLME (Delattre *et al.*, 2014). Our main goal in this work is to propose a model-building strategy to determine which observed individual covariates have a fixed effect on the mechanistic model parameters when the set of available explanatory covariates of size p is high-dimensional ($N \ll p$).

Several approaches have been proposed in recent years to meet this challenge. The most common strategy involves a two-step process. First, the parameters of the NLME model are estimated. Then, a regression model is fitted to relate these parameters to the covariates using either a stepwise method (Lehr *et al.*, 2010) or Lasso (Ribbing *et al.*, 2007), including its variations to deal with problems such as multicollinearity (Haem *et al.*, 2017). Bertrand et Balding (2013) introduced an iterative approach in which the parameter estimates are updated after being related to covariates. However, their method relied on Empirical Bayes Estimates (EBEs), which are prone to shrinkage and sensitive to study design (Nguyen *et al.*, 2017). Yuan *et al.* (2021) proposed a correction method for shrinkage but it may fail in high-dimensional settings. An alternative to address these limitations is to integrate penalized regression into NLME models, which allows simultaneous selection of covariates and parameter estimation (Bertrand *et al.*, 2015). However, the penalized likelihoods in this approach do not incorporate structured penalties, which means that they neither induce structural sparsity nor control information criteria. As a more recent extension, Naveau *et al.* (2023) proposed a Bayesian Lasso approach that controls information criteria and combines spike-and-slab priors with the SAEM algorithm. However, this work focuses on non-linear mixed effect models with closed-form outcomes. In this article, we compare our original method with the performance of their method on the published example. Extending their approach to ODE-based NLME models remains computationally prohibitive, necessitating the development of alternative methods.

In ODE-based NLMEMs, methods for model selection such as SCM (Stepwise Covariate Modeling) (Svensson et Jonsson, 2022) and COSSAC (Conditional Sampling for Stepwise Approach based on Correlation tests) (Ayr al *et al.*, 2021) were proposed. SCM is a stepwise approach that systematically adds or removes covariates to optimize model fit, while COSSAC improves this process by considering correlation tests between covariates and parameter values. The SAMBA method (Stochastic Approximation for Model Building Algorithm) has been shown to be faster with similar performance (Prague et Lavielle, 2022). In the first step, SAMBA performs an efficient screening of covariates by establishing the linear link between covariates and parameters through stochastic approximation. In the second step, it refines the selection by focusing on covariates that have been shown to be significant, thus ensuring a robust final model after several iterations of the two-step process. However, none of these methods performs well in high-dimensional settings : the SCM or COSSAC methods are too time-consuming to be used, and the SAMBA method is based on a stepwise algorithm (Zhang, 2016) that performs poorly

in high-dimensional environments. Indeed, there is a growing risk of overfitting, multicollinearity, and the introduction of bias with the usual variable selection methods such as stepAIC. These issues emphasize the need for variable selection methods that balance model complexity and predictive power, such as penalized methods like Lasso (Tibshirani, 1996; Efron *et al.*, 2004) within existing model-building frameworks. The method presented in this article develops a selection method for covariates in high-dimensional settings for non-linear mixed effect models by replacing the stepwise algorithm for covariate selection in complex ODE-based SAMBA with Lasso-based selection. To improve the robustness of the method, we introduce a stability selection algorithm (Meinshausen et Bühlmann, 2010; Shah et Samworth, 2013). The stepwise algorithm originally used in SAMBA enforces a monotonic decrease in the information criterion across iterations, a property that Lasso does not guarantee on its own. Therefore, to control the information criterion in the new approach, we explore multiple covariate models stochastically to identify the one that minimizes the information criterion.

The paper is organized as follows. Section 2 introduces the statistical model based on a general ODE-based NLMEM and gives an overview of the SAMBA method, hereafter referred to as the step-SAMBA method, focusing on the aspects relevant to our approach. The new Lasso-SAMBA method is then presented. Simulation results are discussed in Section 3, and an application to a Varicella-Zoster Virus (VZV) vaccinology study (Li *et al.*, 2017), revealing biological associations between parameters of the humoral immune response mechanistic model and high-dimensional transcriptomic expressions, is proposed in Section 4. Finally, in Section 5, the potential of the method is highlighted and an outlook for future work is given.

2 Methods

2.1 Non-linear mixed effects model

We consider an outcome variable, eventually multivariate, $Y_i = (Y_{ij})_{j \leq n_i}$ observed at time points t_{ij} , with j indexing the occasion ($j = 1, \dots, n_i$) and i indexing the individual ($i = 1, \dots, N$). We assume that these observations come from an individual dynamic process $y_i = f(\cdot, \psi_i)$ that depends on time and on individual-specific model parameters $\psi_i = (\psi_{il})_{l \leq m}$. These dynamics may arise from a mechanistic model defined by an ODE with parameters ψ . Thus, our observations can be expressed as follows :

$$\forall i \leq N, j \leq n_i, \quad u(Y_{ij}) = u(f(t_{ij}, \psi_i)) + g(t_{ij}, \psi_i, \xi) \varepsilon_{ij},$$

where ε_{ij} are standardized Gaussian random variables with zero mean, and g is a function that possibly depends on the parameters ξ and defines the residual error model. This definition is valid up to a certain predetermined link function u , which can be applied to account for non-linearity or other systematic deviations in the residual structure.

In the context of ODE-based NLME models, it is assumed that the individual parameters

ψ_i are normally distributed, or up to a transformation, around a typical population parameter $\psi_{pop} \in \mathbb{R}^m$, with covariance matrix Ω . If individual i is additionally characterized by covariates $X_i = (X_{i1}, \dots, X_{ip})$, we can include covariate effects in the definition of the individual parameters, which leads to the following equation :

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

where $\beta = (\beta_1, \dots, \beta_m)$, and $\beta_l = (\beta_{l1}, \dots, \beta_{lp})^T$ represents the effect of the p covariates on parameter $l \leq m$, $\eta_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Omega)$, $i \leq N$, and h is a potential transformation function applied to the covariates or parameters.

Such models, denoted hereafter by $\mathcal{M}$, are typically estimated using the Stochastic Approximation Expectation-Maximization (SAEM) algorithm (Delyon *et al.*, 1999; Kuhn *et Lavielle*, 2005), a stochastic variant of the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm aims to maximize the observed-data log-likelihood by alternating between two steps :

- an Expectation step (E-step), which computes the conditional expectation of the complete-data log-likelihood given the current estimate of the parameters,
- a Maximization step (M-step), which updates the parameters by maximizing the expected complete-data log-likelihood obtained from the E-step.

However, for complex models such as ODE-based NLME models, this conditional expectation cannot be computed analytically due to latent random effects η_i . The SAEM algorithm addresses this limitation by introducing a stochastic approximation during the E-step, where individual parameters η_i are sampled from their conditional posterior distribution given the current estimate of population parameters, typically using Markov Chain Monte Carlo (MCMC) methods, thus enabling parameter estimation despite the absence of a closed-form likelihood.

2.2 The original step-SAMBA algorithm

To select covariates in an ODE-based NLME model, the step-SAMBA (Prague *et Lavielle*, 2022) method creates models iteratively, starting with an empty initial model $\mathcal{M}_0$ that contains no covariate effects or correlations, and typically assumes a combined error model $g(t, \psi, \xi = (a, b)) = a + b f(t, \psi)$. At each iteration k , the algorithm updates the model $\mathcal{M}_{k+1}$ by sequentially :

- estimating the parameters $\theta^{(k)} = (\psi_{pop}, \Omega, \beta, \xi)$ of $\mathcal{M}_k$ by maximum likelihood using the SAEM algorithm (Kuhn *et Lavielle*, 2005);
- sampling individual parameters $(\psi_i^{(k)})_{i \leq N}$ using MCMC methods;
- building three submodels with $\theta^{(k)}$ and $(\psi_i^{(k)})_{i \leq N}$: a correlation model, an error model, and a covariate model (denoted $\mathcal{M}^{\text{COV}}$).

These three submodels then form $\mathcal{M}_{k+1}$. A complete description of the method is available in [Prague et Lavielle \(2022\)](#).

Our main focus here is on the construction of the covariate part, $\mathcal{M}_{k+1}^{\text{COV}}$. Using the generated parameters $(\psi_i^{(k)})_{i \leq N} = (\psi_{i1}^{(k)}, \dots, \psi_{im}^{(k)})_{i \leq N}$, a regression model is built for each parameter ψ_l , where $l \leq m$:

$$h_l(\psi_{li}^{(k)}) = \mu_l + \mathbf{X}_i \boldsymbol{\beta}_l + \eta_{il}, \quad i \leq N, \quad (3.1)$$

where, for $l \leq m$, η_{il} are Gaussian errors corresponding to the random effects of individual i on parameter l , μ_l is the intercept of the regression model, and $\mathbf{X}_i$ represents the individual covariates. These are selected either by a stepwise selection procedure if there are more than 10 covariates, or by an exhaustive search among all possible models if there are fewer than 10 covariates.

The SAMBA method is an efficient approach to improve the model-building process in ODE-based NLMEMs, as demonstrated in simulation and real-world examples ([Prague et Lavielle, 2022](#)). We propose to replace the stepwise algorithm for selecting covariates in step-SAMBA with a refined Lasso-based selection method in order to improve the variable selection process in high-dimensional data ($N \ll p$).

2.3 The stability-enhanced Lasso-SAMBA algorithm

Similar to the step-SAMBA method, our approach performs variable selection independently for each parameter ψ_l in the ODE-based model. At each iteration $k \in \mathbb{N}$, the first two steps remain unchanged : the Lasso-SAMBA method first estimates the population parameters and samples the individual parameters. Then, in the Lasso-SAMBA method, we estimate coefficients in Equation (3.1) with a Lasso regression ([Tibshirani, 1996](#)) given a penalty parameter λ_l :

$$(\hat{\mu}_l^{\lambda_l}, \hat{\boldsymbol{\beta}}_l^{\lambda_l}) = \arg \min_{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i \leq N} \left(h(\psi_{li}^{(k)}) - \mu - \mathbf{X}_i \boldsymbol{\beta} \right)^2 + \lambda_l \|\boldsymbol{\beta}\|_1 \right\}. \quad (3.2)$$

For simplicity, we now omit the explicit subscript for each parameter l in the remainder of the paper ; for example, in Equation (3.2), $\lambda_l, \hat{\boldsymbol{\beta}}_l^{\lambda_l}$ are later denoted by $\lambda, \hat{\boldsymbol{\beta}}^\lambda$.

The main goal is to identify, for each parameter of the mechanistic model, the true set $\mathcal{S} = \{c \leq p; \beta_c \neq 0\}$ corresponding to the selection of truly relevant covariates. For a fixed λ , this set can be derived, for example, by :

$$\hat{\mathcal{S}}^\lambda = \{c \leq p \mid \hat{\beta}_c^\lambda \neq 0\}.$$

Usually in Lasso regression, the penalty parameter, and thus the selected set of covariates, is determined through cross-validation. However, this approach can result in unstable selections, particularly in high-dimensional settings ([Bühlmann et van de Geer, 2011](#); [Meinshausen et Bühlmann, 2010](#)). To address this problem, we rely on a Stability Selection Algorithm ([Meinshausen et Bühlmann, 2010](#); [Shah et Samworth, 2013](#)), a method developed to improve the performance

of variable selection algorithms. It has been proven that stability selection achieves consistency in variable selection (Meinshausen et Bühlmann, 2010). The central idea of the stability selection algorithm is to repeat the variable selection process a large number of times for a fixed value of λ on data samples that are half the size of the original dataset. A covariate $c \leq p$ is selected when its selection frequency across batches, denoted as $\hat{\pi}^\lambda(c)$, exceeds a given threshold $\pi \in [0, 1]$.

In our case, we draw uniformly n_{SS} random subsamples of individuals $(I_b)_{b \leq n_{SS}}$ from $\{1, \dots, N\}$, each of size $\lfloor N/2 \rfloor$. For each batch $b \leq n_{SS}$ of sample data $\mathcal{D}_b = (\mathbf{X}_i, h(\psi_i^{(k)}))_{i \in I_b}$, we denote $\hat{S}_b^\lambda = \{c \leq p, \hat{\beta}_c^{\lambda, b} \neq 0\}$ as the set of selected covariates resulting from the Lasso regression presented in Equation (3.2) in batch b with penalty parameter $\lambda > 0$. The final selection set for the hyperparameters (λ, π) is then :

$$\hat{S}^{\lambda, \pi} = \{c \leq p \mid \hat{\pi}^\lambda(c) \geq \pi\}, \quad (3.3)$$

$$\text{where } \hat{\pi}^\lambda(c) = \frac{\#\{b \leq n_{SS} \mid \hat{\beta}_c^{\lambda, b} \neq 0\}}{n_{SS}}.$$

Ultimately, this method is based on two critical hyperparameters for each parameter of the mechanistic model : $\lambda > 0$ and $\pi \in [0, 1]$.

2.4 Calibration of parameters (λ, π) and BICc

In the step-SAMBA method, the selection of covariates by a stepwise algorithm seems to ensure a reduction of the criterion used for model construction. Consequently, the models $\mathcal{M}_k$ constructed during the iterations of the algorithm have a BICc criterion that usually decreases as the iterations progress, allowing the algorithm to stop at a certain point. However, when using a Lasso method, the decrease of the BICc may be lost. This behavior appears to result from the decrease in the BIC (Bayesian Information Criterion) of the individual parameter model (3.1) during the stepwise covariate selection at each iteration of the step-SAMBA method. To preserve this property, we conduct an exploration step for selecting the covariate model $\mathcal{M}^{COV}$ during the covariate model building step of each Lasso-SAMBA iteration k . We explore a set $\Lambda \times \Pi$ of hyperparameters (λ, π) that results in models $\mathcal{M}^{COV}(\lambda, \pi)$, among which the one minimizing the BIC is selected as the candidate covariate model $\mathcal{M}^{COV}$. The covariate model is then chosen as $\mathcal{M}_{k+1}^{COV}$ if it has a lower BIC than the previous model $\mathcal{M}_k^{COV}$; otherwise, $\mathcal{M}_k^{COV}$ is retained. The schematic pipeline of the Lasso-SAMBA procedure is presented in Figure 3.2.

In other words, rather than identifying a single set of parameters (λ, π) , we aim to explore a space that yields several covariate models, searching for one that minimizes the BIC.

To characterize a good set $\Lambda \times \Pi$, we use the upper bound for the expected number of false discoveries derived by Meinshausen et Bühlmann (2010). If we denote by $V^{\lambda, \pi}$ the set of falsely selected variables in $\hat{S}^{\lambda, \pi}$, we have, for $\pi \in]0.5; 1]$:

$$\mathbb{E}[V^{\lambda, \pi}] \leq \frac{1}{2\pi - 1} \times \frac{q_\lambda^2}{p},$$

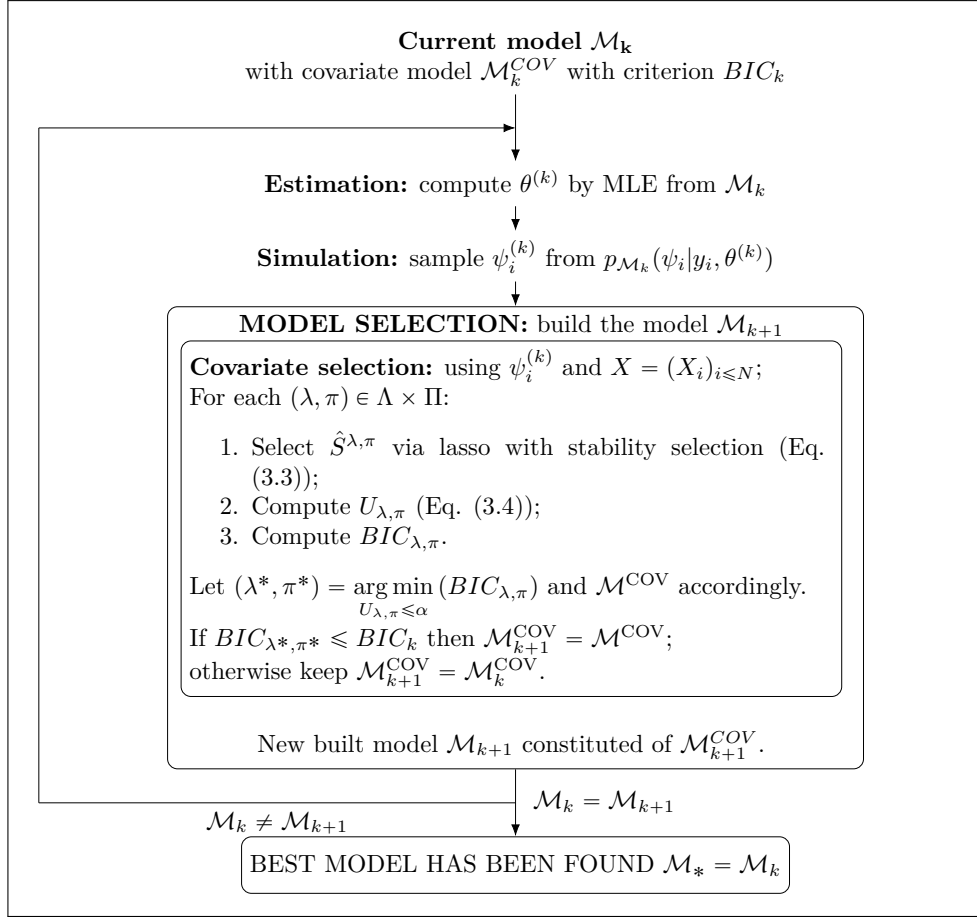


FIGURE 3.2 – Lasso-SAMBA method pipeline.

where $q_\lambda = \overline{|\hat{S}_b^\lambda|}^{b \leq n_{ss}}$ is the average number of features that are selected at least once by the Lasso selection with penalty λ . This upper bound for the expected number of false discoveries can then be transformed into an upper bound for the expected proportion of false discoveries, the False Discovery Rate (FDR) :

$$FDR(\lambda, \pi) = \mathbb{E}[FDP(\lambda, \pi)] \leq U_{\lambda, \pi} = \frac{1}{|\hat{S}_{\lambda, \pi}|} \times \frac{1}{2\pi - 1} \times \frac{q_\lambda^2}{p}. \quad (3.4)$$

Starting from a space of $\lambda \in \Lambda$, which ranges from dense to sparse models, as is usually done (Friedman *et al.*, 2010b), and $\Pi =]0.5; 1]$, we compute $\hat{S}^{\lambda, \pi}$ for each $\lambda \in \Lambda, \pi \in \Pi$ and the induced upper bounds $U_{\lambda, \pi}$. We first remove from $\Lambda \times \Pi$ any set of parameters that does not fulfill the condition $U_{\lambda, \pi} \leq \alpha$ for a fixed hyperparameter $\alpha \in [0; 1]$. For the remaining set of parameters $\Lambda \times \Pi \setminus \{(\lambda, \pi), U_{\lambda, \pi} > \alpha\}$, we choose the final set of selected covariates such that :

$$\hat{S} = \arg \min_{\substack{(\lambda, \pi) \in \Lambda \times \Pi, \\ U_{\lambda, \pi} \leq \alpha}} BIC_{\lambda, \pi},$$

where $BIC_{\lambda, \pi}$ is the information criterion of the linear model in Equation (3.1) with $\hat{S}^{\lambda, \pi}$ as the selected covariates.

2.5 Implementation

The new method is implemented in the LSAMBA R (R Core Team, 2024) package, available on CRAN, using the Rsmlx (Prague et Lavielle, 2022; Mihaljevic, 2023) and lixoftConnectors (LIXOFT, 2023) packages as a base. For the Lasso estimation in Lasso-SAMBA, we use the sharp package (Bodinier, 2024), which implements stability-based variable selection. All code was run with Lixoft Monolix 2023R1.

3 Simulation studies

In this section, we evaluate the performance of the Lasso-SAMBA method in terms of covariate selection through simulation studies. We conducted a simulation study under two distinct frameworks to assess the performance of the proposed method in sparse settings. The aim was to compare the newly developed procedure with step-SAMBA (Prague et Lavielle, 2022) and SAEMVS (Naveau *et al.*, 2023) in terms of variable selection accuracy. The first has been described above in Section 2.2, while the latter consists of iteratively applying a SAEM procedure, where the E-step estimates posterior inclusion probabilities of variables using Monte Carlo sampling, and the M-step updates model parameters by maximizing a penalized likelihood, effectively shrinking irrelevant coefficients toward zero for sparse model recovery. We calibrated both reference methods using the hyperparameters and initialization strategies detailed in the original studies. For our Lasso-SAMBA, the only hyperparameter requiring specification was the constraint on the expected number of false discoveries, which we fixed at 10%. A sensitivity analysis over a range of values is reported in Supplementary Material Section A.3. Data were generated using biologically motivated models presented in each subsection : a toy pharmacokinetics model (Section 3.2) and a humoral immune response model (Section 3.3). For each framework, 100 independent replicates were generated.

3.1 Performance evaluation

In the context of covariate selection, a true positive (TP) is a covariate truly associated with the parameter and correctly identified as such ; a false positive (FP) is not associated but incorrectly selected ; a false negative (FN) is truly associated but missed ; and a true negative (TN) is not associated and correctly excluded. The overarching goal is to minimize false discoveries. In particular, for our method, the aim is to reduce the False Discovery Rate (FDR) without significantly increasing the False Negative Rate (FNR), defined as :

$$FDR = \frac{FP}{FP + TP},$$

$$FNR = \frac{FN}{FN + TP}.$$

To assess the performance of the model, we compare both the FDR and FNR, while also considering the F_1 -score, which provides a single metric that balances the trade-off between FP and FN :

$$F_1 = \frac{2TP}{2TP + FN + FP}.$$

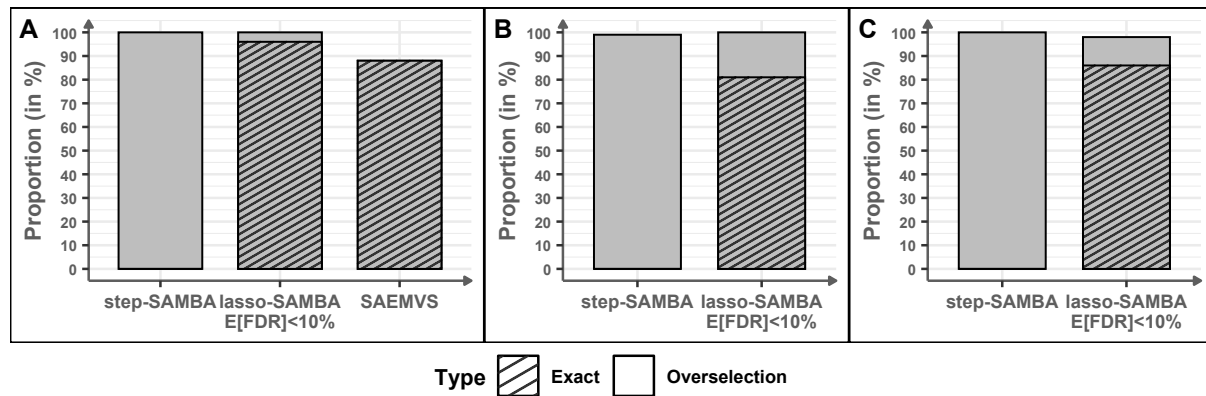


FIGURE 3.3 – Proportion of exact models (no false positives or false negatives) and models that strictly include the exact model (no false negatives but with false positives). (A)

Pharmacokinetic model with categorical covariates ; (B) Vaccinology framework with Gaussian-correlated covariates ; (C) Vaccinology framework with randomly drawn covariates.

Metric	PK model binomial covariates	Vaccinology model Gaussian covariates	Vaccinology model non-Gaussian covariates
FDR (%) :			
- step-SAMBA	87.5 [78.8;91.7]	72.7 [50.0;83.8]	70.0 [25.0;81.2]
- Lasso-SAMBA	0.0 [0.0;14.3]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
FNR (%) :			
- step-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- Lasso-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
F_1-score (%) :			
- step-SAMBA	22.2 [15.4;34.9]	42.9 [27.9;66.7]	46.2 [31.6;85.7]
- Lasso-SAMBA	100.0 [92.3;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- SAEMVS	100.0 [0.0;100.0]	ND	ND

TABLE 3.1 – False Discovery Rate (FDR), False Negative Rate (FNR), and F_1 -score computed for each simulation framework, comparing the original step-SAMBA method and the Lasso-SAMBA method, with an error control threshold of 10%. (ND : not defined).

3.2 A toy pharmacokinetic model

We replicate the simulation study provided in [Naveau et al. \(2023\)](#). We simulate a one-compartment pharmacokinetic model with linear elimination. The model describes the concentration y_{ij} of a drug, in the effect compartment, measured at 12 time points t_j , $j = 1, \dots, 12$, for

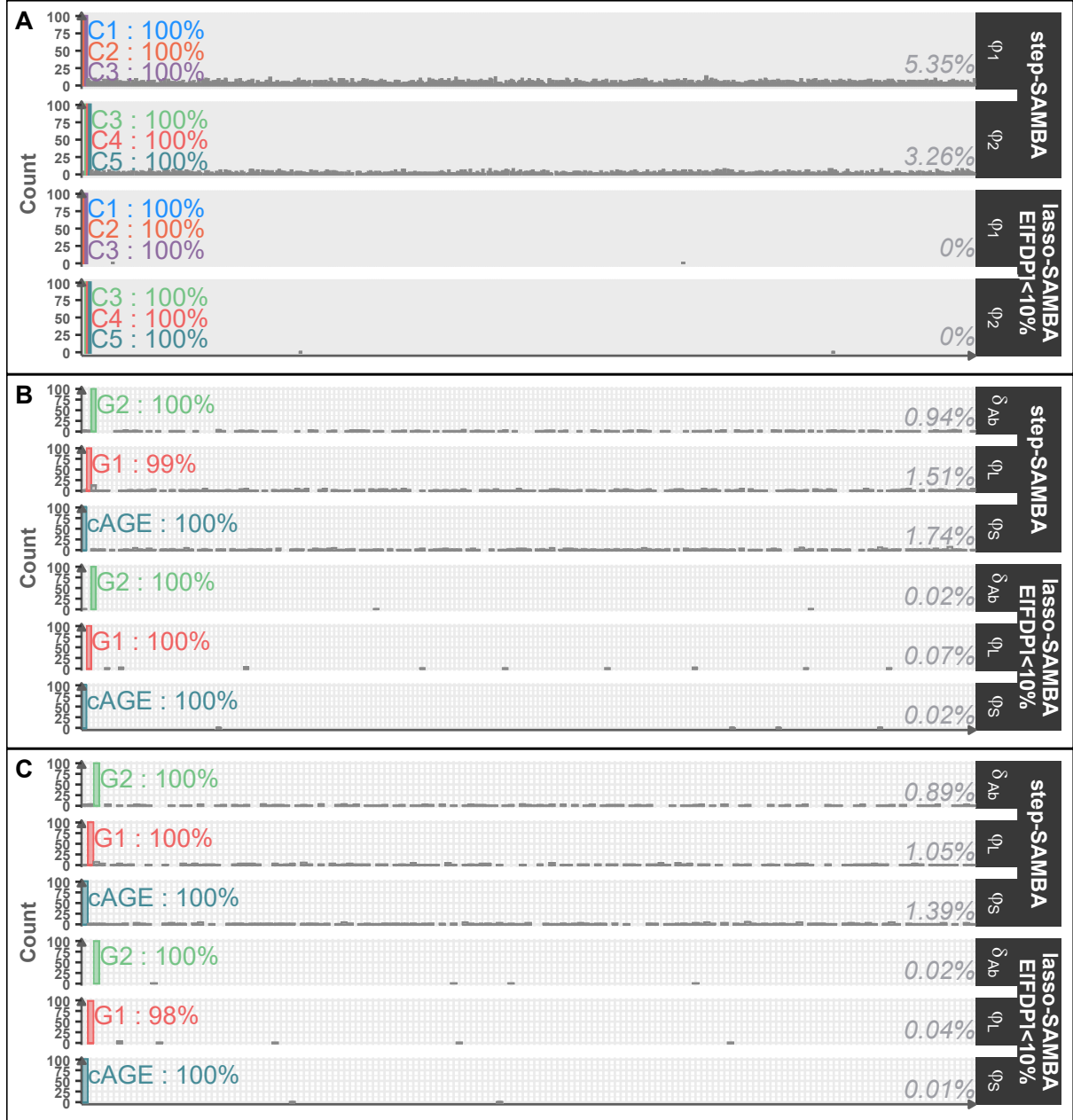


FIGURE 3.4 – Covariate selection frequency across simulation frameworks for the step-SAMBA method and the Lasso-SAMBA method, with an error control threshold of 10%. The mean selection frequency of false discoveries is shown on the right side of each histogram. (A) Pharmacokinetic model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

individual i with $i = 1, \dots, 200$:

$$y_{ij} = \frac{D\varphi_{1i}}{V\varphi_{1i} - \varphi_{2i}} \times \left(e^{-\varphi_{2i} \frac{t}{V}} - e^{-\varphi_{1i} t} \right) + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

and

$$\begin{cases} \varphi_{1i} &= \varphi_{1pop} + \beta_1^T C_i + \eta_i^1 \\ \varphi_{2i} &= \varphi_{2pop} + \beta_2^T C_i + \eta_i^2 \end{cases},$$

where $\boldsymbol{\eta}_i = (\eta_i^1, \eta_i^2) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{\Omega})$ are random effects explaining individual residual heterogeneity, and σ^2 is the variance of measurement error. Time points for all individuals are $t_j \in \{0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40\}$. The administered dose $D = 100$ and the apparent volume of distribution $V = 30$ are known constants. The individual parameters are the individual-specific rate φ_1 and φ_2 . These parameters are to be estimated and related to individual-specific covariates.

For each individual i , the covariate matrix $\mathbf{C}_i \in \mathbb{R}^{500}$ is simulated from independent binomial distributions with success probability 0.2. The covariates are standardized. The effect parameters (β_1, β_2) are such that covariates 1–3 are associated with φ_{1i} , and covariates 3–5 are associated with φ_{2i} , leaving 495 covariates not associated with any parameter. The parameter values used are detailed in Supplementary Material Section A.1.

In this initial simulation scenario, designed to evaluate variable selection under controlled conditions, we allow both correlation and covariate model building through the Lasso-SAMBA and step-SAMBA methods in order to compare results with the SAEMVS algorithm. Here, we observe that the Lasso-SAMBA method markedly outperforms step-SAMBA with respect to FDR control. The median FDR achieved by step-SAMBA was 88% (95% CI : 79%, 92%), while Lasso-SAMBA reduced this to a striking 0.0% (95% CI : 0.0%, 14.3%), as summarized in Table 3.1. This considerable improvement in FDR does not come at the expense of reduced power.

Indeed, as illustrated in Figure 3.4, panel A, which plots covariate selection frequencies across 100 replicated datasets, Lasso-SAMBA retains the same number of true positive selections as step-SAMBA, demonstrating that the improved FDR is achieved without sacrificing sensitivity. Consequently, the F_1 -score shows a substantial increase : from 22% (95% CI : 15%, 35%) with step-SAMBA to 100% (95% CI : 92%, 100%) with Lasso-SAMBA.

Further evaluation of model selection quality is presented in Figure 3.3, panel A. This figure shows the proportion of exact model recoveries and models that strictly include the correct covariates. The step-SAMBA method fails to recover the exact model in any replicate. In contrast, Lasso-SAMBA successfully recovers the exact model structure in 96% of simulations. Notably, neither method yields final models with false negative covariates, indicating consistent inclusion of all relevant associations.

We also compare the performance of Lasso-SAMBA with the SAEMVS method. In this context, SAEMVS achieves exact model recovery in 88% of replicates. However, unlike Lasso-SAMBA, which maintains zero false negatives across all replicates, SAEMVS occasionally omits relevant variables. Table 3.1 also indicates substantial variability in performance, as the F_1 -score can be as low as 0% in some replicates.

3.3 Simulation of humoral immune response

To emulate a realistic biological process relevant to our real-data application (Pasin *et al.*, 2019), we simulated antibody kinetics for 100 individuals based on a mechanistic model incorporating two types of antibody-secreting cells (ASCs) : short-lived (S) and long-lived (L), which secrete antibodies (Ab) at rates φ_S and φ_L , respectively, and are characterized by their distinct half-lives δ_S and δ_L . Antibody concentrations decay over time according to a rate parameter δ_{Ab} . The dynamics are driven by an ordinary differential equation as follows :

$$\frac{d}{dt}Ab_i(t_j) = \varphi_{Si}e^{-\delta_S t_j} + \varphi_{Li}e^{-\delta_L t_j} - \delta_{Abi}Ab_i(t_j),$$

with individual parameters

$$\begin{cases} \log(\varphi_{Si}) &= \log(\varphi_{Spop}) + \beta_{AGE} \times cAGE_i + \eta_i^S, \\ \log(\varphi_{Li}) &= \log(\varphi_{Lpop}) + \beta_{G_1} \times G_{1i} + \eta_i^L, \\ \log(\delta_{Abi}) &= \log(\delta_{Abpop}) + \beta_{G_2} \times G_{2i} + \eta_i^{Ab}, \end{cases}$$

with $\eta_i^S \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \omega_S^2)$, $\eta_i^L \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \omega_L^2)$ and $\eta_i^{Ab} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \omega_{Ab}^2)$, and $(\beta_{AGE}, \beta_{G_1}, \beta_{G_2})$ the effect parameters. For each individual, the centered age $cAGE$ follows a normal distribution with variance 16, and G_1 and G_2 are drawn from standardized normal distributions. The process is observed at times $t_j \in \{0, 7, 21, 123, 180, 300\}$.

To simulate a high-dimensional setting, 200 covariates were generated by adding 197 noise variables to the 3 informative ones. Two data generation scenarios were considered to test the robustness of the methods : one involving purely Gaussian covariates and another with a heterogeneous mix of randomly drawn distributions, including both categorical and continuous variables. In both cases, correlation was introduced among the covariates to mimic realistic dependencies observed in biomedical datasets. Full details of the covariate generation process are provided in Supplementary Material Section A.1.

In the second simulation framework, covariate model building was performed using both Lasso-SAMBA and step-SAMBA. The Lasso-SAMBA method again demonstrates a marked improvement in FDR over the step-SAMBA method. The median FDR decreases from 73% (95% CI : 50%, 84%) to 0.0% (95% CI : 0.0%, 25.0%), as shown in Table 3.1. Figure 3.4, panel B, confirms that this improvement in specificity is not achieved at the cost of sensitivity, as the number of true discoveries remains comparable between the two methods. In turn, the F_1 -score is drastically increased under Lasso-SAMBA, reaching a median of 100% (95% CI : 86%, 100%). Moreover, the 95% confidence intervals for both FDR and F_1 -score do not overlap between the methods. Figure 3.3, panel B, further illustrates that, whereas the step-SAMBA method fails to recover the exact model in any replicate, Lasso-SAMBA identifies the correct model structure in 81% of cases.

To assess the robustness of the method to deviations from classical assumptions, a third simulation setting was implemented where covariates were randomly drawn from heterogeneous

distributions, including categorical and non-Gaussian continuous variables. Even under these more complex and less idealized conditions, Lasso-SAMBA continued to substantially outperform the benchmark method. The median FDR dropped from 70% (95% CI : 25%, 81%) to 0.0% (95% CI : 0%, 25%), as summarized in Table 3.1. Figure 3.4, panel C, indicates that the number of true discoveries remains largely unaffected, while the median F_1 -score increases from 46% (95% CI : 31%, 85%) to 100% (95% CI : 88%, 100%). Model selection quality is again improved : as shown in Figure 3.3, panel C, step-SAMBA fails to identify the true model in any replicate, while Lasso-SAMBA achieves exact recovery in 86% of cases.

Finally, a comparative analysis of computational times across all simulation frameworks reveals that Lasso-SAMBA is generally faster. Although individual iterations of Lasso-SAMBA tend to be slower due to the optimization procedure, the algorithm converges in fewer steps compared to step-SAMBA. In the two more complex vaccinology-inspired frameworks, Lasso-SAMBA required a median of two iterations to converge, versus four for step-SAMBA. These runtime differences are visualized in Supplementary Material Section A.3.

4 Application to VZV vaccination data

To illustrate the utility of our approach in a real-world setting, we applied the proposed methodology to a publicly available dataset from a clinical study of immune response to vaccination against the Varicella-Zoster virus (VZV). This dataset, accessible under the accession number SDY984 on the ImmPort platform (Bhattacharya *et al.*, 2018b), stems from the study titled "Zoster vaccine in young and elderly," and includes gene expression and antibody response data following immunization with ZOSTAVAX, a live attenuated vaccine targeting VZV (Li *et al.*, 2017).

ZOSTAVAX (Arvin, 1996) is employed both for the primary prevention of varicella-a childhood illness marked by viremia, fever, and vesicular skin lesions-and for the prevention of herpes zoster, a reactivation of latent VZV that manifests as a localized dermatomal rash. In this study, 35 adult volunteers were administered the vaccine, with longitudinal sampling of whole blood at baseline and days 1, 3, 7, 14, 30, 90, and 180 post-vaccination. Transcriptomic data were obtained at days 0, 1, 3, and 7, while antibody levels were measured at days 0, 7, 14, 30, 90, and 180. Of the 10,086 profiled genes, a subset of 784 protein-coding genes was selected based on functional annotation with roles in interferon signaling, type I interferon response, neutrophil activation, inflammation, cytokine/chemokine activity, and cell cycle regulation, as defined by Biomart (Durinck *et al.*, 2005) and Li *et al.* (2014).

To characterize antibody kinetics post-vaccination, we utilized the humoral immune response model proposed by Pasin *et al.* (2019), which was studied in simulations in Section 3.3 and is detailed in Supplementary Material Section A.2. The analysis specifically explored which early gene expression signatures are predictive of early humoral responses, focusing on the first day post-vaccination-a time point previously shown to exhibit maximal transcriptomic variation. We

investigated the relationship between these genes and the parameters φ_S , φ_L , and δ_S using the step-SAMBA and Lasso-SAMBA methods.

The gene associations identified by the step-SAMBA and Lasso-SAMBA methods are presented in Table 3.2. Step-SAMBA identified 15 genes, while the Lasso-SAMBA method yielded a sparse model in which LEP and KIFC1 were uniquely associated with the model parameters φ_S and φ_L , respectively. The model resulting from the model-building procedure for parameters φ_S and φ_L respectively follows :

$$\log(\varphi_{S_i}) = \log(\varphi_{S_{pop}}) + \beta_{LEP} \times LEP_i + \eta_i^S,$$

and

$$\log(\varphi_{L_i}) = \log(\varphi_{L_{pop}}) + \beta_{KIFC1} \times KIFC1_i + \eta_i^L.$$

The final model estimates, incorporating these genes, are described in detail in Supplementary Material Section B. These findings implicate LEP and KIFC1 as candidate biomarkers of antibody-secreting cell dynamics during early vaccine-induced immunity. LEP and KIFC1 both belong to the cell cycle pathway. While the first encodes the leptin hormone, the second encodes kinesin family member C1. Their potential role in immunity is discussed in Section 5.

To evaluate model specificity and control for spurious associations, we conducted model selection using a null covariate dataset without any biological signal. This artificial dataset was generated by fitting a multivariate Gaussian distribution to the empirical gene expression values of the 784 preselected genes. Under these conditions, the Lasso-SAMBA method did not identify any significant associations with the antibody model parameters, supporting its robustness to spurious associations. In contrast, the step-SAMBA method produced six false-positive associations—one involving φ_S and five involving φ_L .

Further assessment of model stability was performed via nonparametric bootstrap analysis using 500 resampled datasets (Figure 3.5). The Lasso-SAMBA method consistently identified LEP and KIFC1 as the most stable covariates, with selection frequencies of 25.2% and 22.6%, respectively. The step-SAMBA method also frequently selected LEP (45.8%) and KIFC1 (46.2%) but exhibited greater model variability; no other genes surpassed a selection frequency of 12.8%. Additionally, step-SAMBA failed to converge in 7 out of 500 iterations, primarily due to overfitting-related estimation issues.

Together, these findings underscore the enhanced robustness and reproducibility of the Lasso-SAMBA method in identifying biologically plausible gene–parameter associations in the context of humoral immune modeling.

5 Discussion

Model selection remains a central challenge in the development of reliable statistical models, particularly within the context of personalized medicine, where the accurate identification

Parameter	Genes selected	
	step-SAMBA	Lasso-SAMBA
φ_S	ASPM, CDC45, LEP , NOTCH1, SP110	LEP
φ_L	GDPD5, IKZF2, KIFC1 , PSORS1C1, SLC16A5, SLC50A1, SRP72	KIFC1
δ_S	ISTNA1, RAB2A, SLC31A1	

TABLE 3.2 – Results from the step-SAMBA and Lasso-SAMBA methods for the ZOSTAVAX vaccine. The initial model had an empty covariate structure, a constant error model, and no correlation. Both methods started from the same initial conditions, with covariate model selection being the only difference.

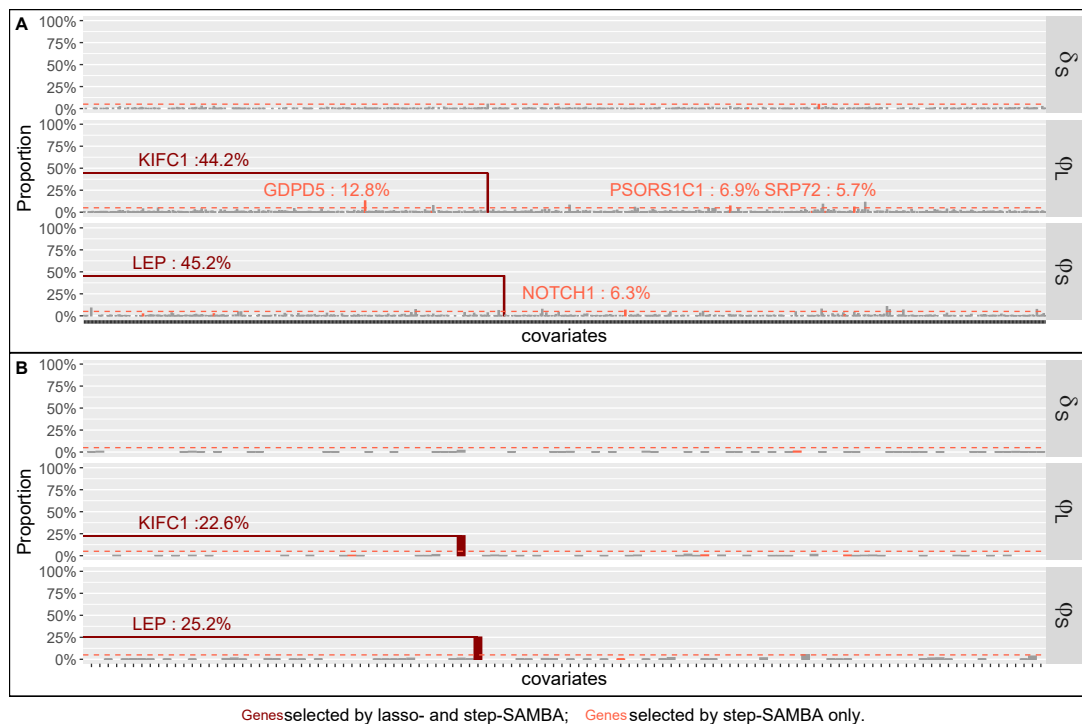


FIGURE 3.5 – Bootstrap selection frequencies for covariates across all parameter models. The top panel shows results for the step-SAMBA method (493 successful runs), and the bottom panel for the Lasso-SAMBA method (500 runs). Selection frequencies above 5% of selected genes by any of the procedures are indicated.

of individual-specific covariates can substantially influence therapeutic decision-making. In this paper, we proposed a method for selecting covariates in high-dimensional settings by integrating a Lasso approach for covariate selection within the framework of non-linear mixed effects models described by ordinary differential equations. Although the method was presented for a univariate longitudinal outcome variable, it also applies to any ODE-based model with multivariate longitudinal data.

The motivation for this enhancement stems from the limitations of traditional stepwise selection procedures as implemented in step-SAMBA which, although computationally tractable, are known to be susceptible to overfitting and the inclusion of non-informative covariates, especially in high-dimensional settings (Heinze *et al.*, 2018). By integrating the Lasso, coupled with a stability selection procedure, we aimed to improve the selection of relevant covariates

while preserving computational feasibility, in contrast to the approach by Naveau *et al.* (2023).

Our simulation studies demonstrated a marked improvement in the control of FDR compared to other procedures, without compromising FNR. This improvement was consistent across diverse pharmacokinetic and vaccinology-inspired model scenarios. These findings reinforce prior conclusions on the robustness of penalized regression approaches in complex hierarchical models (Caillebotte *et al.*, 2024).

Furthermore, in the real-data application involving antibody response to Varicella-Zoster virus vaccination, the Lasso-SAMBA approach exhibited superior performance in identifying biologically plausible biomarkers while producing a sparser covariate selection. This is consistent with the detection of true signals observed in our simulation studies. The LEP gene participates in several common biological pathways involved in the immune response, such as JAK-STAT or NF- κ B. KIFC1 is involved in cell proliferation and therefore is not specific to the immune response. These biomarkers may, however, reflect the involvement of broader pathways in the establishment of the vaccine response (Hagan *et al.*, 2022). KIFC1 also has the ability to promote stable mitotic spindle formation during early B cell development, where centriole duplication is frequent but must be tightly regulated (Schapfl *et al.*, 2024). These results suggest that our method improves the reliability of gene-covariate relationships in biomedical research involving ODE-based modeling.

The methodological integration we propose holds promise for broader applications, particularly in fields such as pharmacogenomics and systems pharmacology, where model interpretability and variable selection are of paramount importance. Future investigations could extend this approach to non-linear relationships between covariates and parameters.

Data and code availability

Code for running the simulations and applications is available at <https://github.com/aurianegbt/LassoSAMBA>.

3.3 Travaux additionnels : Analyse des données de l'essai clinique PREVAC-UP

Nous présentons à présent une application de la méthode Lasso-SAMBA réalisée sur les données recueillies lors de l'étude PREVAC-UP du consortium PREVAC (Badio *et al.*, 2021), présentée dans la Section 1.2.3.1. Cette étude vise à évaluer la sécurité et la durabilité de la réponse immunitaire pour trois stratégies vaccinales contre Ebola, notamment la stratégie rVSV-ZEBOV non boosté. Au total, 1400 adultes et 1401 enfants de Guinée, du Libéria, du Mali et de Sierra Leone ont reçu une dose de vaccin rVSV-ZEBOV contre le virus Ebola puis ont été suivis dans le temps : pour chaque individu, des mesures de concentration en anticorps, illustrées

en Figure 1.6, ont été effectuées le jour de la première dose, puis environ 7, 14, 28 et 56 jours après. Une seconde dose est administrée 56 jours après la première. Les participants ont ensuite été suivis au jour 63, puis à 3, 6 et 12 mois après l'inclusion. Une sous-étude immunologique, menée spécifiquement en Guinée, a inclus 95 participants adultes, dont 30 ont reçu le schéma rVSV-ZEBOV non boosté. Pour chacun d'eux, en plus des mesures d'anticorps, l'expression génique de plus de 28 000 gènes a été mesurée au moment du rappel, au jour 56. Notre objectif ici est de sélectionner, parmi ces gènes, ceux qui permettent de prédire la réponse immunitaire humorale au vaccin.

Nous présentons ci-dessous l'analyse réalisée sur les données des 26 patients de la sous-étude immunologique ayant reçu le bras vaccinal rVSV-ZEBOV non boosté, c'est-à-dire une dose de rVSV-ZEBOV au jour 0 puis une dose de rappel placebo au jour 56. Nous analysons les données d'anticorps mesurées à partir de la visite du jour 28, moment à partir duquel on suppose que les populations de cellules sécrétrices d'anticorps ne font plus que décroître. On peut alors modéliser la dynamique de ces cellules par une décroissance exponentielle. Nous modélisons la production d'anticorps en considérant une population de cellules sécrétrices d'anticorps, notée S , caractérisée par un taux de décroissance δ_S , les données ne permettant pas de différencier plusieurs sous-populations de cellules. Nous supposons que ces plasmocytes produisent des anticorps Ab à un taux θ_S . Les anticorps sont éliminés avec un taux de décroissance δ_{Ab} . Nous considérons ici que les paramètres θ_S et δ_{Ab} varient selon les individus et suivent une distribution log-normale.

Nous cherchons à identifier quels gènes (mesurés au jour 56, premier point de mesure disponible pour les données génétiques) sont associés à la réponse humorale parmi ceux impliqués dans les voies immunitaires des lymphocytes T et B, et des CPAs d'après la classification de Li *et al.* (2014) et de BioMart (Durinck *et al.*, 2005) (soient 757 gènes au total dans nos données). La réponse immunitaire humorale est modélisée par un modèle non-linéaire à effets mixtes (Andraud *et al.*, 2012; Pasin *et al.*, 2019), présenté en Section 2.3.4. Plus formellement, pour chaque individu $i = 1, \dots, 26$, le modèle utilisé est :

- **Modèle Structurel :**

$$\frac{dAb_i}{dt} = \theta_{Si} e^{-\delta_S t} - \delta_{Abi} Ab_i, \quad i = 1, \dots, N,$$

- **Modèle Statistique :**

$$\begin{cases} \log(\theta_{Si}) = \log(\theta_{S0}) + \boldsymbol{\beta}_S \mathbf{X}_i + \eta_{Si}, \\ \log(\delta_{Abi}) = \log(\delta_{Ab0}) + \boldsymbol{\beta}_{Ab} \mathbf{X}_i + \eta_{Abi}, \end{cases}$$

avec $\eta_{Si} \sim \mathcal{N}(0, \omega_S^2)$ i.i.d. et $\eta_{Abi} \sim \mathcal{N}(0, \omega_{Ab}^2)$ i.i.d, $\mathbf{X}_i$ étant le vecteur des 757 gènes considérés pour l'individu i .

- **Modèle d'Observation :**

$$Y_{ij} = \log_{10}(Ab_i(t_{ij})) + \varepsilon_{ij},$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{Ab}^2)$, pour l'individu i et la j^{eme} observation.

Les paramètres à estimer sont donc $(\theta_{S0}, \delta_{Ab0}, \delta_S, \beta_S, \beta_{Ab}, \omega_S^2, \omega_{Ab}^2, \sigma_{Ab}^2)$. Nous nous intéresserons en particulier aux composantes des vecteurs de coefficients β_S et β_{Ab} qui seraient non nulles, c'est-à-dire aux gènes qui influencent significativement la production ou la décroissance des anticorps et qui pourraient ainsi servir de prédicteurs de la réponse vaccinale.

Afin d'estimer le taux de décroissance des plasmocytes δ_S tout en assurant l'identifiabilité du modèle, nous avons réalisé un profil de log-vraisemblance sur ce paramètre. Plus précisément, pour des valeurs possibles de la demi-vie des cellules S $\log(2)/\delta_S$ allant de 1 an à 15 ans, nous avons fixé δ_S à la valeur correspondante puis estimé les autres paramètres du modèle par l'algorithme SAEM, via le logiciel Monolix. La log-vraisemblance du modèle a été calculée pour chaque valeur testée. Le profil de log-vraisemblance ainsi obtenu, donné en Figure 3.6, présente un maximum qui correspond à $\delta_S \approx 0,26 \times 10^{-3} \text{ jour}^{-1}$, soit une demi-vie d'environ 7,3 ans. Nous avons donc fixé δ_S à cette valeur pour la suite de l'analyse.

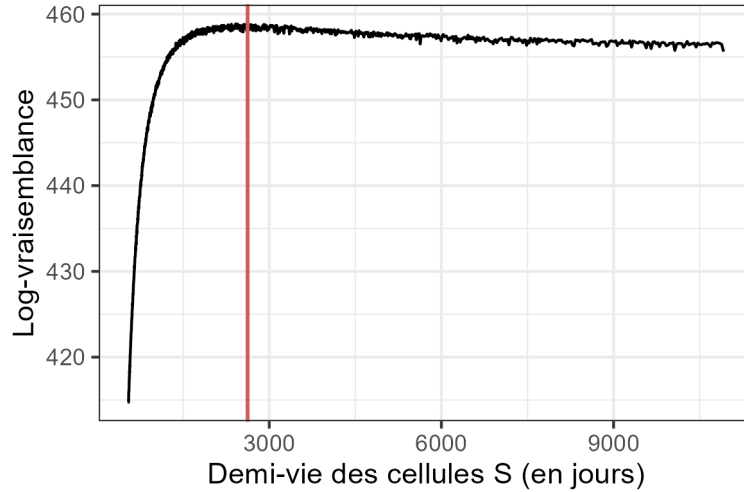


FIGURE 3.6 – Profil de log-vraisemblance en fonction de la demi-vie des cellules S. La ligne rouge indique le maximum atteint en 7.3 ans de demi-vie.

Dans un second temps, afin d'initialiser l'algorithme Lasso-SAMBA, nous utilisons le modèle $\mathcal{M}^{(0)}$ dit vide, sans covariables, avec $\beta_S = \mathbf{0}$ et $\beta_{Ab} = \mathbf{0}$, pour lequel nous avons exploité l'ensemble des données d'anticorps disponibles dans l'étude PREVAC pour les patients ayant reçu le vaccin rVSV-ZEBOV, soit 395 individus. L'utilisation de ce grand échantillon permet d'obtenir des estimateurs plus précis des paramètres initiaux du modèle, améliorant ainsi la convergence de l'algorithme sur les 26 individus de la sous-étude immunologique ensuite. Nous avons donc estimé les paramètres du modèle sans covariables à l'aide de l'algorithme SAEM sur ces 395 sujets. Les valeurs de paramètres obtenues sont présentées dans le Tableau 3.3.

Paramètre	Définition	Valeur [IC95%]	estimée
Effets fixes			
<i>Taux de production d'anticorps</i>			
$\theta_{S,0}$	Taux de production d'anticorps par les cellules S (Eu/mL/jour)	22,29	[16,15 ; 28,43]
<i>Taux de décroissance</i>			
$\delta_{Ab,0}$	Taux de décroissance des anticorps Ab (1/jour)	0,026	[0,023 ; 0,029]
δ_S	Taux de décroissance des cellules S	$0,26 \times 10^{-3}$	(fixé)
Effets aléatoires			
ω_S	Écart-type (log) de $\theta_{S,i}$	0,66	[0,59 ; 0,73]
ω_{Ab}	Écart-type (log) de $\delta_{Ab,i}$	0,45	[0,36 ; 0,54]
Erreur d'observation			
σ_{Ab}	Écart-type de l'erreur ($\log_{10}$) sur les niveaux d'anticorps	0,150	[0,145 ; 0,155]

TABLE 3.3 – Paramètres estimés (modèle sans covariables) pour le modèle de réponse humorale à partir du jour 28, sur les données de l'essai PREVAC (395 individus ayant reçu une dose de vaccin rVSV-ZEBOV).

IC95% : intervalle de confiance à 95%.

À partir de ces valeurs initiales, nous avons appliqué l'algorithme Lasso-SAMBA présenté précédemment sur les 26 individus de la sous-étude immunologique, avec l'hyperparamètre donnant le taux de contrôle de faux positifs à chaque itération à 10%. Au terme de l'algorithme, un seul gène FAM216A a été retenu dans le modèle final, avec un effet significatif sur le taux de décroissance des anticorps. Le Tableau 3.4 donne les valeurs des paramètres estimés par SAEM pour ce modèle final intégrant FAM216A comme covariable sur δ_{Ab} .

Paramètre	Définition	Valeur [IC95%]	estimée
Effets fixes			
<i>Taux de production d'anticorps</i>			
$\theta_{S,0}$	Taux de production d'anticorps par les cellules S (Eu/mL/jour)	26,03	[16,15 ; 35,91]
<i>Taux de décroissance des anticorps</i>			
$\delta_{Ab,0}$	Taux de décroissance des anticorps Ab (1/jour)	2,96	[0 ; 10,04]
$\beta_{Ab,fam216a}$	Effet du gène FAM216A sur $\log(\delta_{Ab,i})$	-1,82	[-2,76 ; -0,88]
δ_S	Taux de décroissance des cellules S	$0,26 \times 10^{-3}$	(fixé)
Effets aléatoires			
ω_S	Écart-type (log) de $\theta_{S,i}$	0,28	[0,06 ; 0,55]
ω_{Ab}	Écart-type (log) de $\delta_{Ab,i}$	0,40	[0,22 ; 0,58]
Erreur d'observation			
σ_{Ab}	Écart-type de l'erreur ($\log_{10}$) sur les niveaux d'anticorps	0,098	[0,086 ; 0,110]

TABLE 3.4 – Paramètres estimés pour le modèle final obtenu par Lasso-SAMBA de la réponse humorale (données à partir du jour 28) sur les 26 individus ayant reçu le vaccin rVSV-ZEBOV (une dose unique).

Le gène FAM216A est inclus comme covariable influençant δ_{Ab} .

IC95% : intervalle de confiance à 95%.

Nous avons évalué la stabilité de ce résultat par une analyse de *bootstrap* non paramétrique. Nous avons généré 1000 jeux de données *bootstrap* à partir des 26 individus initiaux, et appliqué la procédure de sélection sur chacun. Au total, 44% des modèles finaux obtenus ne sont pas vides, c'est-à-dire comportent au moins un gène sélectionné comme covariable. Parmi ceux-ci, dans 28% des cas le gène FAM216A fait partie des gènes sélectionnés, et dans 17% des cas FAM216A est même le seul gène retenu par le modèle.

L'ensemble de ces résultats met en lumière un effet potentiel du gène FAM216A sur la cinétique de décroissance des anticorps : l'expression plus élevée de ce gène est associée à une décroissance plus lente des anticorps après vaccination, et donc à une persistance plus longue des anticorps circulants. Dans notre échantillon, la valeur médiane de l'expression de FAM216A est de 2,544 (avec un intervalle interquartile de 2,390 à 2,626). Concrètement, pour un individu dont l'expression de FAM216A est égale au premier quartile, la demi-vie estimée des anticorps est d'environ 18,5 jours, contre 25,4 jours pour un individu dont l'expression atteint le troisième quartile. Autrement dit, les personnes présentant une forte expression de FAM216A tendent à maintenir des niveaux d'anticorps plus élevés sur une plus longue durée que celles ayant une expression plus faible. Néanmoins, la littérature sur FAM216A est à ce jour très limitée. On sait qu'il est exprimé dans divers types de cellules immunitaires, en particulier dans les lymphocytes B et les plasmocytes ([The Human Protein Atlas, 2025](#); [Uhlén et al., 2015](#)). Des variants génétiques modulant l'expression de FAM216A ont été associés à des différences spécifiques aux cellules myéloïdes dans le contexte du microenvironnement tumoral récemment ([Pagadala et al., 2023](#)). À notre connaissance, aucune implication de ce gène dans la réponse vaccinale humorale n'avait été rapportée jusqu'à présent. Des études complémentaires, menées sur de nouvelles données, permettraient de confirmer le potentiel de FAM216A comme prédicteur de la durabilité de la réponse vaccinale. En parallèle, des investigations expérimentales ciblées pourraient contribuer à élucider le rôle fonctionnel de ce gène dans les cellules B ou les plasmocytes.

Enfin, les performances du modèle final ont été étudiées à l'aide de critères d'adéquation : les graphiques de VPC pour le modèle initial sans covariables et pour le modèle final incluant FAM216A, ainsi que les courbes d'ajustement individuel pour le modèle final, sont fournis en Annexe A.4. Ces diagnostics suggèrent la capacité du modèle à capturer la dynamique observée des anticorps, et indiquent que l'inclusion de FAM216A pourrait contribuer à expliquer une partie de la variabilité inter-individuelle de la réponse humorale. Ces résultats sont préliminaires et restent à confirmer sur de nouvelles analyses exploratoires.

3.4 Perspectives

L'objectif de ce travail de thèse est de développer une stratégie robuste de sélection de covariables dans les modèles mécanistes, dans un contexte de grande dimension. De ce fait, l'article Lasso-SAMBA présente une contribution innovante en couplant la régression Lasso avec un algorithme de sélection stable pour la construction de modèles NLEM. Cette approche étend l'algorithme step-SAMBA à la grande dimension en remplaçant la procédure pas-à-pas

classique utilisée pour la sélection du modèle de covariables par une optimisation pénalisée, calibrée à l'aide de bornes théoriques sur le nombre espéré de faux positifs, réduisant les fausses découvertes et améliorant la robustesse de la sélection. Les études de simulation rapportées montrent que Lasso-SAMBA surpasse nettement, en termes de sélection, de contrôle du taux de faux positifs, les méthodes alternatives step-SAMBA, dont Lasso-SAMBA est l'extension, et SAEMVS, à notre connaissance la seule autre méthode proposant une sélection de covariables dans les modèles NLEM en grande dimension. En outre, son application aux données des études sur la vaccination contre le virus VZV et Ebola a mis en lumière des associations entre gènes et paramètres, cohérentes sur le plan biologique pour la première application et à approfondir pour la seconde. Ces résultats soulignent l'intérêt de la méthode proposée pour l'inférence de covariables dans les modèles basés sur des EDOs, tant sur les plans méthodologique que pratique.

Dans la continuité, plusieurs extensions méthodologiques naturelles pourraient être explorées dans le cadre de ce travail. Tout d'abord, l'utilisation de pénalités de type SCAD (Fan et Li, 2001) et MCP (Zhang, 2010) constitue une extension naturelle pour limiter le biais introduit par le Lasso (le *shrinkage*) sur les estimateurs. Ces pénalités sont définies, pour un coefficient β_k de $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, de la manière suivante :

$$\text{pen}_{\text{SCAD}}(\beta_k) = \begin{cases} \lambda|\beta_k| & , \text{ si } |\beta_k| \leq \lambda, \\ \frac{2a\lambda|\beta_k| - \beta_k^2 - \lambda^2}{2(a-1)} & , \text{ si } \lambda < |\beta_k| < a\lambda, \\ \frac{\lambda^2(a+1)}{2} & \text{ sinon.} \end{cases}, \quad \text{avec } \lambda > 0 \text{ et } a > 2;$$

$$\text{pen}_{\text{MCP}}(\beta_k) = \begin{cases} \lambda|\beta_k| - \frac{\beta_k^2}{2a} & , \text{ si } |\beta_k| \leq a\lambda, \\ \frac{a\lambda^2}{2} & \text{ sinon.} \end{cases}, \quad \text{avec } \lambda > 0 \text{ et } a > 1.$$

Dans le même objectif, l'étape de sélection par Lasso dans l'algorithme Lasso-SAMBA pourrait être remplacée par un Lasso adaptatif (Zou, 2006), incluant des poids pour chaque covariable dans la pénalité :

$$\text{pen}_{\text{la}}(\boldsymbol{\beta}) = \lambda \sum_{k=1}^p w_k |\beta_k|,$$

où $(w_k)_{k \leq p}$ sont des poids ajustés aux données. Il est par exemple possible de définir w_k comme l'inverse du coefficient β_k estimé par la méthode des moindres carrés classique.

Ensuite, une limitation connue du Lasso est sa sensibilité à la présence de covariables fortement corrélées entre elles. Nous nous sommes appliqués à produire des simulations dans ce cas de figure, qui correspond à une réalité biologique certaine. Les matrices de covariables utilisées pour vérifier empiriquement l'efficacité de la méthode ont été générées à l'aide de matrices de corrélation estimées à partir de données réelles. Néanmoins, dans des contextes de corrélations encore plus fortes, ou bien lorsque nos structures biologiques sont intrinsèquement groupées, comme les gènes intervenant dans une même voie, le *group Lasso* (Yuan et Lin, 2006) pourrait être intégré à l'algorithme Lasso-SAMBA. Il s'agit d'une méthode permettant la sélection simultanée de groupes de variables, en appliquant une pénalité à ces groupes prédéfinis plutôt qu'à chaque covariable indépendamment. Si l'on considère nos p covariables réparties en L groupes, le *group*

Lasso utilise une pénalité de la forme :

$$\text{pen}_{\text{gl}}(\boldsymbol{\beta}) = \lambda \sum_{l=1}^L \|\boldsymbol{\beta}_l\|_2,$$

appliquant ainsi la pénalité au niveau des groupes de covariables. Une limite majeure de cette méthode reste néanmoins la nécessité de prédéfinir ces groupes, ce qui peut naturellement induire un biais dans l'estimation. Dans la continuité, afin de tenir compte du caractère groupé des données sans nécessairement sélectionner toutes les variables d'un même groupe, comme le fait le *group Lasso*, le *sparse group Lasso* (Friedman *et al.*, 2010a) pourrait permettre un résultat plus parcimonieux. La pénalité combine à cet effet une pénalité Lasso classique L^1 et la pénalité du *group Lasso* précédemment définie, de sorte que :

$$\text{pen}_{\text{sgl}}(\boldsymbol{\beta}) = \lambda_1 \sum_{l=1}^L \|\boldsymbol{\beta}_l\|_2 + \lambda_2 \|\boldsymbol{\beta}\|_1,$$

où $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_L)$ correspond au vecteur complet des coefficients.

Une autre piste complémentaire serait d'explorer des méthodes de sélection de variables alternatives, comme la sélection de variables par forêts aléatoires (Breiman, 2001; Genuer *et al.*, 2010). Ces méthodes consistent à combiner des arbres de décision construits à partir de multiples *bootstraps* des données initiales. Pour les problèmes de régression, de ces arbres, on dérive des mesures d'importance des variables basé par exemple sur la décomposition des variances (Grömping, 2007), à partir desquelles on retient les covariables dont l'importance est élevée et stable tout au long des réplifications.

Par ailleurs, l'algorithme Lasso-SAMBA permet la construction conjointe des modèles de covariables, de corrélation et d'erreur. Lors de la construction du modèle de corrélation, une matrice de corrélation des effets aléatoires $\boldsymbol{\Omega}$ est estimée avec des corrélations sélectionnées entre certains paramètres du modèle. Dans la méthode actuelle, cette matrice de corrélation n'est toutefois pas prise en compte dans la sélection des covariables, réalisée indépendamment sur chaque paramètre. Nous pourrions tenir compte de cette corrélation en intégrant à notre sélection une étape de blanchiment comme proposé par Perrot-Dockès *et al.* (2018). Cette étape consiste à utiliser la matrice de variance-covariance des effets aléatoires $\boldsymbol{\Omega}$, correspondant à la matrice de covariances résiduelles pour le modèle de régression sur les paramètres individuels, et à travailler sur des quantités transformées. Si l'on considère $h(\boldsymbol{\psi}) = (h_l(\boldsymbol{\psi}_{li}))_{i \leq N, l \leq m}$ la matrice $N \times m$ des paramètres individuels transformés, alors on peut écrire le modèle de régression sous forme matricielle de la manière suivante :

$$h(\boldsymbol{\psi}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$

où $\mathbf{X}$ est la matrice $N \times p$ des covariables individuelles, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$ est la matrice $p \times m$ des coefficients, $\boldsymbol{\beta}_l = (\beta_{l1}, \dots, \beta_{lp})$, et $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_N)$ la matrice $N \times m$ des effets aléatoires avec $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \boldsymbol{\Omega})$. On omet ici l'intercept pour simplifier l'écriture. On peut alors *blanchir* le modèle en multipliant puis vectorisant l'expression précédente, ce qui conduit au modèle de régression

linéaire classique :

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E},$$

avec $\mathcal{Y} = \text{vec}\left(h(\boldsymbol{\psi})\boldsymbol{\Omega}^{-1/2}\right)$, $\mathcal{X} = \left(\boldsymbol{\Omega}^{-1/2}\right)^T \otimes \mathbf{X}$, $\mathcal{B} = \text{vec}(\boldsymbol{\beta})$ et $\mathcal{E} = \text{vec}\left(\boldsymbol{\eta}\boldsymbol{\Omega}^{-1/2}\right)$, où vec correspond à l'opérateur de vectorisation, et $\otimes$ est le produit de Kronecker. Cette formulation permet d'appliquer ensuite une estimation Lasso classique, tout en neutralisant les corrélations et en affinant l'estimation des effets des covariables.

Sur le plan appliqué, il serait également intéressant de pouvoir vérifier de manière plus approfondie les associations mises en lumière par notre méthode. Une validation biologique, permettant d'évaluer la cohérence immunologique et virologique des gènes sélectionnés, pourrait être envisagée soit par la réplication de l'analyse sur des essais indépendants, soit par une confrontation à la littérature, ou encore, si possible, par des validations expérimentales. Nous pourrions également explorer d'autres voies immunologiques, avec des stratégies de présélection différentes de celles considérées dans cette étude. De plus, il serait pertinent d'explorer d'autres temps de mesure pour les gènes étudiés. Dans nos applications, nous avons analysé l'expression génique au jour 56 pour le virus Ebola, première temps de mesure disponible, et au septième jour pour le virus VZV. Des marqueurs précoces, mesurés entre six et vingt-quatre heures après l'injection, pourraient toutefois porter davantage d'information sur le processus étudié. L'augmentation de la disponibilité de ces mesures, comme l'indique [Thiébaud et al. \(2025\)](#), permet désormais d'investiguer ces temporalités et d'affiner l'identification de signatures précoces, ce qui enrichirait notre compréhension des dynamiques étudiées. Par ailleurs, même si nous avons privilégié des applications liées à la réponse humorale post-vaccinale, la méthode reste générale et s'applique à tout modèle à effets mixtes. Des travaux futurs viseront ainsi à analyser des données de pharmacogénomique, afin d'évaluer la capacité du cadre proposé à prédire la variabilité de réponse aux traitements et à mettre en évidence des biomarqueurs précoces pertinents dans ce contexte.

– Chapitre 4 –

Estimation régularisée dans les modèles non-linéaires à effets mixtes en présence de biomarqueurs répétés de grande dimension

Publication :

Ce chapitre présente le travail de l'article suivant :

Gabaut, A., Hejblum, B.P., Bercu, A., Thiébaud, R., Proust-Lima, C.[†] & Prague, M.[†] **Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers.** *En préparation pour être soumis dans Biostatistics*

et

Gabaut, A., Proust-Lima, C.[†] & Prague, M.[†] **REMixed : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers in R.** <https://cran.r-project.org/web/packages/REMixed/vignettes/doc.pdf>.

Un paquet R **REMixed** est disponible sur le CRAN.

[†] : ces auteurs ont contribué de manière égale à ce travail.

Table des matières

4.1	Positionnement et contribution	106
4.2	Intégration des données transcriptomiques dans un modèle mécaniste de la réponse humorale	108
4.2.1	Modèle mécaniste et modèles d'observations	108
4.2.2	Méthode	109
4.2.3	Résultats	111
4.2.4	Discussion	114
4.3	Article : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers	115
4.4	Vignette : REMixed : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers in R	138
4.5	Perspectives	163

4.1 Positionnement et contribution

Ce chapitre s'intéresse au défi d'intégrer des données longitudinales d'expression génique, présentées en Section 1.2.2.2, au sein des modèles mécanistes, introduits en Section 2.1. Nous cherchons à exploiter les profils d'expression génique mesurés chez chaque individu comme mesures observées des compartiments latents de notre modèle mécaniste. Ainsi, contrairement à l'approche précédente Lasso-SAMBA, présentée dans le Chapitre 3, où les données transcriptomiques étaient uniquement utilisées en tant que covariables fixes pour expliquer la variabilité inter-individuelle des paramètres, nous proposons ici de les intégrer via le modèle d'observation en reliant directement les profils mesurés à certains compartiments latents du modèle structurel.

Nous considérons pour cela un modèle mécaniste dont la structure comprend $P + 1$ compartiments, notés $\mathbf{S} = (S_1, \dots, S_P)$ et R , décrits par un système d'EDOs :

$$\begin{cases} \frac{dS_p}{dt}(t) = F_{S_p}(\mathbf{S}(t), R(t), \boldsymbol{\psi}, t), & k = 1, \dots, K, \\ \frac{dR}{dt}(t) = F_R(\mathbf{S}(t), R(t), \boldsymbol{\psi}, t), & p = 1, \dots, P, \end{cases}$$

où F_{S_p} , $p = 1, \dots, P$, et F_R sont des fonctions paramétriques fixées décrivant le processus mécaniste. Pour un individu $i = 1, \dots, N$ dont les paramètres spécifiques sont $\boldsymbol{\psi}_i$, la résolution de ce système fournit des trajectoires $S_{pi}(t)$ pour $p = 1, \dots, P$ et $R_i(t)$. Notons que nous avons ici adapté la notation par rapport au Chapitre introductif 2 afin de distinguer explicitement les compartiments $S_1, \dots, S_P$ du compartiment latent R .

Pour le modèle d'observation, nous supposons d'une part que les compartiments $S_{1i}, \dots, S_{Pi}$, pour l'individu $i = 1, \dots, N$, sont mesurés à travers des marqueurs longitudinaux répétés Y_{pij} , $j = 1, \dots, n_{pi}$, définis par :

$$Y_{pij} = g_p(S_{pi}(t_{pij})) + \epsilon_{pij},$$

où $g_p(\cdot)$ est une fonction d'observation, par exemple l'identité ou une transformation connue comme log, et ϵ_{pij} est l'erreur de mesure iid selon $\mathcal{N}(0, \sigma_p^2)$. Ces observations Y_{pij} correspondent à des mesures classiques que l'on possède usuellement, telles que les concentrations d'anticorps dans le sang au cours du temps dans les essais cliniques présentés en Section 1.2. D'autre part, nous supposons que la trajectoire du processus latent $R_i(t)$ est indirectement observée à des temps répétés via un ensemble de K biomarqueurs. Plus précisément, pour chaque biomarqueur $k = 1, \dots, K$, on dispose pour l'individu i de mesures Z_{kij} , $j = 1, \dots, n_{ki}$, modélisées par une relation linéaire :

$$Z_{kij} = \mu_k + \alpha_k s_k(R_i(t_{kij})) + \varepsilon_{kij},$$

où $s_k(\cdot)$ est une transformation fixée appliquée au compartiment latent R_i , et ε_{kij} représente les erreurs de mesure iid suivant $\mathcal{N}(0, \sigma_k^2)$. Ce modèle d'observation traduit l'idée que chacun des K biomarqueurs transcriptomiques pourrait refléter l'état du compartiment latent R à un instant donné. Les paramètres α_k quantifient l'intensité du lien entre le biomarqueur Z_k et le processus R : un coefficient α_k nul signifie que le biomarqueur k n'apporte pas d'information sur R , tandis qu'un α_k non nul indique que Z_k varie en fonction de R .

L'objectif de notre travail est donc d'identifier quels biomarqueurs Z_k , pour $k = 1, \dots, K$, possèdent des coefficients α_k non nuls, fournissant ainsi de l'information sur le compartiment latent R . En d'autres termes, il s'agit de sélectionner, parmi un grand nombre de candidats, les marqueurs transcriptomiques longitudinaux réellement informatifs sur la dynamique de R . L'apport de ces nouvelles observations issues des données de transcriptomiques permettrait d'enrichir le modèle d'observation usuel et, par conséquent, d'inférer les compartiments latents du modèle mécaniste, tels que par exemple des populations de cellules sécrétrices d'anticorps, qui seraient autrement difficilement accessibles, voire inaccessibles. La considération de ces nouvelles observations permettrait d'améliorer la précision des estimations et l'identifiabilité de certains paramètres du modèle, notamment dans des cas où le modèle purement mécaniste souffre d'identifiabilité insuffisante. Ce bénéfice potentiel a d'ailleurs été mis en évidence par un travail préliminaire présenté en Section 4.2, où le gain de telles données en termes d'identifiabilité pratique a été étudié à l'aide de simulations.

À notre connaissance, sur le plan conceptuel, REMixed (*Regularized Estimation for non-linear Mixed effects models*) est la première approche exploitant des données transcriptomiques longitudinales de grande dimension en tant qu'observations indirectes d'un compartiment latent au sein d'un modèle mécaniste basé sur des EDOs. Notre démarche se distingue clairement des travaux antérieurs où les données transcriptomiques n'étaient utilisées que comme covariables fixes pour expliquer la variabilité inter-individuelle des paramètres des modèles à effets mixtes non-linéaires. C'était par exemple le cas de notre approche Lasso-SAMBA présentée dans le Chapitre 3 et de la méthode de Naveau *et al.* (2023), qui proposent une approche bayésienne de sélection de covariables dans les modèles non-linéaires à effets mixtes, combinant un *prior spike-and-slab* gaussien multivarié sur les coefficients de régression et une estimation par MCMC-SAEM, permettant d'inférer la probabilité *a posteriori* d'inclusion de chaque covariable. Dans le même objectif, Caillebotte *et al.* (2024) proposent une méthode de sélection pénalisée par Lasso dans un modèle conjoint longitudinal-survie, utilisant une optimisation par gradient stochastique avec opérateur proximal pour identifier les biomarqueurs transcriptomiques pertinents tout en estimant simultanément les paramètres du modèle. Elle se démarque également des études ayant incorporé des biomarqueurs de grande dimension dans des cadres semi-paramétriques, des modèles mixtes linéaires ou des modèles conjoints (Ni *et al.*, 2010; Zhao *et al.*, 2014; Pan et Huang, 2014; Li *et al.*, 2018; Sun et Basu, 2024), car aucune de ces approches n'intégrait des données longitudinales de grande dimension à l'intérieur d'un modèle mécaniste gouverné par des EDOs. En ce sens, REMixed propose un nouveau cadre de modélisation qui combine la puissance explicative des modèles mécanistes et la richesse informative des données transcriptomiques longitudinales.

Sur le plan méthodologique, nous développons un algorithme d'estimation régularisée qui combine la sélection de variables par Lasso et l'inférence des paramètres du modèle mécaniste via l'algorithme SAEM présenté en Section 2.1.2.5, appelé REMixed. Concrètement, nous cherchons à maximiser la vraisemblance pénalisée du modèle mécaniste, avec une pénalité de type Lasso, à l'aide d'un algorithme de descente de coordonnées cyclique (Tibshirani, 1996; Tseng, 2001; Friedman *et al.*, 2010b). Autrement, nous alternons l'optimisation de la vraisemblance

en l’optimisant selon chaque coordonnée itérativement jusqu’à convergence. Pour cela, nous regroupons les paramètres du modèle en deux directions : les paramètres de populations θ (comprenant les paramètres classiques de notre modèle mécaniste ainsi que les variances d’erreur et les intercepts des biomarqueurs $\mathbf{Z}_k$) et les paramètres de régularisation $\alpha = (\alpha_k)_{k \leq K}$. Ainsi nous alternons entre la maximisation classique selon θ , dont le terme de pénalité ne dépend pas, et la résolution d’un problème d’optimisation Lasso classique selon α . Ce schéma itératif en deux temps (sélection puis estimation) s’inspire des stratégies introduites dans des travaux antérieurs pour traiter des modèles à effets mixtes de grande dimension (par exemple, sélection de covariables dans les modèles mixtes et joints) qui ont montré leur robustesse (Pan et Huang, 2014; Li et al., 2018; Caillebotte et al., 2024). Nous étendons ici ce type d’approche au contexte des modèles mécanistes basés sur des EDOs. Nous proposons également une implémentation logicielle de la méthode sous la forme d’un paquet R.

4.2 Intégration des données transcriptomiques dans un modèle mécaniste de la réponse humorale

Dans le cadre d’un travail actuellement en préparation (Prague et al., 2025), nous avons étudié l’apport de données transcriptomiques longitudinales dans un modèle mécaniste décrivant l’établissement de la réponse immunitaire humorale après vaccination. L’objectif principal était d’évaluer dans quelle mesure l’intégration de telles données pouvait améliorer l’identifiabilité des paramètres d’un modèle basés sur des EDOs, au-delà des seules observations sérologiques usuelles.

4.2.1 Modèle mécaniste et modèles d’observations

Le modèle structurel considéré repose sur le cadre proposé par Clairon et al. (2023b), présenté en Section 2.3.3, et est inspiré des données mesurées dans l’essai clinique par Rinchai et al. (2022), présenté en Section 1.2.2, qui décrit la dynamique des cellules sécrétrices d’anticorps ou plasmocytes, notées S , et des anticorps Ab après administration d’une dose de vaccin. Chaque injection de vaccin, au temps $t_1 = 0$ et $t_2 = 21$, est supposée disparaître selon une cinétique exponentielle de taux δ_V , tout en produisant des cellules S à un taux f_{Mk} , avec $k = 1, 2$ pour les deux doses vaccinales. Les cellules S produisent des anticorps au taux μ , et disparaissent au taux δ_S . Les anticorps alors produits sont éliminés au taux δ_{Ab} . Le système différentiel est donc après la $k^{\text{ème}}$ injection de vaccin :

$$\begin{cases} \frac{dS}{dt} &= f_{Mk} e^{-\delta_V(t-t_k)} - \delta_S S, \\ \frac{dAb}{dt} &= \mu S - \delta_{Ab} Ab. \end{cases}$$

Pour le modèle statistique de notre modèle mécaniste, nous considérons que les paramètres f_{M2} et μ sont définis à l’échelle de l’individu, générant de la variabilité inter-individuelle dans

nos données, de sorte que pour l'individu $i = 1, \dots, N$, ces paramètres suivent les modèles :

$$\begin{aligned}\log(f_{M2i}) &= \log(f_{M20}) + \eta_i^f, \\ \log(\mu_i) &= \log(\mu_0) + \eta_i^\mu,\end{aligned}$$

avec $\eta_i^f \sim \mathcal{N}(0, \omega_f^2)$ et $\eta_i^\mu \sim \mathcal{N}(0, \omega_\mu^2)$. Pour l'individu $i = 1, \dots, N$, caractérisé par le vecteur de paramètres $\psi_i = (f_{M1}, \mu_i, f_{M2i}, \delta_V, \delta_S, \delta_{Ab})$, nous notons finalement Ab_i, S_i les dynamiques individuelles résultant du système différentiel considéré pour ce vecteur de paramètres.

Les mesures d'anticorps, obtenues à partir de prises de sang, sont en général les seules mesures utilisées pour estimer les paramètres des modèles de réponse immunitaire humorale (Pasin *et al.*, 2019; Alexandre, 2022; Clairon *et al.*, 2023b). Toutefois, la disponibilité croissante de données transcriptomiques offre de nouvelles sources d'information (Thiébaud *et al.*, 2025). Dans notre contexte, nous supposons ici que ces mesures permettent d'obtenir un proxy longitudinal des plasmocytes, ou les cellules S , responsables de la sécrétion d'anticorps, et d'intégrer ainsi une information complémentaire aux mesures sérologiques usuelles. Ainsi, afin de relier ce modèle structurel aux données, deux modèles d'observation sont considérés :

- M_{obs1} : seules les concentrations d'anticorps sont observées :

$$Y_{ij1} = \log_{10}(Ab_i(t_{ij1})) + \epsilon_{ij1},$$

avec $\epsilon_{ij1} \sim \mathcal{N}(0, \sigma_1^2)$.

- M_{obs2} : les concentrations d'anticorps et un proxy des cellules S sont conjointement observés :

$$\begin{cases} Y_{ij1} &= \log_{10}(Ab(t_{ij1})) + \epsilon_{ij1}, \\ Y_{ij2} &= \alpha_S S(t_{ij2}) + \epsilon_{ij2}, \quad \epsilon_{ij2} \sim \mathcal{N}(0, \sigma_2^2). \end{cases}$$

Le paramètre α_S traduit l'intensité du lien entre le proxy mesuré et les cellules S .

4.2.2 Méthode

Selon l'analyse de l'identifiabilité théorique de ce modèle (Prague *et al.*, 2025), les paramètres f_{M1} et δ_V ne peuvent être estimés, et sont donc fixés par la suite à leur valeur réelle. Les paramètres δ_S et δ_{Ab} ne peuvent être conjointement estimés lorsque seuls les anticorps sont observés. Avec la mesure du proxy également présente, l'identifiabilité théorique de ces paramètres est alors satisfaite. Finalement, les paramètres $\theta_1 = (\delta_S, f_{M20}, \mu_0)$ sont identifiables théoriquement lorsque les anticorps seuls sont observés, les paramètres $\theta_2 = (\delta_S, \delta_{Ab}, f_{M20}, \mu_0)$ ne sont en revanche observables que si le proxy des cellules S est également observé avec les données anticorps usuelles.

Notre objectif est maintenant d'étudier l'identifiabilité pratique du système. Pour cela nous avons mené une étude de simulation afin d'évaluer l'identifiabilité pratique des paramètres. Nous avons considéré trois situations : la première correspond au modèle d'observation M_{obs1} où

nous considérons donc seulement les anticorps observés ; la deuxième et la troisième situation correspondent au modèle d'observation M_{obs2} et pour lequel nous considérons deux niveaux de ratio signal-bruit, α_S valant 1 ou 0.1. Pour chaque situation d'observation, nous avons généré $R = 100$ jeux de données indépendants, de $N = 50$ individus. Les valeurs des paramètres utilisées sont données dans les Tables 4.1 et 4.2. Les mesures d'anticorps ont été simulées aux jours 0, 7, 14, 21, 28 et 35 tandis que les mesures transcriptomiques ont été supposées disponibles des jours 0 à 14 et 21 à 35. Les données simulées pour le premier réplicat sont données à titre illustratif en Figure 4.1.

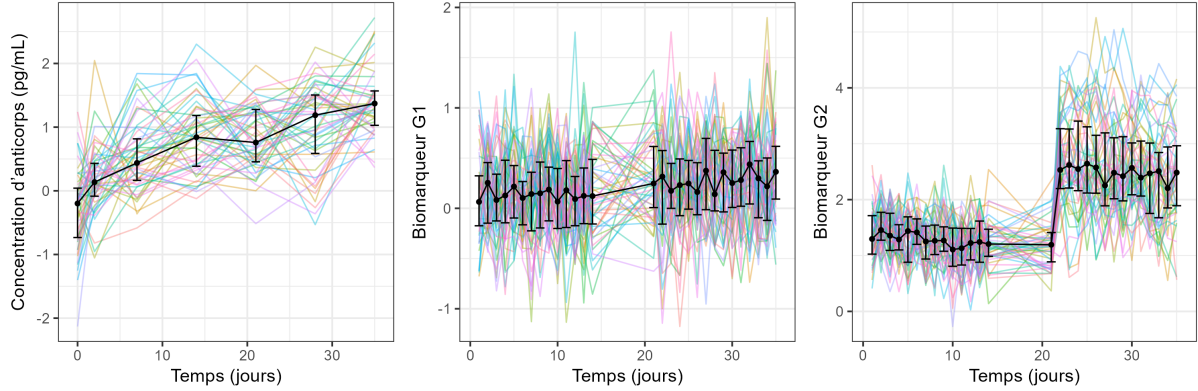


FIGURE 4.1 – Cinétique des anticorps et des biomarqueurs G1 et G2 simulés, pour $N = 50$.
 Les lignes noires indiquent la médiane et l'intervalle interquartile à chaque visite.

Pour chaque réplicat $r = 1, \dots, R$, nous appliquons l'algorithme SAEM (Delyon *et al.*, 1999; Kuhn et Lavielle, 2004) accessible via Monolix (Lixoft, 2024b). Nous réalisons deux estimations distinctes : d'une part les paramètres θ_1 , en maintenant δ_{Ab} fixé à sa valeur réelle, et d'autre part les paramètres θ_2 . Les erreurs standards sur les paramètres sont estimées de deux manières : par inversion de la matrice de Fisher $\hat{s}^{(r)}$ sur Monolix et par *bootstrap* utilisant 200 rééchantillonnages avec remise $\hat{s}_B^{(r)}$. Pour évaluer la qualité de l'estimation, nous utilisons les métriques suivantes, pour un paramètre de valeur réelle θ :

- **Biais relatif moyen** : $\frac{1}{R} \sum_{r=1}^R \left(\frac{\hat{\theta}^{(r)} - \theta}{\theta} \right)$,
 où $\hat{\theta}^{(r)}$ est l'estimation finale issue du réplicat r .
- **Écart-type empirique** : $\sqrt{\frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}^{(r)} - \bar{\hat{\theta}})^2}$,
 avec $\bar{\hat{\theta}} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)}$.
- **Erreur standard estimée** : $\frac{1}{R} \sum_{r=1}^R \widehat{s(\theta)}^{(r)}$.

- **Probabilité de couverture de l'intervalle de confiance à 95% :**

$$\frac{1}{R} \sum_{r=1}^R \mathbb{I} \left(\theta \in \left[\hat{\theta}^{(r)} \pm q_{97.5} \cdot \widehat{s(\theta)}^{(r)} \right] \right),$$

avec $q_{97.5}$ le quantile 97.5% de la loi normale standard.

4.2.3 Résultats

Les résultats des simulations sont reportés dans les Tables 4.1 et 4.2. Dans la première configuration, dont les résultats sont donnés dans la Table 4.1, aucun problème de convergence n'a été observé lors de l'estimation par SAEM et de l'inversion de la matrice d'information de Fisher. Dans la seconde configuration, présentée dans la Table 4.2, un faible pourcentage de non-convergence est observé, atteignant 4% lorsque seul les anticorps sont utilisés (M_{obs1}), et diminuant à 1% lorsque le proxy transcriptomique est intégré (M_{obs2}), quel que soit son niveau de bruit.

Lorsque seules les mesures d'anticorps sont prises en compte, les paramètres structuraux sont estimés avec de forts biais relatifs. En particulier, f_{M20} présente un biais supérieur à 90% et un taux de couverture très faible, atteignant seulement 17% avec la matrice d'information de Fisher estimée et 33% par les erreurs standards estimées par *bootstrap*, comme présenté dans la Table 4.1. De même, la variance inter-individuelle ω_f est largement surestimée (biais de plus de 600%). Les taux de couverture des intervalles de confiance issus de la matrice de Fisher estimée sont très faibles (3% pour δ_S par exemple), alors qu'ils s'améliorent nettement lorsque les erreurs standards sont évaluées par *bootstrap* (80% pour δ_S). Ces résultats mettent en évidence la mauvaise performance de l'inversion directe de la matrice dans ce cadre, et la difficulté d'identifier conjointement δ_S et δ_{Ab} en l'absence de mesures complémentaires.

L'ajout du proxy transcriptomique améliore nettement la qualité des estimations. Les biais relatifs des paramètres structuraux deviennent faibles : pour f_{M20} , le biais passe de 91% sous M_{obs1} à seulement 0,5% dans le cas d'un proxy peu bruité, et à 19% dans le cas d'un proxy plus bruité. La couverture des intervalles de confiance, lorsque les erreurs standards sont calculées par *bootstrap*, est globalement proche du niveau nominal de 95%, et ce pour l'ensemble des paramètres, en particulier dans le cas où le proxy est peu bruité.

Néanmoins, des incohérences persistent lorsque les erreurs standards sont obtenues via l'inversion de la matrice de Fisher : par exemple, pour δ_S , la couverture chute à 22% dans le cas du proxy peu bruité. Ce problème disparaît lorsque les erreurs standards sont estimées par *bootstrap*, où les couvertures se rapprochent de 91-96%.

Paramètre	Valeur réelle	Estimation moyenne	Biais relatif	Erreur Standard		Couverture (95%)	
				$\hat{\sigma}_E$	$\hat{\sigma}$	estimée	bootstrap
M_{obs1} - Anticorps seul							
f_{M20}	4.13	0.335	-0.919	0.968	0.808	1.698	0.17
δ_S	0.009	0.003	-0.656	0.009	0.000	0.030	0.03
μ_0	0.41	0.453	0.106	0.053	0.055	0.124	0.84
ω_f	0.54	3.860	6.147	1.974	4.018	2.125	0.78
ω_μ	0.14	0.206	0.471	0.060	0.197	0.062	0.99
σ_1	0.59	0.580	-0.018	0.021	0.027	0.023	0.94
M_{obs2} - Anticorps et proxy peu bruité des cellules S							
f_{M20}	4.13	4.111	-0.005	0.375	0.342	0.347	0.93
δ_S	0.009	0.009	-0.018	0.001	0.000	0.001	0.22
μ_0	0.41	0.418	0.018	0.037	0.048	0.039	0.99
α_S	1	1.001	0.001	0.017	0.015	0.016	0.91
ω_f	0.54	0.530	-0.019	0.054	0.057	0.057	0.92
ω_μ	0.14	0.249	0.777	0.073	0.180	0.075	0.94
σ_1	0.59	0.578	-0.021	0.021	0.024	0.022	0.93
σ_2	0.44	0.440	-0.001	0.009	0.008	0.008	0.95
M_{obs2} - Anticorps et proxy bruité des cellules S							
f_{M20}	4.13	3.333	-0.193	1.308	1.084	1.405	0.71
δ_S	0.009	0.004	-0.513	0.009	0.000	0.009	0.03
μ_0	0.41	0.409	-0.003	0.047	0.040	0.048	0.92
α_S	0.1	0.103	0.032	0.014	0.013	0.014	0.86
ω_f	0.54	0.621	0.150	0.187	0.194	0.279	0.95
ω_μ	0.14	0.242	0.731	0.078	0.192	0.075	0.95
σ_1	0.59	0.578	-0.020	0.020	0.024	0.022	0.92
σ_2	0.44	0.441	0.002	0.008	0.008	0.008	0.96

TABLE 4.1 – Résultats des simulations pour l’identifiabilité pratique des paramètres θ_2 et α_S selon le modèle d’observation (M_{obs1} ou M_{obs2}) et le niveau de bruit, dans le cas d’un échantillon de taille $N = 50$.

NC correspond à “Non-Convergence” et indique le pourcentage de cas où la variance estimée n’a pas pu être calculée en raison de problèmes d’inversion de la matrice d’information de Fisher.

Paramètre	Valeur réelle	Estimation moyenne	Biais relatif	Erreur Standard		Couverture (95%)	
				$\hat{\theta}$	$\hat{\sigma}_E$	$\hat{\sigma}_B$	estimée bootstrap
M_{obs1} - Anticorps seul							
fM_{20}	4.13	0.349	-0.915	1.032	1.675	0.11	0.29
δ_S	0.009	0.010	0.104	0.021	0.002	0.09	1.00
μ_0	0.41	0.405	-0.013	0.091	0.051	0.70	0.98
δ_{Ab}	0.03	0.012	-0.590	0.026	0.001	0.05	0.76
ω_f	0.54	4.248	6.867	2.259	3.797	0.88	0.73
ω_μ	0.14	0.215	0.532	0.062	0.159	0.98	0.89
σ_1	0.59	0.579	-0.019	0.021	0.024	0.93	0.92
M_{obs2} - Anticorps et proxy peu bruité des cellules S							
fM_{20}	4.13	4.117	-0.003	0.375	0.335	0.95	0.94
δ_S	0.009	0.009	-0.017	0.001	0.000	0.26	0.96
μ_0	0.41	0.475	0.159	0.147	0.046	0.48	0.96
δ_{Ab}	0.03	0.045	0.509	0.039	0.001	0.04	0.95
α_S	1	1.001	0.001	0.017	0.014	0.86	0.96
ω_f	0.54	0.529	-0.021	0.053	0.057	0.93	0.92
ω_μ	0.14	0.249	0.777	0.081	0.140	0.91	0.78
σ_1	0.59	0.577	-0.021	0.021	0.023	0.92	0.90
σ_2	0.44	0.440	-0.001	0.009	0.008	0.94	0.94
M_{obs2} - Anticorps et proxy bruité des cellules S							
fM_{20}	4.13	3.338	-0.192	1.305	1.204	0.76	0.84
δ_S	0.009	0.003	-0.623	0.009	0.000	0.02	0.75
μ_0	0.41	0.438	0.068	0.117	0.048	0.63	1.00
δ_{Ab}	0.03	0.040	0.330	0.034	0.001	0.07	0.99
α_S	0.1	0.102	0.020	0.015	0.014	0.87	0.94
ω_f	0.54	0.636	0.177	0.240	0.243	0.97	0.98
ω_μ	0.14	0.242	0.727	0.075	0.203	0.97	0.83
σ_1	0.59	0.577	-0.021	0.020	0.024	0.93	0.90
σ_2	0.44	0.441	0.002	0.008	0.008	0.96	0.96

TABLE 4.2 – Résultats des simulations pour l’identifiabilité pratique des paramètres θ_1 et α_S selon le modèle d’observation (M_{obs1} ou M_{obs2}) et le niveau de bruit, dans le cas d’un échantillon de taille $N = 50$.

NC correspond à “Non-Convergence” et indique le pourcentage de cas où la variance estimée n’a pas pu être calculée en raison de problèmes d’inversion de la matrice d’information de Fisher.

4.2.4 Discussion

Dans l'ensemble, les résultats indiquent que l'utilisation de données sérologiques seules conduit à une estimation biaisée et peu précise de certains paramètres structuraux, avec une couverture basse des intervalles de confiance. L'ajout d'un proxy transcriptomique des cellules S permet d'améliorer les estimations, en réduisant les biais et en augmentant les taux de couverture. Ces gains sont particulièrement marqués lorsque le proxy est peu bruité, pour lequel la précision et la couverture sont très proches des niveaux nominaux.

Ce travail illustre l'apport des données transcriptomiques longitudinales dans l'analyse de la réponse immunitaire humorale, via des proxies de populations cellulaires, pour lever les problèmes d'identifiabilité pratique des modèles mécanistes, alors que l'utilisation des seules données sérologiques contraint à fixer un nombre plus important de paramètres. De plus, l'intégration d'un marqueur transcriptomique réduit les biais et améliore la précision des intervalles de confiance, à condition que le signal ne soit pas trop bruité. Ces résultats constituent un premier cadre méthodologique validant l'intérêt de l'approche REMixed, et démontrent que l'intégration de données transcriptomiques dans le modèle d'observation peut jouer un rôle clé pour inférer des compartiments latents difficilement mesurables, voire impossibles à mesurer, tels que les populations de plasmocytes.

Il est toutefois à noter que l'estimation de la matrice d'information de Fisher fournie par le logiciel Monolix (version 2023.R1) conduit, dans notre étude, à des erreurs standards et taux de couverture incohérents. Ce constat nous a conduit à privilégier l'approche par *bootstrap* pour obtenir des estimations fiables de l'incertitude. Ces difficultés pourraient en partie s'expliquer par des limitations propres à la version du logiciel utilisée, imposée par les contraintes techniques de la Plateforme Fédérative pour la Recherche en Informatique et Mathématiques (PlaFRIM), qui ne permet pas actuellement une mise à jour vers une version plus récente (version 2024.R1). Dans les dernières versions de Monolix, l'estimation par *bootstrap* est directement implémentée, ce qui constitue un progrès important. Enfin, il convient de souligner la limite liée au logiciel, dont les codes sources ne sont pas accessibles, restreignant ainsi la possibilité de vérifier et d'adapter les procédures internes d'estimation.

Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers

Auriane Gabaut^{1,2,3}, Boris P. Hejblum^{1,2,3}, Ariane Bercu^{1,4}, Rodolphe Thiébaud^{1,2,3,5}, Cécile Proust-Lima^{1†}, Mélanie Prague^{1,2,3*†}

¹ : Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France

² : Univ. Bordeaux, Inria, Bordeaux, France

³ : Vaccine Research Institute, Paris, France

⁴ : Univ. Paris-Cité, INSERM, Inria, U1346, Paris, France

⁵ : CHU de Bordeaux, Service d'Information Médicale, INSERM, U1219, F-33000 Bordeaux, France

* : melanie.prague@inria.fr

† : These authors contributed equally

Abstract : Mechanistic models based on ordinary differential equations (ODEs) effectively capture the dynamics of biological systems but typically require longitudinal observations across multiple compartments to ensure identifiability-data that are often unavailable in practice. We hypothesize that high-dimensional transcriptomic biomarkers can serve as indirect observations, revealing insights into latent biological compartments. To address this challenge, we introduce a regularized estimation for non-linear mixed effects models method, called REMixed, designed to simultaneously estimate the parameters of mechanistic models with latent compartments and select relevant longitudinal biomarkers related to them.

Our method employs an iterative algorithm combining Lasso-penalized maximum likelihood estimation with a proximal coordinate descent procedure. This framework extends previous approaches-originally developed for generalized linear mixed-effects models (GLMMs) and high-dimensional joint Cox models-to non-linear mixed-effects models (NLMEMs) governed by ODEs, which may or may not admit an analytical solution. Specifically, our algorithm alternates between a regularization step, which selects relevant biomarkers by optimizing a penalized likelihood using a second-order Taylor expansion, and a mechanistic inference step, where parameters of the ODE model are estimated via the Stochastic Approximation Expectation-Maximization (SAEM) algorithm, as implemented in the Monolix software.

We validate REMixed through simulations that demonstrate its robustness and efficacy in biomarker selection and parameter estimation. We also illustrate its practical utility for identifying immune response markers from transcriptomic data in a SARS-CoV-2 vaccine clinical trial.

Acknowledgment

This study was carried out within the framework of the University of Bordeaux’s France 2030 program RRI PHDS. Elements of this work were also conducted as part of the SOLVE project, which has received funding from the European Union’s Horizon Research and Innovation program under Grant Agreement No. 101137185.

We thank Simulations Plus, Lixoft division, for granting free academic access to the MonolixSuite. Computational resources for this study were provided by the PlaFRIM (Plateforme Fédérative pour la Recherche en Informatique et Mathématiques). We disclose that generative AI tools were used to polish the English language.

1 Introduction

Mechanistic models, particularly those based on systems of ordinary differential equations (ODEs), are central to the analysis of biological systems. By explicitly capturing the dynamic interactions among compartments, such as immune cell populations, these models have become essential tools in pharmacometrics (Ibarra *et al.*, 2021), immunology (Perelson *et al.*, 2018), and epidemic modeling (Ganser *et al.*, 2024). However, achieving identifiability within such models typically requires longitudinal measurements across multiple compartments (Lavielle, 2014). In many real-world settings, such as clinical trials, this cannot be achieved due to the inaccessibility of certain compartments. In the context of immunology and vaccine studies, for instance, longitudinal measurements are often limited to circulating antibodies, which are routinely sampled across individuals. Yet, the underlying biological processes governing antibody dynamics are more complex, involving latent compartments such as antibody-secreting cells (ASCs), which are typically not observable in practice (Pasin *et al.*, 2019). Mechanistic models can incorporate such latent compartments to provide a biologically meaningful description of antibody production and decay, but this reliance on unobserved variables introduces substantial challenges for parameter estimation and identifiability.

Recent advances in high-throughput technologies have dramatically increased the information available from each biological sample, particularly gene expression profiles (Lowe *et al.*, 2017), which can provide indirect observations of underlying biological processes. This has motivated a growing interest in integrating high-dimensional biomarkers into statistical and mechanistic models.

Several studies have demonstrated that incorporating longitudinal biomarkers enhances the understanding of complex biological systems, whether in semi-parametric frameworks (Ni *et al.*, 2010; Zhao *et al.*, 2014), linear or generalized mixed-effects models (Pan *et al.*, 2014; Li *et al.*, 2018), or joint models of high-dimensional longitudinal biomarkers and survival outcomes (Sun *et al.*, 2024). Deconvolution approaches have also shown that transcriptomic data provide important information about cell abundances (Wang *et al.*, 2019). A number of

these approaches also rely on iterative estimation and selection strategies (Pan et Huang, 2014; Li et al., 2018; Caillebotte et al., 2024), reinforcing the idea that such algorithmic schemes are both robust and effective in high-dimension with complex modeling contexts. However, these approaches are not designed to accommodate mechanistic models based on systems of ODEs, as typically used, for instance, to study the dynamics of the immune response to a vaccine. In this context, high-dimensional transcriptomic measurements, such as gene expression profiles, contain valuable information on immune responses but are not directly linked to the latent compartments defined by the mechanistic model. Capturing this indirect relationship while accounting for the complexity of the ODE-based non-linear mixed-effects (NLMEM) structure requires new methodological approaches capable of performing variable selection and parameter inference in the presence of longitudinal transcriptomic data and unobserved biological processes.

In parallel, for ODE-based NLMEMs, previous works have shown that integrating time-fixed transcriptomic data into non-linear mixed-effects models is feasible (Caillebotte et al., 2024; Gabaut et al., 2025b). However, these contributions leveraged transcriptomic data exclusively as time-fixed covariates rather than time-dependent outcomes, and did not involve dynamics governed by differential equations.

Assuming that transcriptomic data comprise valuable biomarkers for informing biological system dynamics, we develop a general framework to incorporate longitudinal large-dimensional data as measures of a latent compartment over time in non-linear mixed-effects models based on ordinary differential equations (ODEs). The method, called REMixed (Regularized Estimation for non-linear Mixed effects models), relies on an iterative algorithm based on Lasso-penalized maximum likelihood estimation (Tibshirani, 1996). It uses a proximal coordinate descent method (Tseng, 2001; Friedman et al., 2010b), alternating between a Lasso-based selection of the relevant biomarkers and an update of the mechanistic model parameters using the stochastic approximation EM algorithm (Delyon et al., 1999; Kuhn et Lavielle, 2005). Such a two-step strategy has previously demonstrated promising performance in related contexts, including high-dimensional joint survival and mixed-effects models (Caillebotte et al., 2024), and the selection and estimation of fixed or random effects in mixed-effects models (Pan et Huang, 2014; Li et al., 2018).

The paper is organized as follows. Section 2 details the proposed REMixed methodology, including the model formulation, the estimation algorithm, and implementation considerations. In Section 3, we present results from simulation studies designed to assess estimation accuracy and variable selection performance. Section 4 applies the proposed method to real-world transcriptomic data from COVID-19 mRNA vaccine clinical trials (Rinchai et al., 2022). Finally, Section 5 discusses the potentials and limitations of our approach, along with future perspectives.

2 Method

2.1 Model definition

We consider the framework of a NLMEM based on a system of ODEs. Suppose that, for a subject $i = 1, \dots, N$, we have a differential system of equations, called later our structural model, containing $P + 1$ compartments called $(S_p)_{p \leq P}$ and R such that

$$\begin{cases} S'_p(t) &= F_{S_p}(S_p(t), R(t), \boldsymbol{\psi}), \quad p \leq P, \\ R'(t) &= F_R(S_p(t), R(t), \boldsymbol{\psi}), \end{cases}$$

depending on a m -vector of parameters, denoted $\boldsymbol{\psi}$, and with F_{S_p} , $p \leq P$, and F_R fixed parametric functions describing the mechanistic process. We distinguish between compartments S_p and R since they will be associated with different observation models, defined bellow. For individual i ($i = 1, \dots, N$), trajectories resulting from this system of ODEs are denoted S_{pi} , $p \leq P$, and R_i with individual-specific parameters $\boldsymbol{\psi}_i = (\psi_{li})_{l \leq m}$. Each parameter ψ_{li} ($l = 1, \dots, m$) is defined by a generalized linear model :

$$h_l(\psi_{li}) = h_l(\psi_{l \text{ pop}}) + \mathbf{X}_i \boldsymbol{\beta}_l + \eta_{li},$$

with the vector of individual covariates $\mathbf{X}_i$ associated with regression parameters $\boldsymbol{\beta}_l$, $\psi_{l \text{ pop}}$ the typical value of parameter ψ_l and the vector of individual random effects η_{li} normally distributed such that $\boldsymbol{\eta}_i = (\eta_{li})_{l \leq m} \stackrel{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Omega})$. The covariance matrix $\boldsymbol{\Omega}$ has potentially null diagonal elements for the parameters without random effects.

The distinction between compartments S_p , $p = 1, \dots, P$, and R depends on the observation model considered to link the latent compartment S_p and R to their measurements. For S_{pi} , $p \leq P$, we assume the individual trajectories are observed through repeated marker observations, denoted Y_{pij} , up to a fixed transformation g_p at observation times t_{pij} (with $j \leq n_{pi}$) :

$$Y_{pij} = g_p(S_{pi}(t_{pij})) + \epsilon_{pij}, \quad (4.1)$$

with normally distributed error term $\epsilon_{pij} \stackrel{iid}{\sim} \mathcal{N}(0, \varsigma_p^2)$.

We further assume the trajectory of latent process R_i is measured at repeated occasions by a set of large-dimensional markers, denoted Z_{kij} for marker $k = 1, \dots, K$ and individual $i = 1, \dots, N$ at observation time t_{kij} (with $j \leq n_{ki}$), using a linear observation model :

$$Z_{kij} = \mu_k + \alpha_k s_k(R_i(t_{kij})) + \varepsilon_{kij}, \quad (4.2)$$

with a fixed transformation s_k and independent measurement errors $\varepsilon_{kij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$.

For the rest of the manuscript, we separate the complete set of parameters in two : on one hand the large-dimensional marker intercepts $\boldsymbol{\alpha} = (\alpha_k)_{k \leq K}$ comprising the coefficients linking the markers to their underlying latent process that we aim to regularize, and on the other hand $\boldsymbol{\theta} = (\boldsymbol{\psi}_{pop}, \boldsymbol{\beta}, \boldsymbol{\Omega}, (\sigma_k^2)_{k \leq K}, (\varsigma_p^2)_{p \leq P}, (\mu_k)_{k \leq K})$ comprising the mechanistic model parameters.

2.2 Regularized estimation procedure for large-dimension biomarkers

We propose to estimate this NLMEM in a large dimensional setting using a penalized maximum likelihood approach that regularizes the α parameters linking the latent compartment R_i to its observations in large-dimension.

2.2.1 NLMEM log-likelihood

For individual $i \leq N$, the contribution to the likelihood is

$$\mathcal{L}_i(\alpha, \theta) = f(Y_i, Z_i | \alpha, \theta) = \int_{\eta_i} f(Y_i, Z_i | \alpha, \theta, \eta_i) \times f(\eta_i | \theta) d\eta_i,$$

with $Y_i = (Y_{pij})_{p \leq P, j \leq n_{ip}}$ and $Z_i = (Z_{kij})_{k \leq K, j \leq n_{ik}}$ the vectors of observations defined in Equations (4.1) and (4.2) respectively. The latent compartment $(S_{pi})_{p \leq P}$ and R_i , and by extensions their observations Y_i and Z_i , are independent conditionally to the individual random effects so the individual log-likelihood can be rewritten as a product of the : (1) density function of observation Z_i , $f_Z(Z_i | \alpha, \theta, \eta_i)$; (2) density function of observations Y_i , $f_Y(Y_i | \theta, \eta_i)$ not depending on α ; (3) density function of the random effects η_i , $f(\eta_i | \theta)$; so that, with independent individuals :

$$LL(\alpha, \theta) = \sum_{i \leq N} \log \left(\int_{\eta_i} f_Y(Y_i | \theta, \eta_i) f_Z(Z_i | \alpha, \theta, \eta_i) f(\eta_i | \theta) d\eta_i \right). \quad (4.3)$$

Additional details about log-likelihood expression can be found in Supplementary Material B.1.1.

The log-likelihood defined in Equation 4.3 does not have a closed-form formula. In practice, to approximate this log-likelihood as well as its derivatives, we rely on a multidimensional adaptive Gauss-Hermite quadrature (Davidian et Gallant, 1993), combined with a solver that estimates all compartments simultaneously for a given vector of individual parameters ψ_i (Curtiss et Hirschfelder, 1952).

In low-dimensional settings, all the parameters of the NLMEM, namely α and θ , can be directly estimated by maximum likelihood estimation through a standard application of the SAEM algorithm (Dempster et al., 1977; Delyon et al., 1999), proved to be particularly suitable and efficient (Kuhn et Lavielle, 2005). This algorithm is a stochastic version of the classical EM algorithm, specifically designed to handle latent variable models where the complete likelihood is difficult or even impossible to compute explicitly. SAEM alternates between two main steps, repeated until convergence : a stochastic expectation step (E-step) and a maximization step (M-step). During the E-step, the algorithm simulates the latent individual effects conditionally on the current parameter values using a Markov Chain Monte Carlo (MCMC) approach. These simulations provide a numerical approximation of the complete likelihood. In the subsequent M-step, the fixed and variance parameters of the model are updated by maximizing this stochastic approximation. The algorithm offers an optimized and robust procedure for estimating NLMEM parameters. However, in large-dimensional contexts, particularly when multiple biomarkers

are measured simultaneously, the number of parameters to estimate increases substantially, making direct application of SAEM challenging in terms of convergence and numerical stability. Indeed, in practice, our models usually involve 2-3 compartments, with parameters including the population values, variances of random effects, possible covariate effects, and error model parameters. In addition, each biomarker introduces three extra parameters $(\mu_k, \alpha_k, \sigma_k)$, so that analyzing tens of biomarkers quickly leads to several hundreds of parameters to be estimated. This difficulty thus arises mainly from the increased complexity of the parameter space to be explored and the challenge of obtaining accurate estimates of the latent individual effects from a limited number of observations per subject. In the following, we detail our regularized estimation procedure which is adapted to large dimensional settings.

2.2.2 Proximal coordinate descent algorithm

The problem that this work aims to solve is maximizing the log-likelihood of Equation (4.3) under Lasso penalization (Tibshirani, 1996). Thus, we want to estimate $(\alpha_\lambda^*, \theta_\lambda^*)$ so that, for a fixed regularization parameter $\lambda > 0$:

$$(\alpha_\lambda^*, \theta_\lambda^*) = \arg \max_{(\alpha, \theta)} LL_{pen, \lambda}(\alpha, \theta) = \arg \max_{(\alpha, \theta)} (LL(\alpha, \theta) - \lambda \|\alpha\|_1). \quad (4.4)$$

To achieve this, we use a coordinate descent algorithm, meaning we sequentially maximize the penalized log-likelihood of Equation (4.4) in a direction while keeping other parameters at their previous values, and we iterate this in all the directions until convergence. Presently, we consider two directions : at each iteration, the optimization algorithm updates the θ and α parameters in two alternate steps, updating the one while fixing the other to its previous value. The convergence is reached when the penalized log-likelihood and the parameters stabilize below convergence hyper-parameters τ_1 and τ_2 , respectively.

The algorithm is summarized in Algorithm 1 and the two alternate steps are detailed below.

Algorithm 1 : REMixed algorithm.

Let $\lambda > 0$ **be fixed** (See Section 2.2.4 for its outer selection) ;

Let $\alpha^{(0)}, \theta^{(0)}$ be obtained, for instance, by MLE or SAEM from a random subset of biomarkers (see initialization step in Section 2.2.3) ;

Compute $LL(\alpha^{(0)}, \theta^{(0)})$, $\partial_\alpha LL(\alpha^{(0)}, \theta) \big|_{\alpha=\alpha^{(0)}}$, and $\partial_\alpha^2 LL(\alpha^{(0)}, \theta) \big|_{\alpha=\alpha^{(0)}}$;

At iteration k :

Compute $\alpha^{(k+1)}$ using Equation (4.5) with $\hat{\alpha} = \alpha^{(k)}$ and $\hat{\theta} = \theta^{(k)}$ fixed, detailed in Section 2.2.2.1 ;

Estimate $\theta^{(k+1)}$, with α parameters set to $\alpha^{(k+1)}$, using the SAEM algorithm starting from $\theta^{(k)}$, detailed in Section 2.2.2.2 ;

Update $LL(\alpha^{(k+1)}, \theta^{(k+1)})$, $\partial_\alpha LL(\alpha, \theta^{(k+1)}) \big|_{\alpha=\alpha^{(k+1)}}$, and $\partial_\alpha^2 LL(\alpha, \theta^{(k+1)}) \big|_{\alpha=\alpha^{(k+1)}}$ using a new estimation of the posterior distribution of the random effects ;

Evaluate the convergence criteria and stop the algorithm if all are met :

$$\begin{aligned} & - \left| LL_{pen_\lambda}(\alpha^{(k+1)}, \theta^{(k+1)}) - LL_{pen_\lambda}(\alpha^{(k)}, \theta^{(k)}) \right| < \tau_1 ; \\ & - \left| \left(\frac{\alpha^{(k)} - \alpha^{(k+1)}}{\alpha^{(k)} + 1}, \frac{\theta^{(k)} - \theta^{(k+1)}}{\theta^{(k)} + 1} \right) \right| < \tau_2. \end{aligned}$$

2.2.2.1 Updating the biomarker contributions

Given fixed current estimates $(\hat{\alpha}, \hat{\theta}) = (\alpha^{(k)}, \theta^{(k)})$ from iteration k , the first coordinate to update is the regularization parameters α , with fixed $\hat{\theta}$. The optimization problem to update α from Equation (4.4) then reduces to solving a Lasso-type penalized likelihood :

$$\max_{\alpha \in \mathbb{R}^K} LL_{pen_\lambda}(\alpha, \hat{\theta}) = \max_{\alpha \in \mathbb{R}^K} \left(LL(\alpha, \hat{\theta}) - \lambda \|\alpha\|_1 \right).$$

We approximate the log-likelihood function $LL(\alpha, \hat{\theta})$ using a second-order Taylor expansion around the current estimate $\hat{\alpha}$ (details in Supplementary Material B.1.2). This approximation yields an explicit coordinate-wise update rule for each parameter α_k , $k = 1, \dots, K$:

Define $A_k = \partial_{\alpha_k} LL(\alpha, \hat{\theta}) \big|_{\hat{\alpha}} + \hat{x}_{kk} \hat{\alpha}_k$, with $\hat{x}_{kk}$ the k th diagonal elements of the opposite of the Hessian matrix at $(\hat{\alpha}, \hat{\theta})$, assumed negative definite.

The parameter updates follow a standard soft-thresholding rule :

$$\alpha_k = \begin{cases} \frac{A_k + \lambda}{\hat{x}_{kk}}, & \text{if } A_k < -\lambda, \\ \frac{A_k - \lambda}{\hat{x}_{kk}}, & \text{if } A_k > \lambda, \\ 0, & \text{otherwise.} \end{cases} \quad (4.5)$$

Additional mathematical derivations and detailed proofs are provided in the Supplementary Material B.1.2.

2.2.2.2 Updating the mechanistic model parameters

Secondly, for the fixed $\alpha^{(k+1)}$ updated previously, the problem is now equivalent to a classical likelihood maximization problem, that can be efficiently solved using the SAEM algorithm for NLMEM, presented in Section 2.2.1.

2.2.3 Initialization strategy

The estimation and selection procedure requires an initialization step to ensure convergence toward a reasonable solution. We initialized the model by randomly pairing markers Z_k , estimating the model for each pair using a standard SAEM algorithm (described in Section 2.2), and estimating the maximum log-likelihood under $\alpha_k = 0$ for all the other markers. The initialization yielding the best joint log-likelihood for all markers Z_k and Y_p was selected for the final estimation. This strategy helps ensure that the starting point is sufficiently plausible, facilitating likelihood computation via the adaptive Gauss-Hermite quadrature.

2.2.4 Outer selection of the penalty parameter

To select the penalty parameters λ , the algorithm is applied over a grid of hyperparameters Λ defined as

$$\Lambda = \left\{ \lambda_{\max} \times \alpha_{\lambda}^{l/N_{\lambda}} \mid l = 1, \dots, N_{\lambda} \right\},$$

where N_{λ} is the number of grid points, and α_{λ} controls the decay rate of the penalty levels, typically set to $\alpha_{\lambda} = 0.001$ (Park et al., 2007; Friedman et al., 2010b). The maximal penalty $\lambda_{\max}$ is approximate by

$$\lambda_{\max} = \max \left(\partial_{\alpha} LL(\theta^{(0)}, \alpha) \Big|_{\alpha=0} \right).$$

The final model is determined as the one minimizing the BICc, a corrected version of the BIC (Schwarz, 1978) well suited for NLMEM (Delattre et al., 2014) and defined as :

$$BICc(\hat{\alpha}, \hat{\theta}) = -2LL(\hat{\alpha}, \hat{\theta}) + \log(N) \dim(\theta_R) + \log(n_{tot})(\dim(\theta_F) + \#\{k = 1, \dots, K | \alpha_k \neq 0\}).$$

where n_{tot} is the total number of observations, and the population parameter θ is decomposed into (θ_R, θ_F) with θ_R and θ_F the parameters from θ linked to the random effects and to the fixed effects, respectively.

Finally, we select the optimal penalty parameter λ^* such that

$$\lambda^* = \arg \min_{\lambda} BICc(\hat{\alpha}_{\lambda}, \hat{\theta}_{\lambda}),$$

with $\hat{\theta}_{\lambda}$ and $\hat{\alpha}_{\lambda}$ the final estimated parameters by REMixed algorithm with penalty parameter λ .

2.2.5 Implementation in the REMixed R package

The method was implemented in R (R Core Team, 2024) using the `lixoftConnectors` (Lixoft, 2024a) and `marqLevAlg` (Philipps *et al.*, 2021) packages. The package `REMixed` is available on CRAN.

After selection of the penalty parameter λ , a final SAEM estimation is performed on the selected subset of biomarkers Z to prevent the shrinkage of regularization parameters toward 0. Statistical test are also proposed to control the false proportion rate using Wald test (Wald, 1943).

For the implementation of the SAEM algorithm, we rely on Monolix software (Kuhn et Lavielle, 2005; Lixoft, 2024b). For log-likelihood computation, Monolix is used to estimate the posterior distribution of the random effects through a Monte Carlo Markov Chain algorithm in order to adapt the quadrature points, through the `fastGHQuad` (Blocker, 2022) package.

3 Simulation Studies

We evaluated the performance of the proposed REMixed method using simulation studies that reflect the complexity of mechanistic models with large-dimensional markers. We consider well specified model.

3.1 Simulation Design : simulation of the humoral immune response

We modeled the antibody production process for $N = 25$ individuals from 7 day following vaccination, when antibody-secreting cell (ASC) reach a maximum level and only decay after (Pasin *et al.*, 2019). Each individual is characterized by an ASC compartment S_i , which produces antibodies at a rate φ_{S_i} and decays at rate δ_{S_i} . Antibodies Ab_i decay at rate δ_{Ab_i} . The parameters follow a log-normal distribution with random effects of variance $\omega_{\varphi_S}^2$, $\omega_{\delta_S}^2$, and $\omega_{\delta_{Ab}}^2$, respectively. The model is governed by the following ODE system :

$$\begin{cases} \frac{dS_i}{dt}(t) &= -\delta_{S_i}S_i(t), \\ \frac{dAb_i}{dt}(t) &= \varphi_{S_i}S_i(t) - \delta_{Ab_i}Ab_i(t), \end{cases}$$

with individual parameters

$$\begin{cases} \log(\varphi_{S_i}) &= \log(\varphi_{S_{pop}}) + \eta_i^{\varphi_S}, \\ \log(\delta_{S_i}) &= \log(\delta_{S_{pop}}) + \eta_i^{\delta_S}, \\ \log(\delta_{Ab_i}) &= \log(\delta_{Ab_{pop}}) + \eta_i^{\delta_{Ab}}, \end{cases}$$

where $(\eta_i^{\varphi_S})_{i \leq N} \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\varphi_S}^2)$, $(\eta_i^{\delta_S})_{i \leq N} \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\delta_S}^2)$, and $(\eta_i^{\delta_{Ab}})_{i \leq N} \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\delta_{Ab}}^2)$.

Antibody levels are measured at clinical time points (days 0, 7, 21, 123, 180, 300) from 7 days after vaccination, consistent with COVID-19 vaccine trials (Rinchai *et al.*, 2022). The latent compartment $S_i(t)$ is measured from K_1 biomarkers collected daily from day 1 to day 21 post-vaccination. We also generate K_2 noise biomarkers. The generating measurement equations are, for individual $i \leq 25$:

$$\begin{cases} Y_{ij} = \log_{10}(Ab_i(t_j)) + \epsilon_{ij}, & \forall j \leq 6, \\ Z_{kij} = \mu_k + \alpha_k \times S_i(t_{kj}) + \epsilon_{kij}, & \forall j \leq 21, k \leq K_1, \\ Z_{kij} = \mu_k + \epsilon_{kij}, & \forall j \leq 21, K_1 < k \leq K = K_1 + K_2, \end{cases}$$

where $\epsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{Ab}^2)$ and $\epsilon_{kij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$. We simulated two scenarios differing in biomarker dimensionality : one with $K = 20$ genes ($K_1 = 5$ relevant biomarkers) and another with $K = 50$ genes ($K_1 = 10$ relevant biomarkers). Further details, including the exact distributions of parameters and noise levels, are provided in the Supplementary Material B.1.3. The parameter values model are detailed in the Supplementary Material B.1.3.

3.2 Strategy of Estimation

We evaluated the performance of our proposed method using simulation studies under two scenarios reflecting different levels of complexity using 200 replicates. For the 20-gene scenario, a grid of 15 penalty parameters was constructed, and the final penalty parameter used for each replicate was selected by minimizing the BICc. For the 50-gene scenario, to reduce computation time, the penalty parameter was estimated by constructing the grid over 10 replicates only and averaging the penalty values selected via BICc minimization as the optimal penalty parameter for every replicates. In the initialization step, the mechanistic model was estimated by considering random pairs of potentially relevant genes (α_k estimated), while the others being considered as noise (α_0 fixed at 0), as proposed in 2.2.3.

3.3 Evaluation metrics

Performance metrics, including the False Proportion Rate (FPR), False Negative Rate (FNR), defined as follows :

$$FPR = \frac{FP}{FP + TN},$$

$$FNR = \frac{FN}{FN + TP},$$

were computed to assess the variable selection accuracy. To assess the effectiveness of our approach, we compared the FPR and FNR alongside the F_1 -score, which balances precision and recall :

$$F_1 = \frac{2TP}{2TP + FN + FP}.$$

This metric provides a comprehensive evaluation of the trade-off between false positives and false negatives.

Parameter estimation accuracy was assessed by comparing the estimated parameters to their true simulated values using the following metrics : percentage of relative bias, empirical and estimated standard deviations, which are expected to be close, and the coverage probability of the 95% confidence interval, which should ideally reach the nominal 95% level. Explicit formulas for these metrics are provided in Supplementary Material [B.1.4](#).

3.4 Simulation Results

3.4.1 Initialisation Strategy

As shown in Figure [4.2](#), which presents results for the 50-gene scenario, the initialization strategy consistently yielded higher likelihood values for gene pairs that included a significant gene (i.e. genes 1 to 10). The difference in likelihood between the best and worst pairs reached 7000 points. In the example shown in the bottom panel of the figure, a difference of over 4000 points is observed between the top-ranked group and the second one, highlighting a substantial improvement. Similar results were observed in the 20-gene scenario (see Supplementary Figure [5.18](#)). This initialization procedure is computationally efficient, typically taking less than 30 minutes for the 20-gene scenario, with a mean time of 26 minutes, and less than 160 minutes for the 50-gene scenario, with a mean time of 142 minutes.

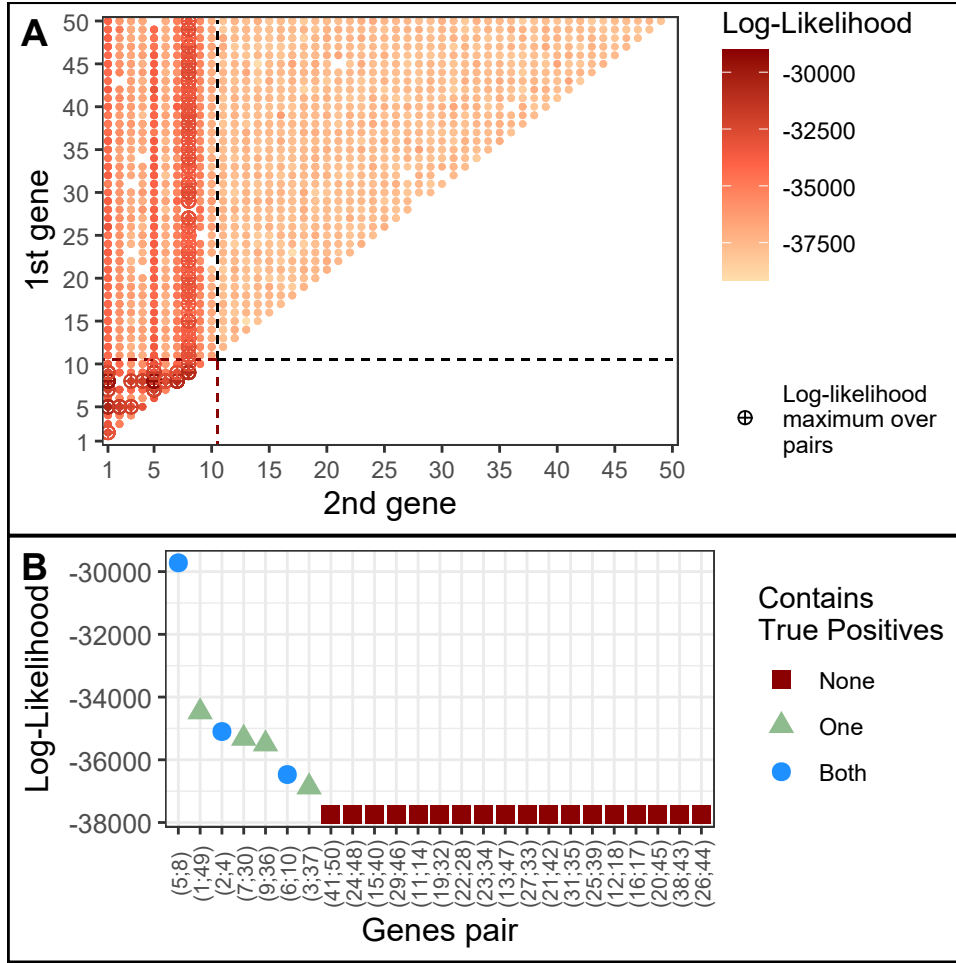


FIGURE 4.2 – (A) Maximum log-likelihood of the mechanistic model with 50 genes but only considering a pair of them as relevant throughout the 200 replicates. Best pairs selected over all 25 pairs for each replicates tested is marked by a rounded cross. The dashed lines delimit the set of significant genes, range from 1 to 10 on each axis. (B) Initialization strategy results for the first replicate of the 50-gene scenario.

3.4.2 Selection accuracy

Figure 4.3 summarizes the selection performance across scenarios and compares the REMixed method with and without the Wald test enhancement.

Firstly, in the 20-gene scenario, the median FPR was 14% [0% ; 33.3%], with a median F_1 -score of 83.9% [66.7% ; 100%]. In the 50-gene scenario, the median FPR was slightly lower at 10.1% [2.4% ; 20.0%], and the median F_1 -score was 83.7% [71.4% ; 95.4%]. In both cases, the median FNR remained at 0%, indicating strong control over false negatives.

After applying the Wald test, the median FPR improved to 5.3% [0% ; 20.0%] for the 20-gene scenario and to 5.0% [0% ; 12.5%] for the 50-gene scenario. The F_1 -score also increased, reaching 93.2% [76.9% ; 100.0%] and 91.2% [80.0% ; 100%], respectively. In both settings, no false negatives were detected across all replicates. While the number of false positives increased with dimensionality, it was substantially reduced after applying the Wald test, as illustrated in

panel A of Figure 4.3.

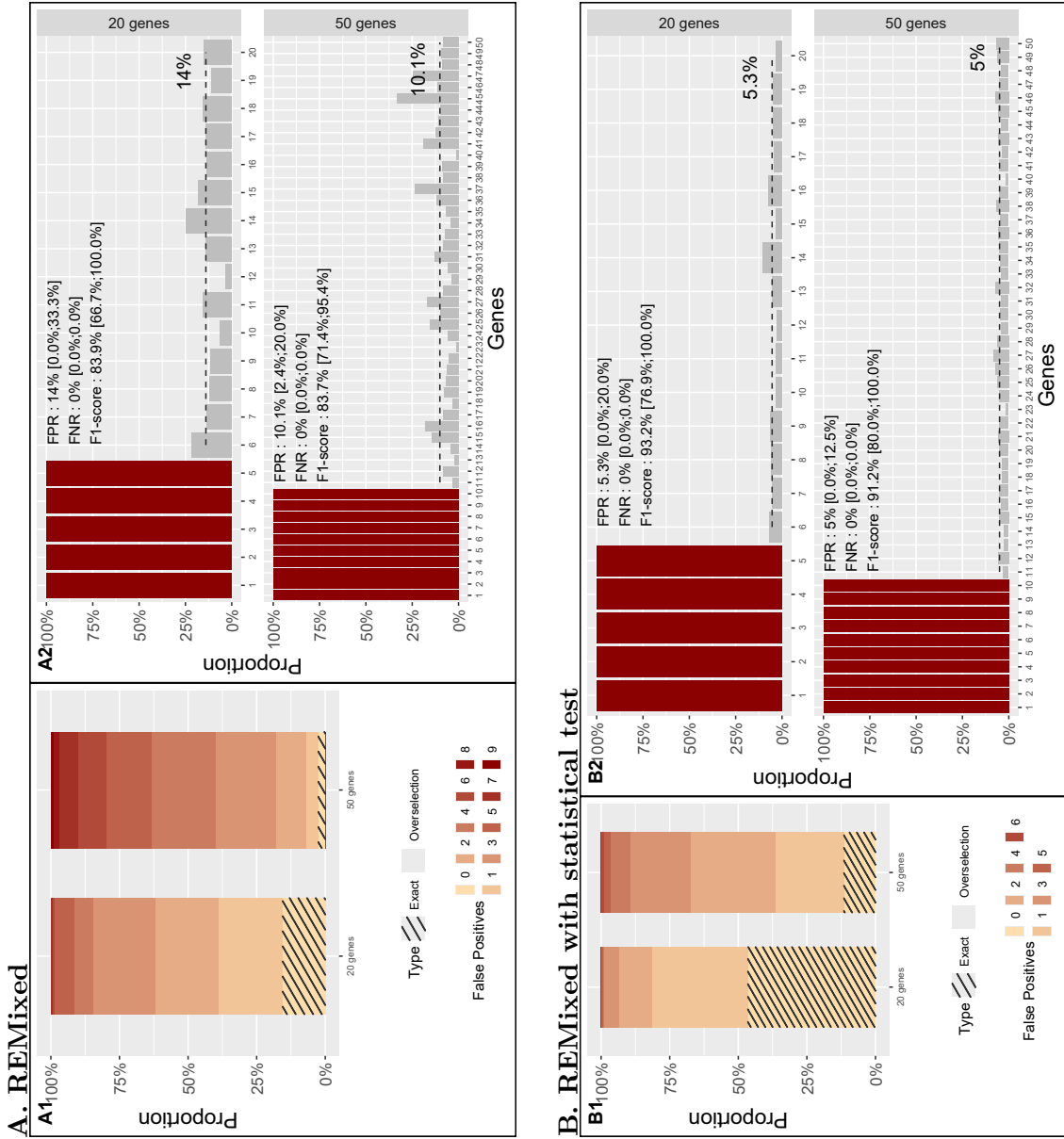


FIGURE 4.3 – Selection results by REMixed algorithm (panel A) and REMixed algorithm with final statistical test (panel B). (1) Number of exact models (no false positives or negatives) and models containing the exact significant biomarker set (no false negatives) across 200 replicates for 20- (A1) and 50-genes scenarios (B1). (2) Selection frequency of each biomarker across 200 replicates for 20-gene (top) and 50-gene (bottom) scenarios with variable selection median metrics (FPR, FNR, F1-score) with 95% interquartiles interval, across 200 replicates for 20- (A2) and 50-gene scenarios (B2).

3.4.3 Parameter Estimation

Table 4.3 shows that in the 20-gene scenario, REMixed accurately recovered both population parameters (δ_{Ab} , δ_S , φ_S and associated variances) and biomarker-specific parameters (μ_k , α_k , σ_k). Relative biases remained below 5%, and 95% coverage probabilities closely matched nominal levels

(median 94%-95%). Results for the 50-gene scenario are similar and reported in Supplementary Material [B.1.5](#).

Supplementary Table [5.9](#) shows that these good estimation properties were preserved after applying the Wald test. The same set of parameters was accurately recovered, with no substantial degradation in relative bias or coverage. Similar observations hold in the 50-gene case (see Supplementary Material [B.1.5](#)).

Tables [4.4](#) present results when non-significant biomarkers were included in the candidate set. For both REMixed alone and with the Wald test detailed in Supplementary Table [5.10](#), non-significant biomarkers were consistently estimated with low effects and selected in fewer than 25% of replicates. This demonstrates robustness against irrelevant biomarker inclusion. The same pattern was observed for the 50-gene scenario (see Supplementary Material [B.1.5](#)).

For the 20-gene scenario, computation times were recorded for the final value of the penalty parameter λ . REMixed algorithm converged on average in 7.5 iterations (with a 95% interquantile range (IQR) between 6 and 8) and the mean time per iteration was 5.0 minutes (IQR between 2.8 and 8.0 minutes). These results show high variability, due to the strong dependence between computation time and the number of genes included in the model : estimations tend to be faster when fewer genes are selected, and slower when multiple genes are included. For the 50-gene scenario, computation time increased, with 9.6 iterations on average (IQR between 9 and 11), and an average of 21.9 minutes per iteration (IQR between 12.0 and 30.9 minutes).

Parameter	Target Value	Mean estimate	Relative Bias (%)	Empirical SD	Estimated SD	95% Coverage
δ_{Ab}	0.025	0.025	0.50	0.0017	0.0015	92.0
δ_S	0.230	0.230	-0.16	0.0071	0.0063	91.5
$\varphi_S \times 10^{-2}$	6.110	6.150	0.71	0.0077	0.0075	91.0
$\omega_{\delta_{Ab}}$	0.300	0.288	-3.90	0.0396	0.0419	93.0
ω_{δ_S}	0.100	0.098	-1.98	0.0184	0.0164	92.5
ω_{φ_S}	0.600	0.573	-4.42	0.1030	0.0903	87.0
α_1	0.650	0.649	-0.15	0.0131	0.0128	93.5
α_2	-0.490	-0.489	-0.22	0.0116	0.0126	96.5
α_3	1.060	1.060	-0.12	0.0149	0.0151	93.5
α_4	-0.720	-0.719	-0.18	0.0140	0.0145	96.0
α_5	0.330	0.329	-0.17	0.0156	0.0148	91.0
μ_1	-0.610	-0.609	-0.14	0.0229	0.0196	91.5
μ_2	0.040	0.041	1.76	0.0189	0.0188	96.0
μ_3	-1.800	-1.800	0.07	0.0253	0.0244	94.5
μ_4	2.400	2.400	-0.04	0.0235	0.0222	95.5
μ_5	1.090	1.090	-0.03	0.0213	0.0214	96.0
σ_{Ab}	0.110	0.110	-0.17	0.0081	0.0077	93.0
σ_1	0.310	0.309	-0.42	0.0087	0.0096	97.0
σ_2	0.310	0.309	-0.17	0.0093	0.0096	96.0
σ_3	0.350	0.348	-0.66	0.0109	0.0109	95.0
σ_4	0.350	0.351	0.16	0.0119	0.0109	94.5
σ_5	0.370	0.368	-0.42	0.0114	0.0114	97.5

TABLE 4.3 – Parameter estimation accuracy for population and biomarker-specific parameters using REMixed algorithm (without additional test) over 200 replicates using samples of 25 subjects and 20 genes.

Gene	Parameter	Target Value	Mean Estimate	Empirical SD	Estimated SD	Selection Frequency (%)
G6	μ_6	-1.660	-1.657	0.0225	0.0147	21.5
	α_6	0.000	-0.004	0.0183	0.0103	
	α_6	0.260	0.258	0.0071	0.0080	
G7	μ_7	-0.160	-0.157	0.0230	0.0167	13.5
	α_7	0.000	-0.005	0.0227	0.0117	
	α_7	0.300	0.294	0.0093	0.0091	
G8	μ_8	-0.210	-0.215	0.0265	0.0159	12.0
	α_8	0.000	0.005	0.0218	0.0111	
	α_8	0.280	0.279	0.0059	0.0086	
G9	μ_9	-0.960	-0.960	0.0252	0.0169	11.5
	α_9	0.000	-0.003	0.0257	0.0118	
	α_9	0.300	0.297	0.0079	0.0092	
G10	μ_{10}	-0.520	-0.531	0.0255	0.0208	6.5
	α_{10}	0.000	0.011	0.0264	0.0144	
	α_{10}	0.360	0.362	0.0151	0.0112	
G11	μ_{11}	0.620	0.623	0.0173	0.0142	15.5
	α_{11}	0.000	-0.003	0.0166	0.0099	
	α_{11}	0.250	0.249	0.0078	0.0077	
G12	μ_{12}	0.800	0.799	0.0363	0.0239	3.5
	α_{12}	0.000	0.016	0.0443	0.0167	
	α_{12}	0.420	0.419	0.0130	0.0129	
G13	μ_{13}	0.070	0.069	0.0259	0.0163	14.0
	α_{13}	0.000	0.003	0.0221	0.0114	
	α_{13}	0.290	0.287	0.0091	0.0088	
G14	μ_{14}	0.130	0.129	0.0205	0.0130	24.5
	α_{14}	0.000	-0.001	0.0181	0.0091	
	α_{14}	0.230	0.228	0.0065	0.0070	
G15	μ_{15}	-0.180	-0.179	0.0189	0.0141	18.0
	α_{15}	0.000	0.002	0.0163	0.0098	
	α_{15}	0.250	0.248	0.0073	0.0077	
G16	μ_{16}	-0.170	-0.181	0.0227	0.0175	13.5
	α_{16}	0.000	0.012	0.0242	0.0122	
	α_{16}	0.310	0.307	0.0091	0.0095	
G17	μ_{17}	-0.450	-0.450	0.0263	0.0163	14.0
	α_{17}	0.000	0.001	0.0213	0.0115	
	α_{17}	0.290	0.288	0.0095	0.0089	
G18	μ_{18}	0.500	0.497	0.0250	0.0159	15.5
	α_{18}	0.000	0.004	0.0203	0.0111	
	α_{18}	0.280	0.279	0.0074	0.0086	
G19	μ_{19}	-2.270	-2.270	0.0283	0.0173	11.0
	α_{19}	0.000	-0.001	0.0256	0.0121	
	α_{19}	0.310	0.305	0.0094	0.0094	
G20	μ_{20}	-2.020	-2.022	0.0213	0.0153	15.0
	α_{20}	0.000	0.004	0.0181	0.0107	
	α_{20}	0.270	0.269	0.0080	0.0083	

TABLE 4.4 – Parameter estimation accuracy for the irrelevant biomarker (Gaussian random noise) using REMixed algorithm (without additional test) over 200 replicates using samples of 25 subjects and 20 genes.

4 Application to COVID-19 mRNA vaccination data

We illustrate our statistical method using transcriptomic data from a clinical trial assessing immune responses to the Pfizer/BioNTech COVID-19 mRNA vaccine (BNT162b2). The original dataset, collected by [Rinchai et al. \(2022\)](#), consists of 23 healthy volunteers, of whom 17 received two doses of the Pfizer vaccine at day 0 and 21. Blood samples were collected daily, beginning just before each vaccine dose and continuing for nine consecutive days thereafter, as well as 14 days after, allowing high-resolution monitoring of gene expression profiles over time. Additionally, serological samples were obtained at baseline and at 7 and 14 days following each vaccination to quantify the humoral response via SARS-CoV-2-specific antibodies. This analysis was restricted to the 15 subjects who have not been previously infected with the virus.

The datasets contains 8 172 gene expressions regrouped in 34 pathways defined by Biomart ([Durinck et al., 2005](#)) and [Li et al. \(2014\)](#). For illustrative purpose, we restricted the analysis to one representative variable of each pathway, defined as the average of the pathway-specific expression levels. The individual observed trajectories of antibody and 34 transcriptomics representative are displayed in Supplementary Figures 5.20 and 5.19.

The mechanistic model describing the humoral response follows previous works by ([Clairon et al., 2023b](#)) ([Balelli et al., 2020](#)) :

$$\left\{ \begin{array}{l} \frac{dS_i}{dt} = \rho_i^{(d)} e^{-\delta_V(t-t_d)} - k_m S_i, \\ \frac{dF_{1i}}{dt} = k_m S_i - k_m F_{1i}, \\ \vdots \\ \frac{dF_{6i}}{dt} = k_m F_{5i} - \delta_S F_{6i}, \\ \frac{dA_{bi}}{dt} = \theta_i F_{6i} - \delta_{Ab} A_{bi}, \end{array} \right.$$

with $d = 1, 2$, $t_1 = 0$, and $t_2 = 21$. For each individual $i = 1, \dots, 15$, the individual parameters $\rho^{(2)}$ and θ were defined using the following transformation, in order to capture inter-individual variability observed in the data :

$$\left\{ \begin{array}{l} \log(\rho_i^{(2)}) = \log(\rho_0^{(2)}) + \eta_i^\rho, \\ \log(\theta_i) = \log(\theta_0) + \eta_i^\theta, \end{array} \right.$$

with random effects $\eta_i^\rho \stackrel{iid}{\sim} \mathcal{N}(0, \omega_\rho^2)$ and $\eta_i^\theta \stackrel{iid}{\sim} \mathcal{N}(0, \omega_\theta^2)$. Specifically, it models antibody production according to the two doses at times t_d , $d = 1, 2$, of the BNT162b2 vaccine as an antigen denoted V , characterized by its decay rate δ_V . This antigen produces ASCs S at rate $\rho^{(d)}$, $d = 1, 2$, with $\rho^{(1)} = 1$ by definition of the model [Clairon et al. \(2023b\)](#). To account for the observed delay in the onset of the humoral response between vaccine injection and the increase in antibody titers, resulting from ASCs maturation, we introduced a delay compartment between the generation of S cells and the production of antibodies Ab ([Prague et al., 2025](#)). Antibodies

are then produced at rate θ and decay at rate δ_{Ab} .

In our analysis, we fixed the parameters $\delta_{Ab} = 0.03$ and $\delta_S = 0.1$ according to previous studies (Clairon *et al.*, 2023b; Prague *et al.*, 2025). The parameter δ_V was fixed to $0.46, \text{days}^{-1}$, corresponding to a vaccine dose half-life of 1.5 days. This implies that after 5 days the antigen has essentially disappeared, consistent with its role in driving ASC production. Based on the analysis and estimation by Prague *et al.* (2025), ASCs secrete antibodies Ab after a six-step maturation process with an average duration of $1/k_m = 4$.

In this analysis, we aim to link the pathway representatives to the ASC production term $\rho^{(d)} e^{-\delta_V(t-t_{k-1})}$, introducing a one-day delay. This delay accounts for the fact that gene expression is measured on the day of vaccination, prior to injection, so that any variation becomes observable only on the following day. The underlying assumption is that the observed pathway dynamics may provide information on the dynamics of ASC production, and we seek to identify which pathways are most informative in this context. As a consequence, we consider the following observation models, for each dose $d = 1, 2$:

$$\begin{cases} Y_{ij} = \log(Ab_i(t_{ij})) + \epsilon_{ij}, & t_{ij} = \{0, 7, 14, 21, 28, 35\}, \\ Z_{kij} = \mu_k + \alpha_k \times \rho^{(d)} e^{-\delta_V(t_{kij}-t_k-1)} + \varepsilon_{kij}, & k = 1, \dots, 34, \quad t_{kij} = \{0, \dots, 9, 14, 21, \dots, 30, 35\}, \end{cases}$$

with $\epsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{Ab}^2)$ and $\varepsilon_{kij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$.

Initialization was performed as proposed in Section 4.3. Using 10 different random seeds, we created 17 groups, each containing 2 pathways. The results are displayed in Figure 4.4. The best initialization point was obtained with the Monocytes-Interferon pair, with a total log-likelihood near -14,278, about 10 points above the second-best group. We also observed that the Interferon and Type I Interferon pathways, when included, yielded higher total log-likelihoods than other groups. Indeed, the 18 best total log-likelihoods were obtained from groups containing one of these two pathways, ranging from -14,278 to -14,327, whereas other groups achieved total log-likelihoods below -14,373.

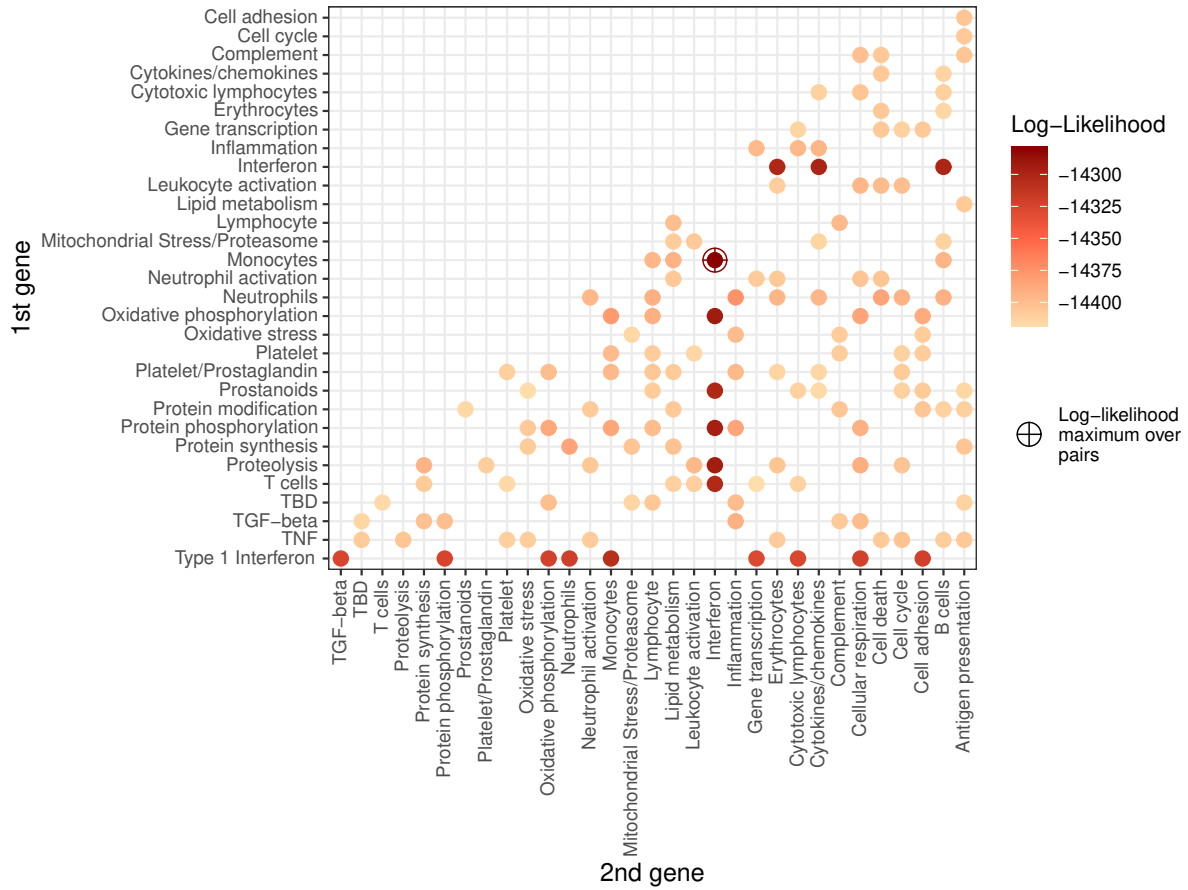


FIGURE 4.4 – Maximum log-likelihood estimates of the mechanistic model for the real-data application (Rinchai *et al.*, 2022) including the first and second genes in the observation model only. Best pairs selected over all pairs tests is marked by a rounded cross.

We then ran the REMixed algorithm starting from this initialization point, over a grid of 30 penalty parameters ranging from 2.5 to 350. Above $\lambda = 250$, all the biomarkers were removed from the selected set, leading to an unidentifiable model. We thus retained the 20 first estimation for $\lambda < 250$. The calibration plot is displayed in Figure 4.5. The minimal BICc was reached for the penalty parameter $\lambda^* = 25$, with $BICc^* = 29,015$. Additional information concerning the tuning of the convergence criteria τ_1 and τ_2 and the number of points used for the Gauss-Hermite quadrature are provided in Supplementary Material B.1.6.

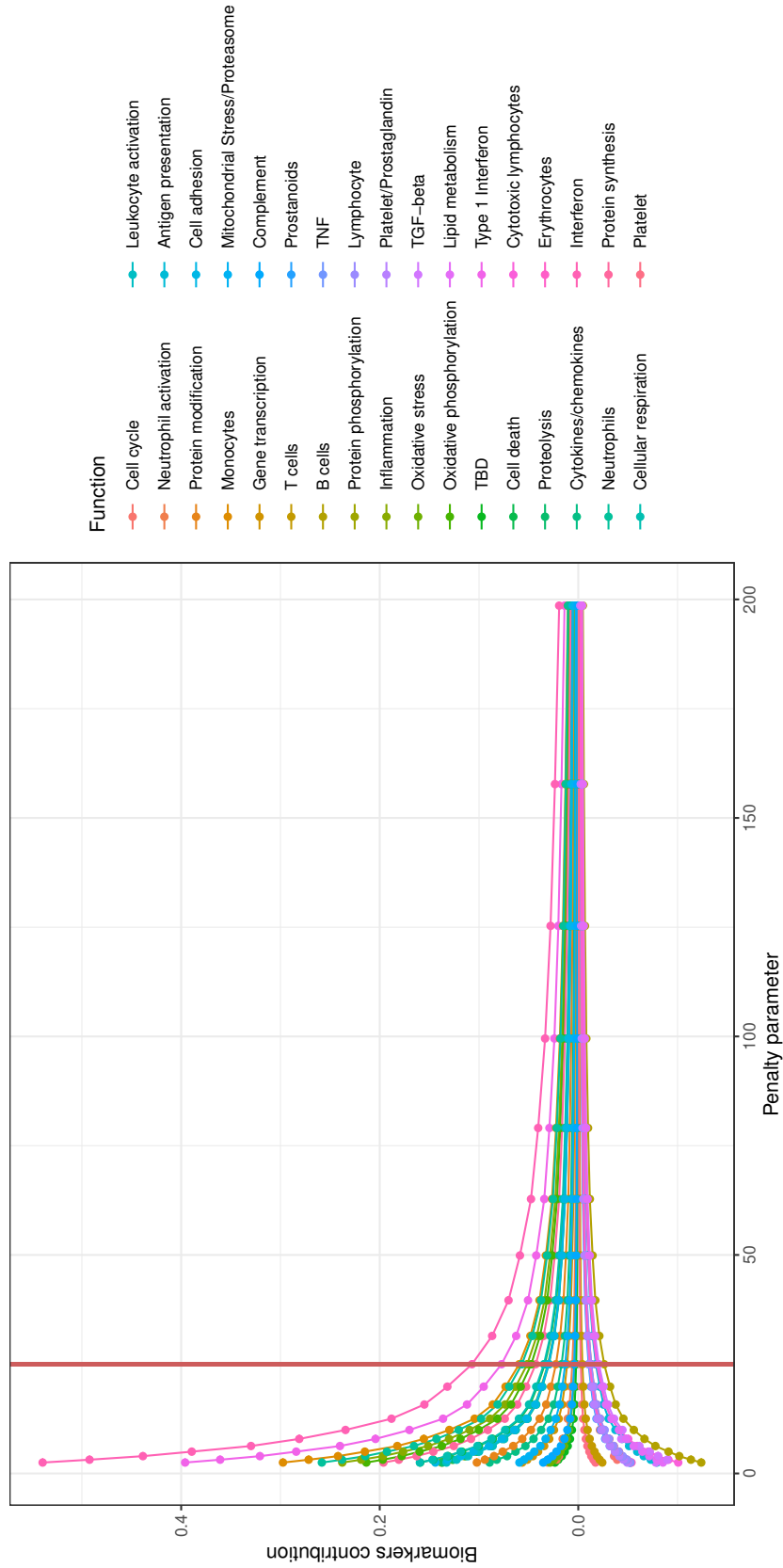


FIGURE 4.5 – Final biomarkers contributions estimates by REMixed algorithm, launched along a grid of penalty parameters, for the real-data application (Rinchai *et al.*, 2022). The red lines indicates the penalty parameter achieving the best BICc ($\lambda^* = 25$, with BICc of 29 015).

In the end, all the pathways were retained along the calibration path. The Interferon and Type I Interferon pathways exhibited the strongest standardized biomarker contribution, with $\alpha_k = 0.105$ and 0.076 , respectively. All other pathway contributions remained small, ranging from -0.026 to 0.058 . Regarding biomarker variability, the estimated standard deviations σ_k were consistently high (0.775 - 1.448), before and after re-estimation by the SAEM algorithm on the final selected set of pathway representatives, reflecting the substantial level of noise in the data.

After the final Wald tests, 21 pathways were retained, including Interferon, Type I Interferon, Monocytes, Neutrophils, Inflammation, Oxidative Phosphorylation, Protein Synthesis, Cellular Respiration, Cell Death, Protein Phosphorylation, Cell Adhesion, Leukocyte Activation, Proteolysis, Protein Modification, Mitochondrial Stress/Proteasome, TNF Signaling, Lymphocytes, Antigen Presentation, Lipid Metabolism, TGF-beta, and B cells. Notably, a large part of the selected pathways are classically involved in immune responses (e.g. type I interferon, inflammation, antigen presentation, B cells), as well as to cell activation and replication processes, which may explain their selection and thereby providing support for their selection (Thiébaud *et al.*, 2019; Hagan *et al.*, 2022). Parameter values obtained throughout all steps of this illustration are reported in Supplementary Material B.1.6.

5 Discussion

In this work, we proposed REMixed, a novel methodology for the integration of large-dimensional biomarker data into non-linear mixed effects models based on systems of ordinary differential equations. Our approach extends existing regularized estimation frameworks -traditionally applied to generalized linear or joint models- to more mechanistic models without necessarily closed-form solution and allows regularization on the repeated measures of the underlying compartments rather than their time-fixed prediction. By developping a proximal coordinate descent algorithm to maximize the penalized log-likelihood of such model, REMixed enables the simultaneous estimation of model parameters and selection of biomarkers dynamically linked to latent biological compartments.

Simulation results supported the relevance of this framework. In large-dimensional settings, REMixed was able to recover the true set of biomarkers, although false positives were also selected. Moreover, parameter estimation remained accurate and stable across replicates, as shown by low bias and appropriate coverage rates. The addition of a post-selection inference step via Wald tests further reduced the false discovery rate. In our simulation study, however, the scenarios considered were rather favorable, with a relatively high signal-to-noise ratio. We relied in our work on the BICc to calibrate the penalty parameter. Although it was found well-suited for NLMEMs (Delattre *et al.*, 2014), our selection procedure based on BICc did not exhibit fully satisfactory performance, with some overselection remaining. Alternative calibration procedures of the penalty parameter could be explored, including cross-validation, although standard cross-validation is known to over-select (Bühlmann comments on Tibshirani (2011)), and consistency in model selection requires more careful schemes as demonstrated by Feng

et Yu (2019). Such refinements could mitigate the overfitting observed here, but they would be computationally prohibitive with REMixed’s current implementation. In this framework, we relied instead on post-selection Wald tests. These performed satisfactorily under our high signal-to-noise simulation scenarios, but this remains a limitation of this work since they do not account for the post-selection inference context.

Moreover, several additional advantages of REMixed arise from its mixed model structure combined with maximum likelihood theory. Notably, the method handles missing data under the Missing At Random assumption. It also accommodates variations in measurement times across subjects and biomarkers, thereby enhancing flexibility and applicability in realistic longitudinal data settings.

We present an application to real SARS-CoV-2 vaccination data. The results underscore the limitations of our method in terms of the realism of the observation model considered. Indeed, assuming constant biomarker effects μ_k throughout time may negatively impact the applicability of the method in real-data context. As a consequences, replacing the constant biomarker effects μ_k in our model with flexible temporal structures -e.g., time-varying coefficients- could better capture the dynamic nature of gene expression. Such refinements would align the model more closely with biological reality, particularly in contexts where gene responses are transient or delayed. Moreover, the current error structure assumes independent and homoscedastic Gaussian noise, for all markers. Future developments may involve incorporating more complex error models, such as temporally correlated noise or heteroscedastic variance structures. These enhancements would increase robustness and reflect the complex variability observed in omics data. Finally, while the current REMixed implementation focuses on linking transcriptomic biomarkers to a single latent compartment through constant effects over time, several natural extensions are worth exploring. First, incorporating multiple latent compartments into the regularized inference framework would allow the identification of biomarkers associated with distinct and potentially interacting biological processes. All these extensions would improve REMixed’s applicability, and could be further assess by additional simulation scenarios.

Additional initialization strategies could also be explored. In our approach, genes were grouped in pairs to balance computational cost and the feasibility of SAEM estimation. However, adding genes individually or in groups larger than two could also be considered. Moreover, for models where the directly observed compartment - such as the antibody compartment - is practically identifiable from the measurements alone, direct estimates could be used to initialize the algorithm without relying on additional genes. A robustness analysis could also be conducted to assess the sensitivity of the method to the choice of initialization.

It is important to highlight a methodological limitation regarding dimensionality. Our simulations considered up to 50 biomarkers, which, although substantially larger than typical longitudinal modeling scenarios (with biomarker count K exceeding the number of individuals N but still lower than the total number of observations), does not represent classical high-dimensional cases involving thousands of biomarkers. In such scenarios, an initial preprocessing step to select a manageable subset of markers would be necessary. Indeed, a preprocessing step

was applied in our real-data application.

In conclusion, REMixed provides a general and efficient framework for integrating large-dimensional longitudinal biomarkers into mechanistic models with latent dynamics. It fills a methodological gap between systems biology and statistical learning, and is a first step towards the identification of transcriptomic signatures of biological processes that are only partially observed. Its extensions to more elaborate temporal structures and latent architectures constitute promising directions for future work.

REMixed : Regularized Estimation for non-linear Mixed effects models in the presence of large-dimension biomarkers in R

Auriane Gabaut^{1,2,3}, Cécile Proust-Lima^{1†}, Mélanie Prague^{1,2,3*†}

¹ : Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France

² : Univ. Bordeaux, Inria, Bordeaux, France

³ : Vaccine Research Institute, Paris, France

* : melanie.prague@inria.fr

† : These authors contributed equally

Abstract : The R package **REMixed** (Regularized Estimation for non-linear Mixed Effects Models) implements an iterative algorithm to estimate population parameters in non-linear mixed effects models (NLMEM) based on a structural mechanistic system and simultaneously regularizes observed biomarkers linked linearly to latent compartments of the mechanistic model. The algorithm uses a cyclical coordinate descent approach, alternating between regularization of the biomarker-compartment parameters through Lasso-penalized log-likelihood maximization and estimation of the remaining parameters through maximum likelihood estimation. The package relies on the Monolix software, which implements estimation methods for NLMEMs based on ordinary differential equations (ODEs). The package is illustrated on a simulated dataset of the humoral response after Ebola virus vaccination strategy using transcriptomics and viral load.

Acknowledgment

This study was carried out in the framework of the University of Bordeaux's France 2030 program RRI PHDS (*Public Health Data Science*). Elements of this work have been carried out as part of the SOLVE project which has received funding from the European Union's Horizon Research and Innovation program under Grant Agreement No. 101137185. Computer time for this study was provided by the computing facilities of the PlaFRIM (Plateforme Fédérative pour la Recherche en Informatique et Mathématiques).

We thank Simulations Plus, Lixoft division for the free academic use of the MonolixSuite. We disclose that generative AI tools were used to polish English language.

4.4 Introduction

Mechanistic models based on ordinary differential equations (ODEs) are widely used to describe biological systems, particularly in fields such as pharmacometrics, immunology, and epidemic modeling (Ibarra *et al.*, 2021; Perelson et Ribeiro, 2018; Ganser *et al.*, 2024). These

models capture the temporal dynamics of multiple compartments, both observed and unobserved, involved in biological processes such as immune responses. However, achieving identifiability often requires repeated measurements across all compartments, which is typically not feasible in clinical studies (Lavielle, 2014).

In parallel, diverse longitudinal data such as transcriptomic profiles, increasingly accessible through high-throughput technologies (Lowe *et al.*, 2017), provide indirect but rich information on underlying biological processes. This has led to a growing interest in incorporating large-dimensional longitudinal markers into statistical models to improve the estimation of such mechanisms, with encouraging results reported in semi-parametric frameworks, mixed-effects models, and joint models with survival outcomes (Ni *et al.*, 2010; Zhao *et al.*, 2014; Pan et Huang, 2014; Li *et al.*, 2018; Sun et Basu, 2024). Many of these approaches rely on iterative estimation and variable selection strategies, demonstrating robustness in complex settings (Pan et Huang, 2014; Li *et al.*, 2018; Caillebotte *et al.*, 2024). Previous work has also demonstrated the feasibility of integrating transcriptomic data into non-linear mixed-effects models (NLMEMs). However, existing contributions remain limited to time-fixed covariates and do not capture dynamics through ODEs (Caillebotte *et al.*, 2024; Gabaut *et al.*, 2025b). The REMixed algorithm proposes a new approach that combines the estimation of ODE-based NLMEMs with the selection of time-varying longitudinal biomarkers linearly linked to one latent compartment of the ODE system Gabaut *et al.* (2025a).

We developed the R package **REMixed**, available on CRAN, to provide a flexible implementation of this methodology. The package offers tools for estimating ODE-based NLMEMs with longitudinal outcomes while simultaneously selecting transcriptomic markers dynamically associated with one latent biological state from the same mechanistic model. It relies on the **Monolix software** (Lixoft, 2024b), which is accessible in R via the **lixoftConnectors** package and is freely available for academic use.

The proposed algorithm is an iterative procedure designed to solve Lasso-penalized maximum likelihood estimation (Tibshirani, 1996), implemented via cyclical coordinate descent (Tseng, 2001; Friedman *et al.*, 2010b). This strategy, previously applied in large-dimensional joint models for survival and mixed effect (Caillebotte *et al.*, 2024), as well as for the selection and estimation fixed or random effects (Pan et Huang, 2014; Li *et al.*, 2018), has shown strong performance in similar frameworks. The algorithm alternates between two steps : a regularization step and a mechanistic inference step. In the regularization step, coefficients linking biomarkers to latent compartments are updated by solving the first-order optimality conditions of the penalized log-likelihood, using second-order Taylor approximations. In the mechanistic inference step, model parameters are estimated with the Stochastic Approximation Expectation-Maximization (SAEM) algorithm (Dempster *et al.*, 1977; Delyon *et al.*, 1999), as implemented in Monolix (Kuhn et Lavielle, 2005; Lixoft, 2024b).

In this vignette, we first describe the REMixed algorithm in Section 2. We then outline its implementation in the **REMixed** package and present the code structure and its integration with Monolix in Section 3. Finally, we illustrate the use of the main functions on a simulated

dataset of humoral immune response following Ebola vaccination (Pasin *et al.*, 2019) in Section 4.

2 REMixed algorithm

2.1 ODE-based NLMEMs

2.1.1 Model definition

Let us consider a mechanistic model defined by a system of ODEs involving at least $P + 1$ compartments, or variables, denoted $(S_p)_{p \leq P}$ and R :

$$\begin{cases} S'_p(t) &= F_{S_p}(S_p(t), R(t), \boldsymbol{\psi}), \quad p \leq P, \\ R'(t) &= F_R(S_p(t), R(t), \boldsymbol{\psi}), \end{cases} \quad (4.6)$$

where $\boldsymbol{\psi}$ is an m -dimensional parameter vector, and F_{S_p} , $p \leq P$, and F_R are fixed parametric functions characterizing the mechanistic process. For each individual i , $i = 1, \dots, N$, the trajectories generated by this ODE system are written S_{pi} ($p \leq P$) and R_i , with individual-specific parameters $\boldsymbol{\psi}_i = (\psi_{li})_{l \leq m}$. Each parameter ψ_{li} ($l = 1, \dots, m$) is modeled through a generalized linear model :

$$h_l(\psi_{li}) = h_l(\psi_{lpop}) + \mathbf{X}_i \boldsymbol{\beta}_l + \eta_{li},$$

where $\mathbf{X}_i$ denotes the vector of individual covariates with regression coefficients $\boldsymbol{\beta}_l$, and the random effects satisfy $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$. In particular, it is possible to fix some parameters to a common value across individuals by setting the corresponding variances in $\boldsymbol{\Omega}$ to zero.

The individual compartments S_{pi} , $p \leq P$, are observed through repeated measurements Y_{pij} at time points t_{pij} , with $j \leq n_{pi}$:

$$Y_{pij} = g_p(S_{pi}(t_{pij})) + \epsilon_{pij}, \quad (4.7)$$

for a fixed transformation g_p , with independent errors $\epsilon_{pij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_p^2)$.

The latent compartment R_i is measured repeatedly through a large number of markers Z_{kij} for $k \leq K$, at times t_{kij} , $j \leq n_{ki}$, via a linear observation model :

$$Z_{kij} = \mu_k + \alpha_k s_k(R_i(t_{kij})) + \varepsilon_{kij}, \quad (4.8)$$

with fixed transformation s_k and independent errors $\varepsilon_{kij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_k^2)$. We further denote the observation vectors $\mathbf{Y}_{pi} = (Y_{pij})_{j \leq n_{pi}}$ for $p = 1, \dots, P$ and $\mathbf{Z}_{ki} = (Z_{kij})_{j \leq n_{ki}}$ for $k = 1, \dots, K$.

with a fixed transformation s_k and independent measurement errors $\varepsilon_{kij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_k^2)$. We denoted later $\mathbf{Y}_{pi} = (Y_{pij})_{j \leq n_{pi}}$, $p = 1, \dots, P$, and $\mathbf{Z}_{ki} = (Z_{kij})_{j \leq n_{ki}}$, $k = 1, \dots, K$, the compartment observation vectors for individual i .

For the remainder of the manuscript, we distinguish between the large-dimensional biomarker slopes $\alpha = (\alpha_k)_{k \leq K}$, the coefficients linking markers to the latent process and subject to regularization, and the vector $\theta = (\psi_{\text{pop}}, \beta, \Omega, (\sigma_k^2)_{k \leq K}, (\varsigma_p^2)_{p \leq P}, (\mu_k)_{k \leq K})$, which gathers all remaining model parameters.

2.1.2 Log-likelihood of the model

We denote the individual observations for subject $i = 1, \dots, N$ as $\mathbf{Y}_i = (\mathbf{Y}_{pi})_{p \leq P}$ from Equation (4.7) and $\mathbf{Z}_i = (\mathbf{Z}_{ki})_{k \leq K}$ from Equation (4.8). For individual i , the contribution to the likelihood is :

$$\mathcal{L}_i(\alpha, \theta) = f(\mathbf{Y}_i, \mathbf{Z}_i \mid \alpha, \theta) = \int_{\eta_i} f(\mathbf{Y}_i, \mathbf{Z}_i \mid \alpha, \theta, \eta_i) \times f(\eta_i \mid \alpha, \theta) d\eta_i.$$

Conditionally on the individual random effects, the trajectories $(S_{pi})_{p \leq P}$ and R_i are determined by the differential system (4.6). Hence, Z_{kij} ($k = 1, \dots, K$; $j = 1, \dots, n_{ki}$) are independent with distribution $\mathcal{N}(\mu_k + \alpha_k s_k(R_i(t_{kij})), \sigma_k^2)$, and Y_{pij} ($p = 1, \dots, P$; $j = 1, \dots, n_{pi}$) are independent with distribution $\mathcal{N}(g_p(S_{pi}(t_{pij})), \varsigma_p^2)$. We then have

$$\begin{aligned} \mathcal{L}_i(\alpha, \theta) &= \int_{\eta_i} \left\{ \prod_{p \leq P} f(\mathbf{Y}_{pi} \mid \alpha, \theta, \eta_i) \right\} \times \left\{ \prod_{k \leq K} f(\mathbf{Z}_{ki} \mid \alpha, \theta, \eta_i) \right\} \times f(\eta_i \mid \alpha, \theta) d\eta_i \\ &= \int_{\eta_i} \left\{ \prod_{p \leq P} \prod_{j \leq n_{pi}} f(Y_{pij} \mid \alpha, \theta, \eta_i) \right\} \times \left\{ \prod_{k \leq K} \prod_{j \leq n_{ki}} f(Z_{kij} \mid \alpha, \theta, \eta_i) \right\} \times f(\eta_i \mid \alpha, \theta) d\eta_i \\ &= \int_{\eta_i} \left\{ \prod_{p \leq P} \prod_{j \leq n_{pi}} \frac{1}{\varsigma_p \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{pij} - g_p(S_{pi}(t_{pij}))}{\varsigma_p} \right)^2 \right] \right\} \\ &\quad \times \left\{ \prod_{k \leq K} \prod_{j \leq n_{ki}} \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))}{\sigma_k} \right)^2 \right] \right\} \\ &\quad \times \left\{ \frac{1}{(2\pi)^{\frac{m}{2}} \det \Omega^{\frac{1}{2}}} \exp \left[-\frac{1}{2} \eta_i^T \Omega^{-1} \eta_i \right] \right\} d\eta_i. \end{aligned}$$

Thus, the individual likelihood can be expressed as a product of : (1) the density of $\mathbf{Z}_i$, $f_Z(\mathbf{Z}_i \mid \alpha, \theta, \eta_i)$; (2) the density of $\mathbf{Y}_i$, $f_Y(\mathbf{Y}_i \mid \theta, \eta_i)$ (which does not depend on α); and (3) the density of the random effects $f(\eta_i \mid \theta)$. With independent individuals, the log-likelihood is :

$$LL(\alpha, \theta) = \sum_{i \leq N} \log \left(\int_{\eta_i} f_Y(\mathbf{Y}_i \mid \theta, \eta_i) f_Z(\mathbf{Z}_i \mid \alpha, \theta, \eta_i) f(\eta_i \mid \theta) d\eta_i \right). \quad (4.9)$$

The calculation of the log-likelihood and its derivatives is detailed in Section 2.3.2.

2.2 Overview of the estimation process

REMixed aims to estimate both the structural parameters θ and the sparse vector of biomarker effects α by maximizing the penalized log-likelihood under Lasso-penalization :

$$LL_{\text{pen},\lambda}(\alpha, \theta) = LL(\alpha, \theta) - \lambda \|\alpha\|_1, \quad (4.10)$$

for a given penalty $\lambda > 0$, selected by minimizing BICc (Delattre *et al.*, 2014) over a grid (see Section 2.3.4). Estimation combines penalized maximum likelihood with a cyclical coordinate descent scheme. Each iteration alternates between :

1. **Regularization step** : with θ fixed at the previous iteration estimates, update α by solving a Lasso-type problem based on a second-order Taylor approximation of the log-likelihood around the current α . This yields a closed-form update for each α_k involving the gradient $\partial_\alpha LL$, the Hessian $\partial_\alpha^2 LL$, and the penalty λ (see Section 2.3.1).
2. **Estimation step** : with α fixed at its updated value, estimate θ using the SAEM algorithm (Delyon *et al.*, 1999; Kuhn *et al.*, 2005), implemented in Monolix and accessed from R via **lixoftConnectors** (see Section 2.3.3).

Convergence is reached when the penalized log-likelihood and the parameters stabilize below user-defined tolerances τ_1 and τ_2 , respectively. The pipeline is shown in Figure 4.6, and pseudocode is given in Algorithm 1.

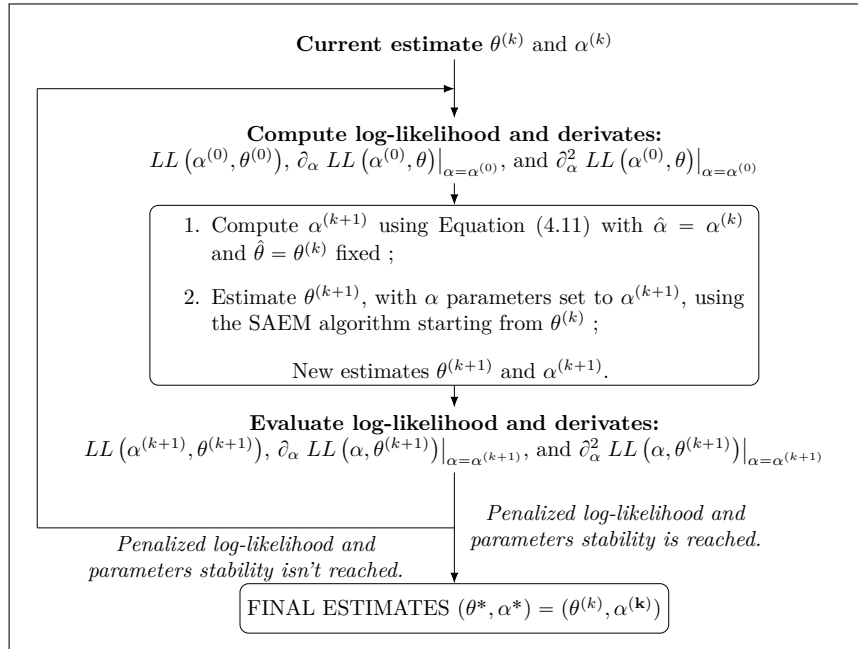


FIGURE 4.6 – Pipeline of the REMixed algorithm.

Algorithm 1 : REMixed algorithm :

Let $\lambda > 0$ be fixed ;

Obtain initial values $\alpha^{(0)}, \theta^{(0)}$, e.g., by MLE or SAEM on a random subset of biomarkers (see Section 2.3.6).

Estimate the posterior distribution of random effects via MCMC in Monolix.

Compute $LL(\alpha^{(0)}, \theta^{(0)})$, $\partial_\alpha LL(\alpha, \theta^{(0)})|_{\alpha=\alpha^{(0)}}$, and $\partial_\alpha^2 LL(\alpha, \theta^{(0)})|_{\alpha=\alpha^{(0)}}$.

At iteration l :

Compute $\alpha^{(l+1)}$ using Equation (4.11) with $\hat{\alpha} = \alpha^{(l)}$ and $\hat{\theta} = \theta^{(l)}$ fixed (Section 2.3.1).

Estimate $\theta^{(l+1)}$ with α fixed at $\alpha^{(l+1)}$ using SAEM initialized at $\theta^{(l)}$.

Update $LL(\alpha^{(l+1)}, \theta^{(l+1)})$, $\partial_\alpha LL(\alpha, \theta^{(l+1)})|_{\alpha=\alpha^{(l+1)}}$, and $\partial_\alpha^2 LL(\alpha, \theta^{(l+1)})|_{\alpha=\alpha^{(l+1)}}$ using a new posterior approximation of the random effects.

Evaluate convergence and stop if all criteria are met :

$$|LL_{\text{pen}_\lambda}(\alpha^{(l+1)}, \theta^{(l+1)}) - LL_{\text{pen}_\lambda}(\alpha^{(l)}, \theta^{(l)})| < \tau_1 ;$$

$$\left| \left(\frac{\alpha^{(l)} - \alpha^{(l+1)}}{\alpha^{(l)} + 1}, \frac{\theta^{(l)} - \theta^{(l+1)}}{\theta^{(l)} + 1} \right) \right| < \tau_2.$$

2.3 Implementation details

2.3.1 Regularization step

Given $\hat{\theta}$ and the current $\hat{\alpha}$, maximizing (4.10) corresponds to solving the Lasso-penalized problem

$$\max_{\alpha} \{LL(\hat{\theta}, \alpha) - \lambda \|\alpha\|_1\}.$$

A second-order Taylor approximation of $LL(\alpha, \hat{\theta})$ around $\hat{\alpha}$ yields

$$LL(\alpha, \hat{\theta}) \approx L(\hat{\alpha}, \hat{\theta}) + (\alpha - \hat{\alpha})^T \frac{\partial}{\partial \alpha} LL(\alpha, \hat{\theta}) \Big|_{\alpha=\hat{\alpha}} + \frac{1}{2} (\alpha - \hat{\alpha})^T \frac{\partial^2}{\partial \alpha^2} LL(\alpha, \hat{\theta}) \Big|_{\alpha=\hat{\alpha}} (\alpha - \hat{\alpha}).$$

Let's define for $k = 1, \dots, K$, $A_k = \partial_{\alpha_k} LL(\alpha, \hat{\theta}) \Big|_{\hat{\alpha}} + \hat{x}_{kk} \hat{\alpha}_k$, where $\hat{x}_{kk}$ is the k th diagonal element of the opposite of the Hessian at $(\hat{\alpha}, \hat{\theta})$, assumed negative definite. The updates follow the standard soft-thresholding rule :

$$\alpha_k = \begin{cases} \frac{A + \lambda}{\hat{x}_{kk}} & \text{if } A < -\lambda, \\ \frac{A - \lambda}{\hat{x}_{kk}} & \text{if } A > \lambda, \\ 0 & \text{otherwise.} \end{cases} \quad \text{with } A = \partial_{\alpha_k} LL(\hat{\theta}, \hat{\alpha}) + \hat{x}_{kk} \hat{\alpha}_k \quad (4.11)$$

Additional derivations and proofs are provided in Appendix B.2.1 and log-likelihood and derivative

evaluations are detailed in the next Section 2.3.2.

To ensure numerical stability in case of model misspecification, the Hessian may fail to be negative definite; in that case we inflate its diagonal by adding a scaled identity matrix, in the spirit of the Marquardt-Levenberg method (Philipps *et al.*, 2021). The pseudocode of the regularization step appears in Algorithm 2.

Algorithm 2 : Regularization step at iteration l , with penalty parameter $\lambda > 0$:

Let $\theta^{(l)}$ be fixed, and $\alpha^{(l)}$ be the current estimate ;
Compute $LL(\alpha^{(l)}, \theta^{(l)})$, $\partial_{\alpha} LL(\alpha, \theta^{(l)})|_{\alpha=\alpha^{(l)}}$, and $\partial_{\alpha}^2 LL(\alpha, \theta^{(l)})|_{\alpha=\alpha^{(l)}} = -(x_{lc})_{l,c \leq K}$ and inflate the Hessian if needed.
For $k \leq K$:
 Compute $A_k = \partial_{\alpha_k} LL(\alpha^{(l)}, \theta^{(l)}) + x_{kk} \alpha_k^{(l)}$.
 if $A_k < -\lambda$:
 if $A_k < -\lambda$: $\alpha_k \leftarrow (A_k + \lambda)/x_{kk}$.
 else if $A_k > \lambda$: $\alpha_k \leftarrow (A_k - \lambda)/x_{kk}$.
 else : $\alpha_k \leftarrow 0$.
Compute $LL(\alpha, \theta^{(l)})$ with the updated α .
If $LL_{\text{pen},\lambda}(\alpha^{(l)}, \theta^{(l)}) > LL_{\text{pen},\lambda}(\alpha, \theta^{(l)})$,
recalibrate α by searching for a better candidate in its neighborhood.

2.3.2 Log-likelihood estimation

2.3.2.1 Log-likelihood derivatives

The log-likelihood of the model is given in Section 2.1.2. We now derive the first and second derivatives with respect to the coordinate α used in REMixed (see Figure 4.6 and Section 2.3.1). For every $k \leq K$,

$$\frac{\partial f_Y(Y_i | \theta, \eta_i) f_Z(Z_i | \alpha, \theta, \eta_i) f(\eta_i | \theta)}{\partial \alpha_k} = \frac{\partial f_Z(Z_i | \alpha, \theta, \eta_i)}{\partial \alpha_k} \times f_Y(Y_i | \theta, \eta_i) f(\eta_i | \theta),$$

and

$$\frac{\partial f_Z(Z_i | \alpha, \theta, \eta_i)}{\partial \alpha_k} = \frac{1}{\sigma_k^2} \left\{ \sum_{j \leq n_{ki}} s_k(R_i(t_{kij})) (Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))) \right\} \times f_Z(Z_i | \alpha, \theta, \eta_i).$$

Similarly,

$$\frac{\partial^2 f_Y(Y_i | \theta, \eta_i) f_Z(Z_i | \alpha, \theta, \eta_i) f(\eta_i | \theta)}{\partial \alpha_k^2} = \frac{\partial^2 f_Z(Z_i | \alpha, \theta, \eta_i)}{\partial \alpha_k^2} \times f_Y(Y_i | \theta, \eta_i) f(\eta_i | \theta),$$

and

$$\frac{\partial^2 f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta})}{\partial \alpha_k^2} = \left(\frac{1}{\sigma_k^4} \left(\sum_{j \leq n_{ki}} s_k(R_i(t_{kij})) (Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))) \right)^2 - \frac{1}{\sigma_k^2} \left(\sum_{j \leq n_{ki}} s_k(R_i(t_{kij}))^2 \right) \right) f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i).$$

Finally, for every $k_1, k_2 \leq K$, $k_1 \neq k_2$:

$$\frac{\partial^2 f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta})}{\partial \alpha_{k_1} \alpha_{k_2}} = \frac{\partial^2 f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i)}{\partial \alpha_{k_1} \alpha_{k_2}} \times f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta}),$$

and

$$\frac{\partial^2 f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta})}{\partial \alpha_{k_1} \alpha_{k_2}} = \prod_{k=k_1, k_2} \left\{ \frac{1}{\sigma_k^2} \left[\sum_{j \leq n_{ki}} s_k(R_i(t_{kij})) (Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))) \right] \right\} \times f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i).$$

Therefore, for $k \leq K$, we can then derivate the log-likelihood with respect to α_k :

$$\begin{aligned} \frac{\partial LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_k} &= \sum_{i \leq N} \frac{\int_{\boldsymbol{\eta}_i} \partial_{\alpha_k} f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta}) d\boldsymbol{\eta}_i}{\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\theta})} \\ &= \sum_{i \leq N} \frac{1}{\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\theta})} \times \int_{\boldsymbol{\eta}_i} \frac{\partial f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i)}{\partial \alpha_k} \times f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta}) d\boldsymbol{\eta}_i, \end{aligned} \quad (4.12)$$

and

$$\begin{aligned} \frac{\partial^2 LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_k^2} &= - \left(\frac{\partial LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_k} \right)^2 \\ &\quad + \sum_{i \leq N} \frac{1}{\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\theta})} \int_{\boldsymbol{\eta}_i} \partial_{\alpha_k}^2 f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) \times f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta}) d\boldsymbol{\eta}_i, \end{aligned} \quad (4.13)$$

and, for $k_1 \neq k_2$,

$$\begin{aligned} \frac{\partial^2 LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_{k_1} \partial \alpha_{k_2}} &= - \frac{\partial LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_{k_1}} \times \frac{\partial LL(\boldsymbol{\alpha}, \boldsymbol{\theta})}{\partial \alpha_{k_2}} \\ &\quad + \sum_{i \leq N} \frac{1}{\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\theta})} \times \int_{\boldsymbol{\eta}_i} \partial_{\alpha_{k_1} \alpha_{k_2}} f_Z(\mathbf{Z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) \times f_Y(Y_i | \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \boldsymbol{\theta}) d\boldsymbol{\eta}_i. \end{aligned} \quad (4.14)$$

2.3.2.2 Log-likelihood approximation by Gauss-Hermite quadrature

The integrals in Equations (4.9), (4.12), (4.13), and (4.14) are evaluated at $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ using adaptive Gauss-Hermite quadrature, combined with numerical ODE solvers to compute R_i and $(S_{pi})_p$ for $i = 1, \dots, N$ jointly across all integrals. We first use Monolix to approximate the posterior distribution of the random effects, which concentrates the mass of the integrands.

The posterior $f(\boldsymbol{\eta}_i \mid \mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\theta})$ of the individual random effects $\boldsymbol{\eta}_i$ given the data, is approximated in Monolix by a Metropolis-Hastings algorithm (Hastings, 1970). For each subject i , a chain $\{\boldsymbol{\eta}_i^{(m)}\}_{m=1}^M$ targeting this posterior is generated using a Gaussian proposal; a candidate $\boldsymbol{\eta}_i^*$ is accepted with a probability depending on the proposal density, the observation density $f(\mathbf{Y}_i, \mathbf{Z}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\alpha}, \boldsymbol{\theta})$, and the prior $f(\boldsymbol{\eta}_i \mid \boldsymbol{\theta})$. After a burn-in phase to reach stationarity, the empirical mean and covariance of a Gaussian approximation to the posterior are computed from the last M accepted samples $(\boldsymbol{\eta}_i^{(m)})$.

We then compute multidimensional adaptive Gauss-Hermite quadrature to evaluate the integrals in (4.9), (4.12), (4.13), and (4.14) at $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})$.

Log-likelihood and derivative evaluations are carried out at each iteration by the function ‘gh.LL’ in the **REMixed** package. The pseudocode of this step is given in Algorithm 3.

Algorithm 3 : Log-likelihood and derivates computation :

Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}$ be fixed ;

Approximate the Gaussian posterior $f(\boldsymbol{\eta}_i \mid \mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\theta})$ via MCMC in Monolix.

For each individual $i = 1, \dots, N$ (can be execute in parallel) :

 Compute the adaptive Gauss-Hermite grid $(\boldsymbol{\eta}_{il}, w_l)_{l \leq m}$ from the Gaussian posterior approximation.

 Evaluate the integrands of the log-likelihood and derivatives at $\boldsymbol{\eta}_{il}$ for $l \leq m$, with m the number of quadrature points, and take the corresponding weighted sums.

 For instance, for the individual likelihood, are computed :

$$f_Z(\mathbf{Z}_i \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_{il}), f_Y(\mathbf{Y}_i \mid \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_{il}), f(\boldsymbol{\eta}_{il} \mid \hat{\boldsymbol{\theta}})$$

 and then $\sum_{l \leq m} w_l f_Z(\mathbf{Z}_i \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_{il}) f_Y(\mathbf{Y}_i \mid \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_{il}) f(\boldsymbol{\eta}_{il} \mid \hat{\boldsymbol{\theta}})$ is returned as $\hat{\mathcal{L}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$.

Aggregate individual contributions to obtain model log-likelihood and derivatives.

2.3.3 NLME parameters estimation

Population parameters $\boldsymbol{\theta}$ are estimated at each iteration via the SAEM algorithm (Delyon *et al.*, 1999; Kuhn *et al.*, 2005), as implemented in Monolix. During this step, $\boldsymbol{\alpha}$ is held fixed at the previously estimated value $\hat{\boldsymbol{\alpha}}$ (see Figure 4.6).

The SAEM is an iterative algorithm for maximum-likelihood estimation in models with latent variables, particularly well-suited to NLMEMs. Each iteration l alternates :

- **a Stochastic Simulation step (S-step) :** sample the individual random effects $\boldsymbol{\eta}^{(l)}$ from their conditional distribution given the data and current parameter estimates, i.e., $\boldsymbol{\eta}^{(l)} \sim f(\boldsymbol{\eta} \mid \mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}^{(l)})$, to approximate the expected complete-data log-likelihood $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)})$.
- **a Maximization step (M-step) :** update $\boldsymbol{\theta}$ by maximizing the approximation of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)})$ obtained in the S-step.

The SAEM algorithm has been demonstrated to be highly efficient and reliable for a broad scope of complex models, including those involving categorical data, count data, time-to-event data, mixture models, and models based on differential equations or censored data. Its convergence properties have been rigorously established, ensuring robustness in parameter estimation.

In our implementation, Monolix is interfaced with R via **lixoftConnectors**. In **REMixed**, the SAEM and Lasso updates alternate until convergence in the ‘**remix**’ function for a fixed penalty, or across a grid via ‘**cv.remix**’.

2.3.4 Selection of the penalty parameter

To select the optimal penalty λ , we run the algorithm over the grid

$$\Lambda = \left\{ \lambda_{\max} \times \beta_{\lambda}^{l/N_{\lambda}}, l = 1, \dots, N_{\lambda} \right\},$$

where N_{λ} is the grid size and β_{λ} is user-defined (typically 0.001 (Park et Hastie, 2007; Friedman et al., 2010b)), and

$$\lambda_{\max} = \max \left(\partial_{\alpha} LL(\alpha, \theta^{(0)})|_{\alpha=0} \right). \quad (4.15)$$

The final model is chosen by minimizing the corrected Bayesian Information Criterion (BICc, Dellattre et al. (2014)), accounting for fixed and random effects and for the number of selected biomarkers :

$$\text{BICc}(\hat{\alpha}, \hat{\theta}) = -2LL(\hat{\alpha}, \hat{\theta}) + \log(N) \dim(\theta_R) + \log(n_{\text{tot}}) (\dim(\theta_F) + \#\{\alpha_k \neq 0\}),$$

where θ is decomposed as (θ_R, θ_F) , the components related to random and fixed effects, respectively.

Let $\hat{\theta}_{\lambda}$ and $\hat{\alpha}_{\lambda}$ denote the REMixed estimates for a given $\lambda \in \Lambda$. We then select

$$\lambda^* = \arg \min_{\lambda \in \Lambda} \text{BICc}(\hat{\alpha}_{\lambda}, \hat{\theta}_{\lambda}), \quad (\hat{\alpha}, \hat{\theta}) = (\hat{\alpha}_{\lambda^*}, \hat{\theta}_{\lambda^*}).$$

To compute $\lambda_{\max}$ defined in (4.15), the gradient in Section 2.3.2 simplifies. For $k \leq K$ and $\theta = \hat{\theta}$,

$$f_Z(Z_i | \alpha = 0, \hat{\theta}, \eta_i) = \prod_{k \leq K} \prod_{j \leq n_{ki}} \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{Z_{kij} - \mu_k}{\sigma_k} \right)^2 \right),$$

does not depend on η_i . Hence, for $i = 1, \dots, N$,

$$\mathcal{L}_i(\hat{\theta}, 0) = f_Z(Z_i | \hat{\theta}, \alpha = 0) \int_{\eta_i} f_Y(Y_i | \hat{\theta}, \eta_i) f(\eta_i | \hat{\theta}) d\eta_i,$$

and therefore $\partial_{\alpha_k} LL(\hat{\boldsymbol{\theta}}, \boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\mathbf{0}}$ is equal to

$$\sum_{i \leq N} \frac{1}{\int_{\boldsymbol{\eta}_i} f_Y(\mathbf{Y}_i | \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \hat{\boldsymbol{\theta}}) d\boldsymbol{\eta}_i} \times \int_{\boldsymbol{\eta}_i} \frac{1}{\sigma_k^2} \left(\sum_{j \leq n_{ki}} s_k(R_i(t_{kij})) (Z_{kij} - \mu_k) \right) \times f_Y(\mathbf{Y}_i | \hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i | \hat{\boldsymbol{\theta}}) d\boldsymbol{\eta}_i.$$

2.3.5 Final tests

After convergence, we assess the statistical significance of each estimated coefficient $\hat{\alpha}_k$ using Wald tests based on the asymptotic normality of maximum-likelihood estimators, testing $H_0 : \alpha_k = 0$. This procedure allows us to identify the parameters that contribute meaningfully to the model, providing insight into relevant biomarkers.

These tests can be launch in **REMixed** using ‘`computeFinalTest`’ function.

2.3.6 Initialization strategy

To reduce computation time, we propose an initialization that sets $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\alpha}^{(0)}$ from randomly partitioning the K biomarkers into Γ groups $(\mathcal{G}_\gamma)_{\gamma \leq \Gamma}$ of approximately equal size, fitting each group-specific model independently, and selecting the best.

Specifically, for each group $\mathcal{G}_\gamma$, define the model $\mathcal{M}_\gamma$ as follows :

- For $k \notin \mathcal{G}_\gamma$, the marker is considered as Gaussian noise (i.e., not selected). We set $\alpha_k = 0$ and use the simplified observation model

$$Z_{kij} = \mu_k + \varepsilon_{kij}, \quad \varepsilon_{kij} \sim \mathcal{N}(0, \sigma_k^2).$$

The parameters (μ_k, σ_k^2) are then estimated via Maximum Likelihood Estimation (MLE) using

$$\hat{\mu}_k = \overline{Z_{kij}}^{i,j}, \quad \hat{\sigma}_k^2 = \frac{1}{\sum_{i=1}^N n_i} \sum_{i=1}^N \sum_{j=1}^{n_i} \left(Z_{kij} - \overline{Z_{kij}}^{i,j} \right)^2, \quad (4.16)$$

where $\overline{Z_{kij}}^{i,j}$ denotes the overall mean of Z_{kij} across subjects and times.

- For $k \in \mathcal{G}_\gamma$, we use the full observation equation and estimate α_k jointly with $\boldsymbol{\theta}$ via the SAEM algorithm.

Finally, the total log-likelihood LL_γ of the fitted model $\mathcal{M}_\gamma$, considering the K markers of $\mathcal{G}_\gamma$ and $\overline{\mathcal{G}}_\gamma$, is then computed via importance sampling using Monolix software. At this stage, gradient and Hessian evaluations are not required, making this estimation computationally possible through Monolix, contrary to later estimation of the log-likelihood (see details in Section 2.3.2). The group $\mathcal{G}$ that yields the highest log-likelihood is then selected, and the corresponding estimates are used to initialize the REMixed algorithm.

This initialization is available through ‘`initStrat`’ function in the **REMixed** package, and its pseudocode is provided in Algorithm 4. Its used is illustrated in Section 4.

Algorithm 4 : Initialization strategy :

Randomly split biomarkers into Γ groups $\mathcal{G}_\gamma$ of (approximately) equal size.

For each $\gamma \leq \Gamma$:

 Compute MLEs $(\hat{\mu}_k, \hat{\sigma}_k)_{k \notin \mathcal{G}_\gamma}$ as in (4.16).

 With $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\alpha}_k = 0)_{k \notin \mathcal{G}_\gamma}$ fixed, estimate θ and $(\alpha_k)_{k \in \mathcal{G}_\gamma}$ by SAEM.

 Compute the log-likelihood LL_γ .

Set $(\alpha^{(0)}, \theta^{(0)})$ to the estimates from $\mathcal{M}_G$ with $G = \arg \max_{\gamma \leq \Gamma} LL_\gamma$.

3 Project architecture and connection with Monolix

3.1 Monolix project

The **REMixed** package relies on the Monolix software suite for non-linear mixed-effects model estimation, interfaced through the **lixoftConnectors** R package [R Core Team \(2024\)](#). This package provides a direct link between R and the MonolixSuite modeling tools, enabling users to automate workflows, control MonolixSuite software, and perform model development and simulation processes.

To initialize the Monolix connection in R, the user needs to load the library and call the initialization function :

```
library(lixoftConnectors)
initializeLixoftConnectors(software = "monolix")
```

Once initialized, the **lixoftConnectors** functions allow creating, editing, and running Monolix projects from R. A Monolix project is defined by a set of key files :

- ‘data.txt’ - the dataset used for estimation,
- ‘model.txt’ - the structural model, written in Mlxtran syntax,
- ‘project.mlxtran’ - the main project file specifying model settings,
- ‘project.mlxproperties’ - optional preferences for visualization.

See [Lixoft \(2024b\)](#) for details about the Monolix software.

When a task is executed (e.g., running the SAEM algorithm or computing conditional distributions), Monolix creates a results folder within the project directory, including various output files such as individual predictions and estimation summaries.

After executing a task, a subfolder is created with output files including ‘summary.txt’, which contains estimation and diagnostic results.

3.2 Summary files and temporary projects

During execution, the **REMixed** package creates temporary Monolix projects and stores summary files for traceability. These are saved in a dedicated ‘remix’ folder within the Monolix project directory.

- **Initialization step** (via ‘initStrat’) : temporary Monolix projects are stored in the ‘tmp_init’ folder. These are deleted by default to save space but can be preserved by setting `unlinkTemporaryProject = FALSE`. A log file called ‘summary_init.txt’ is saved in the ‘remix’ folder, summarizing the full initialization run (see Appendix [B.2.4](#)).
- **REMixed algorithm for a grid of penalty parameters** (via ‘cv.remix’) : one summary file per penalty value is saved as ‘summary_l.txt’ (where l is the index of the penalty λ_l), and a global summary file ‘summary_cv.txt’ is also created (see Appendix [B.2.5](#)). Temporary Monolix projects corresponding to each penalty value are named ‘Build_l.mlxtran’, and are deleted by default unless `unlinkBuildProject = FALSE`. The summary file produced for each penalty parameter has the same format as the one provided for a single penalty parameter through the ‘remix’ function.
- **REMixed algorithm for a given penalty parameter** (via ‘remix’) : a summary file named ‘summary.txt’ is created to monitor the iterations (see Appendix [B.2.6](#)). A copy of the initial project is also saved in the ‘remix’ folder to allow consistent reruns. The final Monolix project is saved in the same directory as the original input project.

After computing final tests and running the SAEM algorithm, the final ‘summary.txt’ is saved in the ‘remix’ folder of the specified final project. For instance, in the ‘project_upd’ folder (see Appendix [B.2.2](#)).

4 Usage

4.1 Demo project overview

We illustrate the use of the **REMixed** package through a demo project in a low-dimensional setting for fast computation. To access the Monolix project provided with the package :

```
library(ggplot2)
theme_set(theme_bw())

library(REMixed)
library(lixoftConnectors)
initializeLixoftConnectors()

project <- getMLXdir()
loadProject(project)
```

This example is based on a simulated antibody production model describing the dynamics over the 7 days following vaccine injection, during which antibody-secreting cells (ASC, denoted S) reach their peak and subsequently decline (Pasin *et al.*, 2019). These cells produce antibodies (AB) at a rate φ_S , while S and AB decay at rates δ_S and δ_{AB} , respectively. The simulation was generated using Simulx (Lixoft, 2024c), and the corresponding Monolix project is provided in the directory returned by `getMLXdir()`. For the structural model, written in Mlxtran syntax, is provided in Appendix B.2.3 (See Lixoft (2024b) for details about the Monolix software and Mlxtran syntax).

The mechanistic model is defined for subject $i = 1, \dots, N$ as :

$$\left\{ \begin{array}{l} \frac{dS_i}{dt}(t) = -\delta_{S_i} S_i(t), \\ \frac{dAB_i}{dt}(t) = \varphi_{S_i} S_i(t) - \delta_{AB_i} AB_i(t), \\ (S_i(0), AB_i(0)) = (0, 1000). \end{array} \right. \quad (4.17)$$

The individual parameters are assumed to follow log-normal distributions :

$$\left\{ \begin{array}{l} \log(\varphi_{S_i}) = \log(\varphi_{S_{pop}}) + \eta_i^{\varphi_S}, \\ \log(\delta_{AB_i}) = \log(\delta_{AB_{pop}}) + \eta_i^{\delta_{AB}}, \\ \log(\delta_{S_i}) = \log(\delta_{S_{pop}}) + \eta_i^{\delta_S}. \end{array} \right. \quad (4.18)$$

To simulate observations, antibody levels are measured at six time points t_{ij} , $j = 1, \dots, 6$,

post-vaccination (days 0, 7, 21, 123, 180, 300), following :

$$Y_{ij} = \log_{10}(AB_i(t_{ij})) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{Ab}^2). \quad (4.19)$$

In parallel, five biomarkers $(G_k)_{k \leq 5}$ are observed daily from day 0 to 21, at times t_{kij} , $j = 1, \dots, 22$, with only one biomarker truly associated with the ASC compartment S :

$$G_{kij} = \alpha_k S_i(t_{kij}) + \epsilon_{kij}, \quad \epsilon_{kij} \sim \mathcal{N}(0, \sigma_{G_k}^2). \quad (4.20)$$

Parameter values used in the demo are summarized in Table 4.5. The simulated data are shown in Figure 4.7.

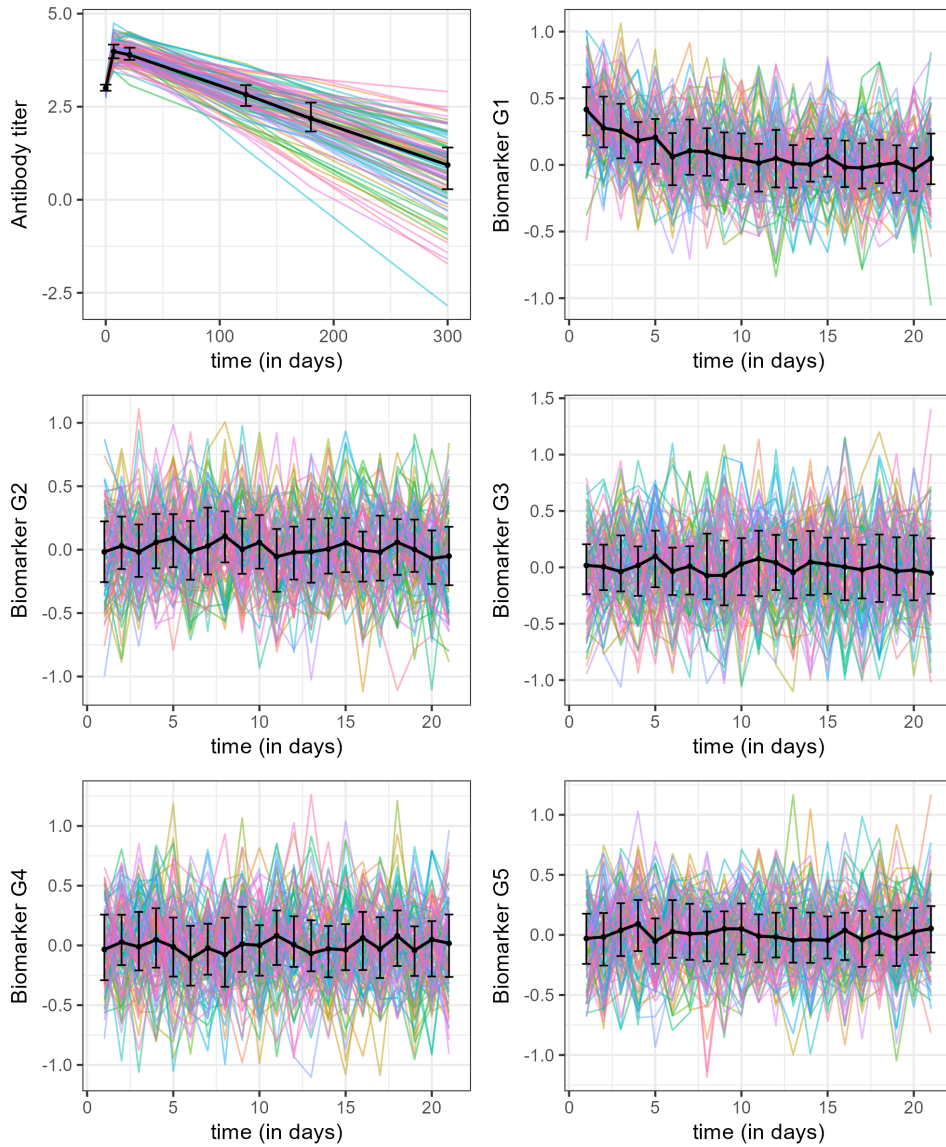


FIGURE 4.7 – Simulated observations for 100 individuals from the demo project included in the **REMixed** package, representing antibody dynamics and biomarkers dynamics over 21 days following vaccination. In this scenario, only one of the five biomarkers, G_1 , is truly associated with the ASC compartment S .

Parameter	Definition	Value
Fixed Effects		
<i>Influx rates</i>		
φ_S	ASC-S antibody production (Eu/mL/day)	6.11×10^2
<i>Decay rates</i>		
δ_{Ab}	Antibody decay rate (1/day)	2.50×10^{-2}
δ_S	ASC-S decay rate	2.30×10^{-1}
Random Effects		
ω_{φ_S}	SD for φ_S	0.60
ω_{δ_S}	SD for δ_S	0.10
$\omega_{\delta_{Ab}}$	SD for δ_{Ab}	0.30
Biomarkers related parameters		
α_{11}		0.10
$\alpha_{1k}, k > 1$		0
Observation Error		
σ_{Ab}	SD of error on log antibody levels	0.11
σ_{G1}	SD of error on biomarkers levels	0.28
$\sigma_{Gk}, k > 1$		0.34, 0.39, 0.36, 0.32

TABLE 4.5 – Parameter values used in the demo project simulation scenario, describing antibody and antibody-secreting cell (ASC) dynamics in 100 individuals. The scenario assumes five biomarkers $(G_k)_{k \leq 5}$, of which only G_1 is associated with the ASC compartment.

For **REMixed** to interface with Monolix, users must specify how to generate system dynamics and how to map outputs to observed data. This involves three elements :

1. The structural ODE model : it must return a table with time and named outputs. In the demo, ‘dynFUN_demo’ is pre-defined using the **deSolve** package (Karline Soetaert *et al.*, 2010) :

```

dynFUN_demo <- function(t, y, parms) {
  parms <- c(parms, S0 = y[["S"]])
  out <- deSolve::ode(y[["AB"]], t, function(t, y, parms) {
    with(as.list(c(y, parms)), {
      dAB <- phi_S * S0 * exp(-delta_S * t) - delta_AB * AB
      list(c(dAB))
    })
  }, parms)
  out <- cbind(out, S = y[["S"]] * exp(-parms[["delta_S"]] * out[, "time"]))
  class(out) <- "deSolve"
  return(out)
}

out <- dynFUN_demo(t = 1:300, y = c(S = 5, AB = 1000),
  parms = c(delta_S = 0.23, delta_AB = 0.025, phi_S = 611))
plot(out)

```

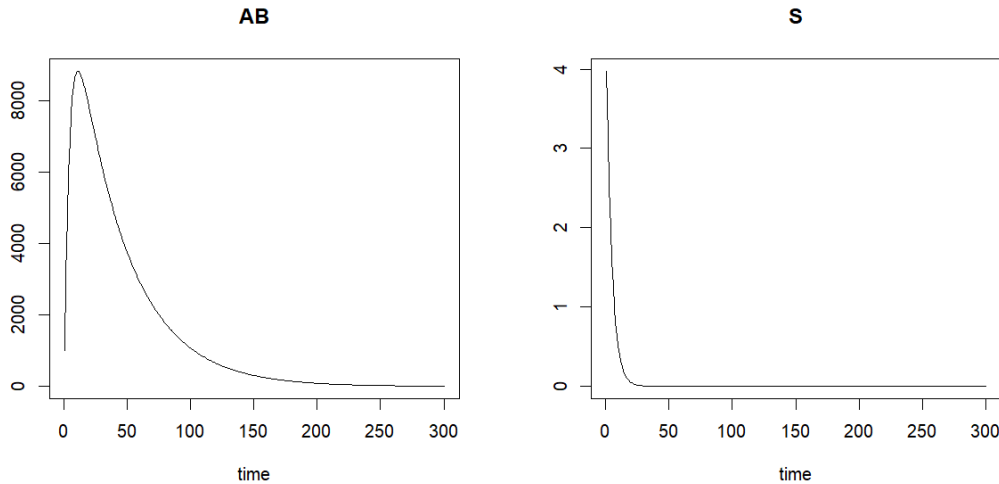


FIGURE 4.8 – "Deterministic dynamics of antibody and antibody-secreting cell (ASC) populations for the demo project, obtained by solving the system of differential equations with typical parameter values (random effects set to zero). These trajectories represent the expected population-level dynamics, without inter-individual variability.

2. The observation model : since Monolix does not store transformations used in the observation model, these must be explicitly specified via the 'ObsModel.transfo' object.

```
getContinuousObservationModel()$prediction
#   yAB   yG1   yG2   yG3   yG4   yG5
# "tAB"  "G1"  "G2"  "G3"  "G4"  "G5"

ObsModel.transfo = list(S = list(AB = log10),
                        linkS = "yAB",
                        R = rep(list(S = function(x){x}), 5),
                        linkR = paste0("yG", 1:5))
```

3. The biomarker Z_k statistical model : parameters must be explicitly linked to Monolix observation names. If no intercepts μ_k are included in the latent regression model, set `alpha0 = NULL`. Here 'alpha0' corresponds to the $\boldsymbol{\mu} = (\mu_k)_{k \leq K}$ parameters and 'alpha1' corresponds to $\boldsymbol{\alpha} = (\alpha_k)_{k \leq K}$ parameters.

```
alpha = list(alpha0 = NULL,
             alpha1 = setNames(paste0("alpha_1", 1:5), paste0("yG", 1:5)))
```

These settings are critical for translating the model between the **REMixed** package and the Monolix software.

4.2 Initialization

As described in Section 2.3.6, we implement an initialization strategy that performs multiple runs of the SAEM algorithm on randomly formed small groups of biomarkers, selecting the configuration yielding the highest log-likelihood. This ensures that early parameter estimates are sufficiently close to the true values to allow stable log-likelihood computation.

```
init <- initStrat(project, alpha, ObsModel.transfo, seed = 1710)
plotInit(init)
```

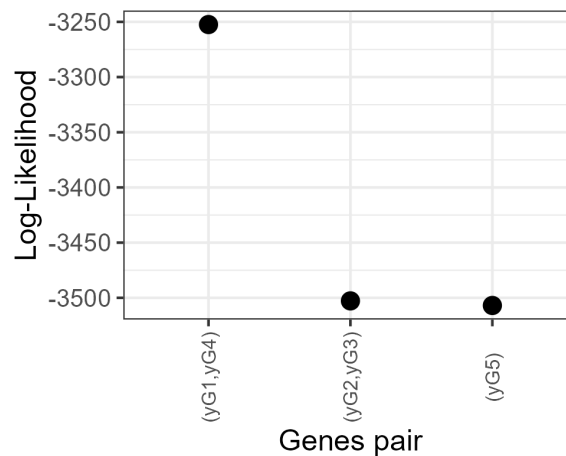


FIGURE 4.9 – Log-likelihood of each group of genes tested for initialization.

In this demo, the biomarkers were randomly grouped into (G1,G4), (G2,G3), and (G5). As shown in Figure 4.9, the group including G1-known to be truly associated with the latent compartment-produced the highest log-likelihood across initialization runs. This provides empirical support for the effectiveness of the initialization procedure.

A comprehensive summary of the initialization results, including the estimated parameters for each group and scores, is available in Appendix B.2.4. These logs are stored in the ‘summary_init.txt’ file described in Section 3 and are automatically generated in the ‘remix’ subfolder of the Monolix project directory.

4.3 Grid Search with REMixed

After initialization, we apply the REMixed algorithm to a grid of penalty values to select an optimal regularization parameter. The following command performs the grid search using 10 automatically chosen λ values. We retained 10 for illustration purposes to ensure reasonable computation time; in practice, a finer grid may be considered to refine the search. The corresponding outputs, including selection paths and BICc plots, together with the penalty parameter λ minimizing BICc, are illustrated in Figure 4.10 and obtained through the functions ‘plotCalibration’ and ‘plotIC’. The last function takes as an argument the information criterion to be used; the package provides the ‘BIC’, ‘AIC’, and ‘BICc’ functions by default, but user-defined criteria can also be provided.

```
resCV = cv.remix(project = project,
  dynFUN = dynFUN_demo,
  y = c(AB=1000,S=5),
  ObsModel.transfo = ObsModel.transfo,
  alpha = alpha,
  selfInit = TRUE,
  nlambda = 10,
  eps1=10**(-2),
  eps2=1)

plotCalibration(resCV, legend.position="right")
plotIC(resCV, criterion = BICc)
```

The tested grid $\Lambda = \{5.09, 10.17, \dots, 2554.29\}$ was automatically determined based on the maximal penalization value estimated internally. The optimal penalty $\lambda^* = 80.77$ was selected as the value minimizing the corrected Bayesian Information Criterion (BICc), equal to 6616.76 in this case, as described in Section 2.3.4. This criterion balances model complexity and fit. In this example, two genes were selected, G1 and G2, with $\alpha_1 = 0.10$ and $\alpha_2 = 0.006$, respectively.

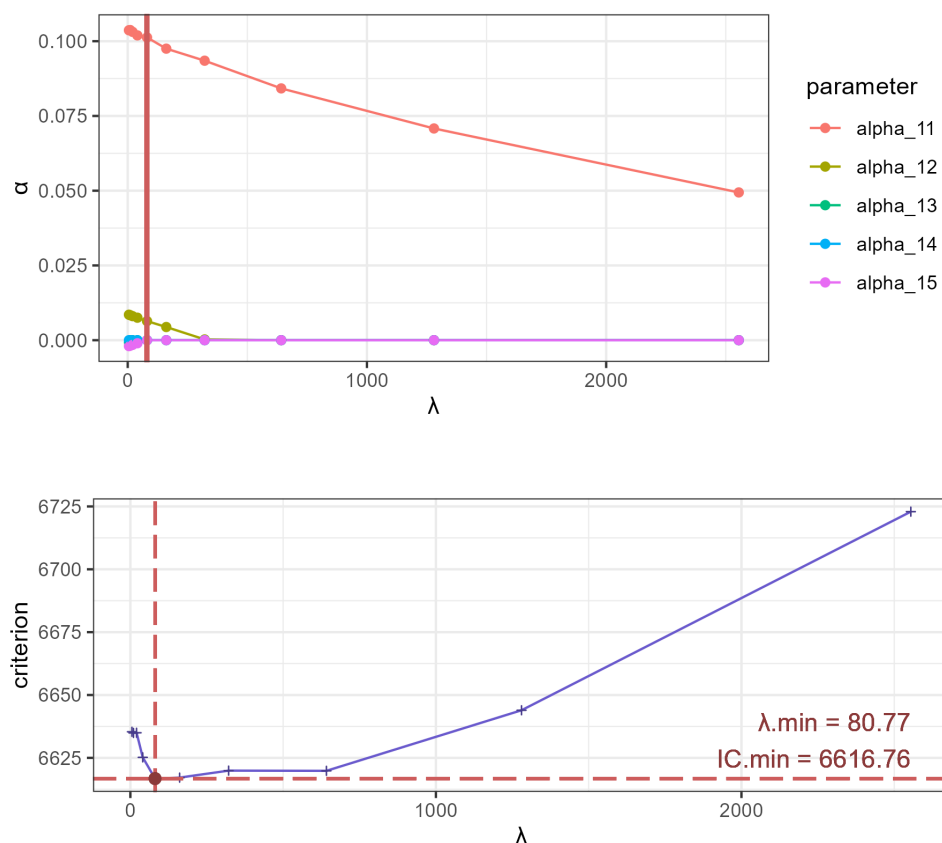


FIGURE 4.10 – Calibration curve (top panel) and corrected Bayesian Information Criterion (BICc) values across the grid of penalty parameters (bottom panel) in the REMixed procedure for the demo project with 10 penalty parameters tested.

```
selected_res <- retrieveBest(resCV)
summary(selected_res)
```

Outputs of remix algorithm from the project :

.../demoMLX/remix/demoMLX_init.mlxtran

with :

- 100 individuals;

- $\lambda = 80.77$.

• Final Estimated Parameters :

delta_S_pop	phi_S_pop	delta_AB_pop	omega_delta_S	omega_phi_S
0.25	596.32	0.02	0.16	0.61
omega_delta_AB	sigma_AB	sigma_G1	sigma_G2	sigma_G3
0.29	0.11	0.28	0.34	0.39
sigma_G4	sigma_G5			

```

          0.36          0.32
-----
• Final Selected Biomarkers and Estimate:
alpha_11_pop alpha_12_pop
    0.10100      0.00638
-----
• Final Scores :
  · OFV   : 6500.482
  · BIC   : 6564.954
  · BICc  : 6616.759

```

Executed in 2 iterations, 106 s.

All results are logged in the summary files ‘summary_cv.txt’, for the overall grid construction, and ‘summary_l.txt’, for each penalty parameter $l = 1, \dots, N_\lambda$ tested, described in Section 3, which are automatically saved in the ‘remix’ subfolder. Complete logs for this example are provided in Appendix B.2.5.

4.4 REMixed with fixed penalty parameter λ

If a specific penalty parameter λ is chosen, it is possible to run the REMixed algorithm for that fixed value to refine the final estimates. This is done using the ‘remix’ function, here with $\lambda = 80.77$:

```

res_lambda = remix(project = project,
                    dynFUN = dynFUN_demo,
                    y = c(AB=1000,S=5),
                    ObsModel.transfo = ObsModel.transfo,
                    alpha = alpha,
                    selfInit = TRUE,
                    lambda = 80.77,
                    eps1=0.01,
                    eps2=0.1)
summary(res_lambda)

```

Outputs of remix algorithm from the project :

```

...demoMLX_upd.mlxtran
with :
```

```

- 100 individuals;
- \textlambda = 80.77.
-----

• Final Estimated Parameters :
      delta_S_pop      phi_S_pop      delta_AB_pop      omega_delta_S      omega_phi_S
            0.24            570.69            0.02            0.11            0.62
omega_delta_AB      sigma_AB      sigma_G1      sigma_G2      sigma_G3
            0.30            0.11            0.28            0.34            0.39
      sigma_G4      sigma_G5
            0.36            0.32
-----

• Final Selected Biomarkers and Estimate:
alpha_11_pop alpha_12_pop
      0.09760      0.00668
-----

• Final Scores :
  · OFV   : 6499.817
  · BIC   : 6564.29
  · BICc  : 6616.094

Executed in 8 iterations, 352.2 s.
-----

```

```

trueValue <- read.csv(paste0(dirname(getMLXdir()),
                             "/demoSMLX/Simulation/populationParameters.txt"))

plotSAEM(res_lambda, trueValue = trueValue)
plotConvergence(res_lambda)

```

Figures 4.11 and 4.12 illustrate the convergence of the SAEM algorithm and the behavior of the log-likelihood across iterations. These are obtained with calls to ‘plotSAEM’ and ‘plotConvergence’. The argument ‘trueValue’, if known (as in simulation studies), allows displaying the true values of each parameter alongside their estimates. This step refines the parameter estimates based on the selected biomarkers and updates the Monolix project accordingly. As with previous steps, a full log is saved in the file ‘summary.txt’ and included in Appendix B.2.6.

4.5 Final Statistical Tests

To balance the shrinkage introduced by the Lasso penalty and validate the relevance of the selected biomarkers, we conduct a final run of the SAEM algorithm followed by Wald-based statistical testing, using the ‘computeFinalTest’ function.

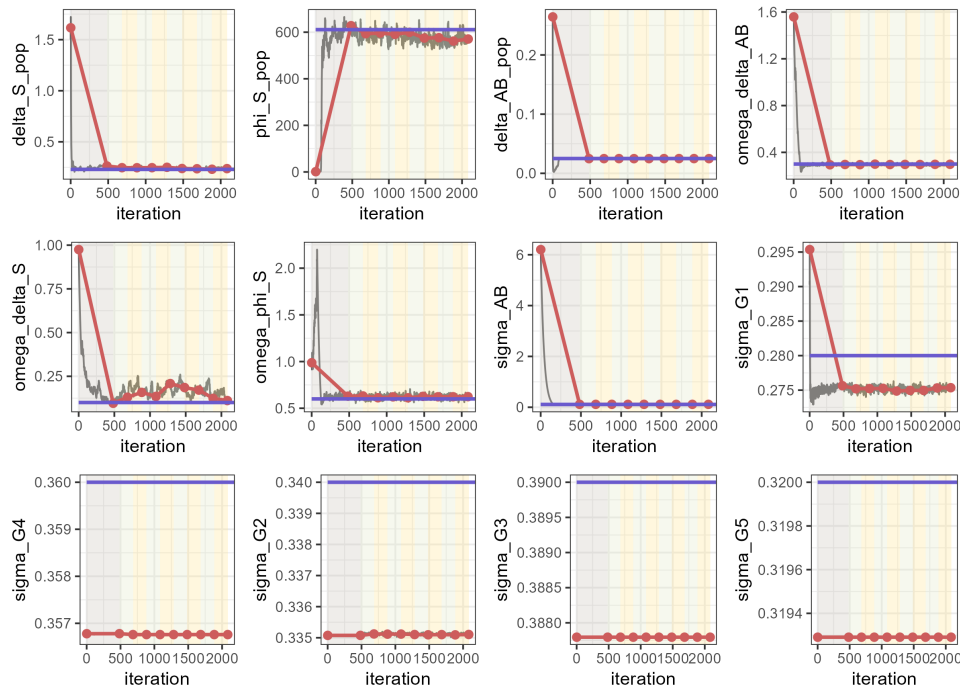


FIGURE 4.11 – Evolution of parameter estimates across iterations of the SAEM algorithm within the REMixed procedure for the demo project. Initialization iterations are shown in grey, and REMixed iterations are displayed in alternating green and yellow segments.

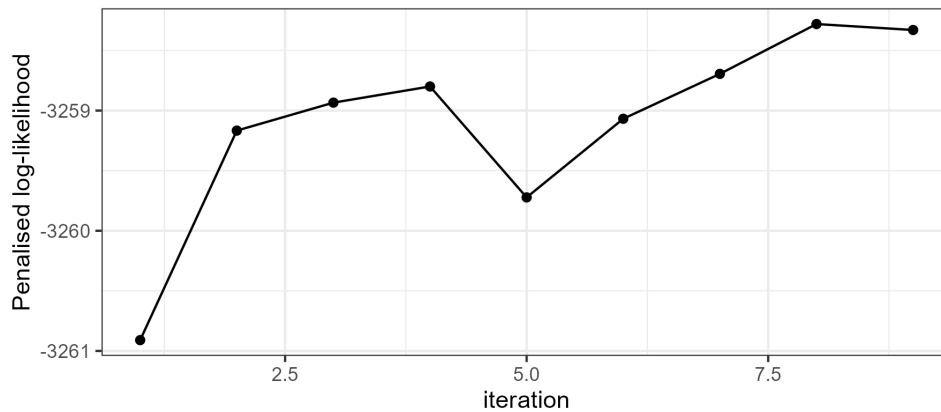


FIGURE 4.12 – Trajectory of penalized log-likelihood evolution over SAEM iterations in the REMixed procedure for the demo project.

This step recomputes the SAEM estimates without penalization on the selected subset of biomarkers and provides standard errors and 95% confidence intervals. Wald tests are then used to confirm the significance of each selected biomarker, ensuring that only those with robust support remain in the final model. If the set of selected biomarkers changes, a final estimation step is carried out using the SAEM algorithm.

In the demo project, biomarker G2 was removed by the final test, confirming the validity of the previous selection of G1 only. Parameter estimates, standard errors, confidence intervals and p-value are reported in Appendix B.2.7, along with full logs saved in the final ‘summary.txt’ file within the ‘remix’ folder of the corresponding final project. After this test, and since the selected

set had changed, a final run of the SAEM algorithm was performed, jointly estimating NLME parameters and the parameters related to biomarker G1.

```
res_final <- computeFinalTest(res_lambda, dynFUN_demo, c(AB=1000,S=5),
                             ObsModel.transfo)
```

This final validation step not only provides inferential guarantees but also prepares the model for interpretation and reporting, making it a critical component of the REMixed workflow. The final outputs can be investigated as an object from the ‘remix’ function.

```
res_final <- computeFinalTest(res_lambda, dynFUN_demo, c(AB=1000,S=5),
                             ObsModel.transfo)

summary(res_final)
```

```
-----
Outputs of remix algorithm from the project :
```

```
    .../demoMLX_upd_final.mlxtran
```

```
with :
```

- 100 individuals;
- \textlambda = 382.22;
- final SAEM computed;
- final test computed.

```
-----
• Final Estimated Parameters :
```

	EstimatedValue	se	CI_95
delta_S_pop	2.4e-01	9.5e-03	[2.2e-01;2.6e-01]
phi_S_pop	5.9e+02	4.1e+01	[5e+02;6.7e+02]
delta_AB_pop	2.5e-02	7.5e-04	[2.3e-02;2.6e-02]
omega_delta_S	1.7e-01	5e-02	[6.7e-02;2.6e-01]
omega_phi_S	6.1e-01	5e-02	[5.2e-01;7.1e-01]
omega_delta_AB	3e-01	2.2e-02	[2.5e-01;3.4e-01]
sigma_AB	1.1e-01	4.1e-03	[1.1e-01;1.2e-01]
sigma_G1	2.8e-01	4.3e-03	[2.7e-01;2.8e-01]
sigma_G2	3.4e-01	5.2e-03	[3.2e-01;3.5e-01]
sigma_G3	3.9e-01	6e-03	[3.8e-01;4e-01]
sigma_G4	3.6e-01	5.5e-03	[3.5e-01;3.7e-01]
sigma_G5	3.2e-01	4.9e-03	[3.1e-01;3.3e-01]

```
-----
• Final Selected Biomarkers and Estimate:
```


	EstimatedValue	se	CI_95
alpha_11_pop	1e-01	4.9e-03	[9.2e-02;1.1e-01]

-
- Final Scores :
 - OFV : 6502.757
 - BIC : 6562.624
 - BICc : 6609.719

Executed in 8 iterations, 429.1 s.

5 Conclusion

In this paper, we presented **REMixed**, a new framework implemented in R for integrating large-dimensional longitudinal biomarkers into non-linear mixed-effects models based on ODEs. This approach extends regularized regression techniques to a mechanistic modeling context, enabling simultaneous parameter estimation and biomarker selection within ODE-based mixed-effects models. Using a cyclical coordinate descent algorithm to maximize a penalized likelihood, **REMixed** dynamically links a large set of candidate biomarkers to latent biological compartments. We described the algorithm step by step and applied it to a simulated dataset, illustrating how **REMixed** can identify relevant biomarkers driving the latent processes in an ODE system while accurately estimating model parameters.

The performance of **REMixed** depends on initialization, as the convergence of SAEM can be sensitive to starting values. In practice, grouping biomarkers into small sets proved effective in balancing computational feasibility with estimation stability. Other strategies could further improve robustness, such as sequentially including biomarkers or exploiting directly observed compartments when available and sufficient for parameter estimation. Moreover, computational cost increases with both the number of biomarkers and the complexity of the mechanistic model. An initial preselection step based on prior biological knowledge or simpler models may therefore help reduce this burden.

The reliance on adaptive Gauss-Hermite quadrature for likelihood approximation can also become prohibitive when many random effects are included. Exploring alternative integration schemes, such as stochastic approximation or refined quadrature methods, could improve scalability. Similarly, inference approaches like particle filtering, variational inference, or sequential Monte Carlo may offer greater stability in challenging scenarios, but at the cost of additional implementation complexity. Finally, selection of the penalty parameter in our implementation relied on the corrected Bayesian Information Criterion (BICc), but the framework can easily accommodate alternatives such as AIC or BIC. Cross-validation could also be used to calibrate the penalty parameter and help mitigate overfitting; however, the repeated model estimations it requires were not computationally feasible in our context.

Interoperability is another important dimension. The current implementation relies on Monolix via **lixoftConnectors**, but broader compatibility with tools such as NONMEM or **saemix** would facilitate adoption across pharmacometrics, epidemiology, and immunology. Moreover, it would be interesting to integrate or compare the method with established pharmacometric software such as NONMEM, or fully R-based frameworks such as the **saemix** package.

The current implementation assumes independent, homoscedastic Gaussian errors and constant intercepts in the biomarker model. Allowing heteroscedasticity or temporal correlation in the residuals could improve realism for complex data. For instance, introducing time-varying intercepts $\mu_k(t)$ or more flexible variance functions $\sigma_k(\cdot)$ could better capture time-dependent biomarker influences on the latent process.

At present, biomarkers are linked to a single latent compartment, but the **REMixed** algorithm could be extended to multiple latent compartments in the ODE system, allowing identification of distinct biomarker signatures associated with different biological processes, such as diverse immune cell populations. The methodology also opens perspectives for broader applications. Extensions to multi-omics data, including transcriptomics, proteomics, or metabolomics, would support integrative analyses of complex biological responses. By identifying dynamic biomarker signatures associated with latent processes, **REMixed** may contribute to advances in precision medicine.

4.5 Perspectives

L'objectif de ce travail de thèse était de proposer une approche méthodologique permettant d'intégrer les données transcriptomiques longitudinales dans des modèles mécanistes, en utilisant ces données comme proxies de compartiments latents. À cette fin, la méthode **REMixed** propose un cadre permettant de modéliser simultanément l'évolution temporelle d'un compartiment latent $R(t)$ au sein d'un modèle mécaniste défini par des EDOs, et l'observation indirecte de ce compartiment par des profils d'expression génique, ou d'autres biomarqueurs, longitudinaux. La méthode permet ainsi de sélectionner et d'estimer les associations gène-compartiment, ainsi que l'ensemble des paramètres du modèle, à l'aide d'un algorithme de descente cyclique couplant régression Lasso et estimation par maximum de vraisemblance. Cette approche a montré de bons résultats en simulation dans un contexte de dimension large, avec un nombre de biomarqueurs plus important que le nombre d'individus. Elle repose néanmoins sur certaines hypothèses fortes, comme observé dans l'application proposée, biologiquement discutables, qui ouvrent de nombreuses perspectives d'amélioration pour des travaux futurs.

En particulier, une première limite concerne l'hypothèse d'indépendance des erreurs $\varepsilon_{kij} \sim \mathcal{N}(0, \sigma_k^2)$. Étant donnée la forte corrélation entre gènes observée dans les données transcriptomiques, supposer un modèle d'erreur indépendant et homoscedastique ne semble pas refléter fidèlement la réalité biologique. Une extension naturelle consisterait à supposer des erreurs

corrélées :

$$Z_{kij} = \mu_k + \alpha_k s_k(R_i(t_{kij})) + \varepsilon_{kij},$$

avec $\varepsilon_{ij} = (\varepsilon_{kij})_{k \leq K} \sim \mathcal{N}(0, \Sigma)$, avec ε_{kij} indépendant de $\varepsilon_{kij'}$ pour $j \neq j'$. Ceci permettrait ainsi de modéliser explicitement les dépendances entre les différents biomarqueurs observés. Une autre alternative serait d'introduire une structure de corrélation temporelle dans les erreurs, en supposant par exemple $\varepsilon_{ki} = (\varepsilon_{kij})_{j \leq n_{ki}} \sim \mathcal{N}(0, \Sigma_k)$.

De la même manière, l'hypothèse d'un lien linéaire direct entre chaque biomarqueur et un unique compartiment latent constitue une approximation forte. Il serait souhaitable de modéliser l'intercept moyen μ_k comme une fonction du temps, $\mu_k(t)$, ou éventuellement avec une composante individuelle, afin de capturer des effets biologiques indirects, non stationnaires, et potentiellement indépendants de la dynamique immunitaire.

Dans la version actuelle de REMixed, un seul compartiment R est relié aux biomarqueurs longitudinaux Z_k , $k = 1, \dots, K$. Une extension naturelle de la méthode serait donc de considérer le cas où les biomarqueurs sont liés à plusieurs compartiments. Un tel modèle d'observation permettrait de modéliser des relations plus complexes entre le modèle mécaniste et les biomarqueurs observés. Par exemple, dans le cas de la réponse immunitaire humorale, il serait alors possible de modéliser différentes populations de plasmocytes, comme dans le modèle d'Andraud *et al.* présenté en Section 2.3.4, avec chacune un effet éventuellement différent sur les biomarqueurs. Cette extension est en cours de développement au sein de l'équipe SISTM (Inria, Inserm, BPH), dans le cadre de la thèse de Lisa Crépin (encadrée par Mélanie Prague (Inria, Inserm, BPH, Université de Bordeaux) et Morgan Craig (Centre de recherche du CHU Sainte-Justine, Université de Montréal)). Elle vise donc, formellement, à généraliser la méthode REMixed au cas multi-compartiments, où l'idée est de considérer plusieurs compartiments latents $R_1, \dots, R_L$ contribuant simultanément à l'expression des biomarqueurs. Le modèle d'observation devient alors :

$$Z_{kij} = \mu_k + \alpha_{1k} s_{1k}(R_{1i}(t_{1kij})) + \dots + \alpha_{Lk} s_{Lk}(R_{Li}(t_{Lkij})) + \varepsilon_{kij},$$

où $s_{lk}(\cdot)$ sont des transformations fixées appliquées aux L compartiments latents R_l , et $\varepsilon_{kij} \sim \mathcal{N}(0, \sigma_k^2)$. Cette structure permettrait de mieux représenter la complexité des interactions entre les compartiments biologiques et les profils d'expression génique, en particulier dans le cas de réponses multifactorielles.

Plusieurs extensions méthodologiques naturelles peuvent également être envisagées. Une première piste consisterait à permettre l'utilisation de pénalités alternatives dans l'étape de sélection des coefficients $\alpha = (\alpha_1, \dots, \alpha_p)$, telles que les pénalités SCAD (Fan et Li, 2001) et MCP (Zhang, 2010), présentées formellement en Section 3.4. Ces pénalités permettent de réduire le biais de sélection introduit par le Lasso sur les coefficients importants, autrement dit limiter l'effet de *shrinkage*. On pourrait également envisager l'utilisation d'une pénalité *elastic net* (Zou et Hastie, 2005), qui combine les propriétés des normes L_1 et L_2 , et s'écrit pour un vecteur α :

$$\text{pen}_{\text{EN}}(\alpha) = \lambda_1 \|\alpha\|_1 + \lambda_2 \|\alpha\|_2^2.$$

Cette pénalité est particulièrement adaptée aux cas de colinéarité entre les variables explicatives, comme c'est fréquemment le cas dans les données transcriptomiques. En ajustant les poids λ_1 et λ_2 , elle permet une sélection plus stable et plus équilibrée entre covariables redondantes. De plus, plusieurs extensions pourraient être envisagées pour remplacer la calibration par le BICc du paramètre de pénalisation λ . Cela pourrait permettre de surmonter la principale limitation de la méthode proposée, à savoir le surapprentissage observé dans les résultats, qui motive le recours aux tests de Wald. Une meilleure calibration permettrait de réduire le surapprentissage et de conduire à une sélection plus parcimonieuse des biomarqueurs $\mathbf{Z}_k$, limitant ainsi le recours à des tests statistiques supplémentaires. En effet, les tests statistiques finaux, permettant un éventuel contrôle, sont réalisés sur un ensemble sélectionné par l'algorithme au préalable, ce qui soulève des problématiques d'inférence post-sélection. Pour cela, des méthodes alternatives pourraient être envisagées, comme l'utilisation de la validation croisée telle que proposée par [Feng et Yu \(2019\)](#), ce qui pourrait permettre de s'affranchir de tests complémentaires en calibrant plus adéquatement le paramètre de pénalisation. Néanmoins, les contraintes computationnelles de REMixed constituent un frein majeur à la mise en œuvre pratique d'une telle procédure, qui pourrait ainsi être explorée dans de futurs travaux.

Enfin, sur le plan appliqué, plusieurs perspectives peuvent être mentionnées. Une première consiste à affiner les stratégies de pré-sélection des gènes utilisés comme proxies, en explorant d'autres bases de données, d'autres méthodes de filtrage, ou des approches plus intégratives combinant connaissances biologiques et sélection automatique. Une autre piste concerne l'amélioration des technologies de mesure transcriptomique elles-mêmes : la forte variabilité technique actuelle constitue un frein à la robustesse de la sélection. À mesure que des données longitudinales de meilleure qualité deviendront disponibles, les performances de REMixed pourront être améliorées. Si nos illustrations concernent la réponse humorale après vaccination, l'approche proposée est générique et valable dans des contextes variés utilisant les modèles à effets mixtes. Ainsi, nous projetons d'appliquer la méthode à des données de protéomique, pour lequel nous pensons que le lien entre biomarqueurs et compartiments latents serait plus direct.

– Chapitre 5 –

Discussion générale

Cette thèse s’est attachée à développer de nouvelles méthodes statistiques dans le cadre des modèles NLEM basés sur des EDOs pour l’analyse de données complexes issues de la vaccinologie. L’ambition était de rendre possible l’intégration de données de plus grande dimension dans les modèles mécanistes, en raison du développement récent et important de ce type de données, comme les données transcriptomiques. Les contributions principales se déclinent selon deux axes complémentaires : la proposition d’un algorithme de sélection de variables explicatives adapté à la grande dimension, Lasso-SAMBA, et le développement d’un cadre méthodologique original pour intégrer des données transcriptomiques longitudinales en tant que proxies de compartiments latents, REMixed.

La première contribution de cette thèse concerne ainsi le développement d’une méthode de sélection de covariables dans les modèles mécanistes, adaptée aux données de grande dimension. À notre connaissance, il n’existait pas de procédure permettant d’effectuer cette tâche lorsque le modèle est défini par un système d’EDO, en particulier en contexte $p \gg n$. L’algorithme proposé, appelé Lasso-SAMBA, étend la méthode step-SAMBA (Prague et Lavielle, 2022), performante dans un cadre de dimension modérée ou faible, mais inadaptée dès lors que le nombre de covariables explose. En parallèle, la méthode SAEMVS (Naveau *et al.*, 2023), permettait déjà une sélection de covariables dans les modèles NLEM analytiques, mais n’était pas applicable aux modèles basés sur des EDOs. Lasso-SAMBA comble ainsi un vide méthodologique en introduisant une approche de sélection pénalisée et stable dans ce cadre complexe des modèles mécanistes. L’approche repose sur une combinaison entre pénalisation de type Lasso (Tibshirani, 1996) et sélection stable (Meinshausen et Bühlmann, 2010; Bodinier *et al.*, 2023), garantissant une sélection robuste aux fluctuations d’échantillonnage. Les résultats de simulations montrent des performances élevées en termes de contrôle du FDR et de F1-score, surpassant les approches alternatives. De plus, l’algorithme est rapide, peu sensible aux hyperparamètres et propose une calibration automatique de la regression Lasso via un seuil sur le taux espéré de faux positifs interprétable (Meinshausen et Bühlmann, 2010; Bodinier *et al.*, 2023). Une implémentation logicielle a été développée dans un paquet R lié à Monolix, rendant la méthode facilement utilisable. Ainsi, Lasso-SAMBA constitue une avancée importante pour construire des modèles mécanistes parcimonieux expliquant la variabilité inter-individuelle à partir de données de grande dimension. D’un point de vue appliqué, nous nous sommes intéressés à la réponse immunitaire humorale à la suite d’injection de vaccin contre le virus Ebola (Badio *et al.*, 2021) et Varicella-Zona (Li *et al.*, 2017). Pour cela, nous avons analysé les dynamiques anticorps des patients ayant reçu les différents vaccins et nous avons sélectionné parmi des gènes pré-sélectionnés, mesurés

chez ces patients, lesquels étaient des prédicteurs de la réponse immunitaire humorale, mettant en lumière des associations permettant une meilleure compréhension de celle-ci.

La seconde contribution, appelée REMixed, propose une approche originale d'intégration de données transcriptomiques longitudinales dans les modèles mécanistes. L'idée centrale est de considérer des biomarqueurs comme des proxies de compartiments latents du modèle, permettant ainsi d'enrichir les observations disponibles pour prédire les compartiments d'un modèle mécaniste et aussi d'améliorer l'identifiabilité des paramètres. À notre connaissance, aucune méthode n'avait encore été proposée pour relier directement profils transcriptomiques longitudinaux et compartiments latents d'un modèle mécaniste compartimental. L'approche repose sur une régularisation, de type Lasso (Tibshirani, 1996), pour sélectionner les biomarqueurs apportant de l'information sur le compartiment latent et permettre l'estimation conjointe des paramètres du modèle et des contributions des biomarqueurs dans un algorithme de descente cyclique (Tibshirani, 1996; Tseng, 2001; Friedman *et al.*, 2010b). Elle fournit ainsi une première brique méthodologique pour explorer le lien entre dynamiques observées et processus cellulaires sous-jacents dans le cas de la vaccinologie. Toutefois, la méthode présente des limites. Elle reste coûteuse sur le plan computationnel et difficile à appliquer en très grande dimension : pour l'instant, son utilisation s'est limitée à de petits échantillons (avec $p \approx 50$ et $N \approx 20$), en particulier du fait du recours au SAEM qui rend son extension directe aux données transcriptomiques de grande taille complexe. Par ailleurs, elle repose sur des hypothèses biologiques fortes concernant le lien entre biomarqueurs et compartiments latents, dont notamment l'idée que les biomarqueurs non sélectionnés ne sont qu'un bruit Gaussien. Ces hypothèses ouvrent néanmoins la voie à de futures extensions vers des modèles plus flexibles et réalistes. Un paquet R, accompagné d'une vignette détaillée, a été développé afin de favoriser la reproductibilité et l'appropriation par d'autres chercheurs. D'un point de vue appliqué, nous nous sommes intéressés à la réponse immunitaire humorale à la suite d'injection de vaccin contre le virus SARS-CoV-2 (Rinchai *et al.*, 2022). Nous avons analysé les dynamiques anticorps ainsi que les profils d'expressions géniques de ces patients à l'aide de cette méthode.

Les deux approches proposées dans cette thèse contribuent, chacune à leur manière, à améliorer les problèmes d'identifiabilité rencontrés dans les modèles mécanistes. La méthode Lasso-SAMBA permet de réduire la complexité des modèles construits en éliminant les covariables non informatives, là où des approches classiques comme le SCM, COSSAC, ou step-SAMBA, tendent à inclure un nombre excessif de covariables et à générer des modèles difficiles à identifier en pratique. À noter que dans des contextes de grande dimension, SCM et COSSAC deviennent inutilisables. Bien qu'elle n'améliore pas directement l'identifiabilité structurelle du modèle, elle conduit à des modèles plus parcimonieux et interprétables. À l'inverse, REMixed, en intégrant de nouvelles observations longitudinales considérées comme proxies des compartiments latents, enrichit l'information disponible et contribue à lever certaines ambiguïtés, renforçant ainsi l'identifiabilité pratique et théorique des paramètres du modèle. Dans le domaine de la vaccinologie, les deux contributions présentent un intérêt complémentaire : l'algorithme Lasso-SAMBA permet d'identifier des covariables associées à la variabilité inter-individuelle des réponses vaccinales, en reliant par exemple certains gènes à des paramètres mécanistes

spécifiques ; l’algorithme REMixed, quant à lui, offre la possibilité de reconstruire dynamiquement des compartiments immunitaires non observables, tels que les plasmocytes, à partir de données transcriptomiques longitudinales. Ces approches combinées ouvrent des perspectives pour une compréhension plus fine des mécanismes de la réponse vaccinale, en reliant variabilité observée aux dynamiques immunitaires sous-jacentes. Un point commun entre les deux méthodes concerne leur dépendance au logiciel Monolix, qui constitue un outil puissant mais dont le caractère propriétaire et non libre d’accès limite la transparence et la reproductibilité. Si Lasso-SAMBA reste relativement rapide, et directement utilisable dans un contexte pratique, REMixed souffre de coûts computationnels élevés et de limites fortes en grande dimension, ce qui en restreint l’applicabilité immédiate. L’ouverture vers des implémentations libre d’accès et des algorithmes plus efficaces apparaît ainsi comme un enjeu transversal pour la poursuite de ces travaux. Il est également important de souligner que les problèmes d’identifiabilité rencontrés dans ces modèles peuvent être de nature différente et donc soulèvent différentes problématiques. Lorsque l’identifiabilité est pratique, elle peut être améliorée par l’acquisition de données plus riches et de meilleures qualités : par exemple, une mesure plus précise de la charge virale, un dénombrement plus fiable des cellules immunitaires, ou encore une augmentation du nombre de points temporels et de biomarqueurs observés. En revanche, lorsque l’identifiabilité est théorique, elle ne peut pas être résolue par l’ajout de données et nécessite de simplifier le modèle, ou de fixer certains paramètres. Dans les deux cas, l’exploitation la plus complète possible de l’information disponible à travers plusieurs jeux de données reste une voie privilégiée pour limiter ces difficultés, même si elle alourdit inévitablement la modélisation et les calculs associés.

Les perspectives de ces travaux sont multiples, et nous avons détaillé des perspectives propres à chacune de ces contributions dans les Sections 3.4 et 4.5. D’un point de vue statistique, les méthodes proposées ont été validées empiriquement par de nombreux schémas de simulation mais pourraient bénéficier de garanties théoriques. Nous pourrions nous intéresser aux propriétés de convergence des méthodes proposées. Pour l’algorithme Lasso-SAMBA, la décroissance recherchée du critère d’information, comme le BICc, est observée au fur et à mesure des itérations de l’algorithme global. Les simulations montrent qu’à une itération donnée, sans la décroissance de ce critère d’information sur le modèle de covariables à l’étape de sélection de celui-ci, la décroissance globale du critère d’information était perdue, engendrant des problèmes de convergence. Il serait donc intéressant dans de futurs travaux d’explorer le lien entre le critère d’information du modèle de covariable à une itération, sur le modèle des paramètres individuels, et le critère d’information du modèle mécaniste complet. D’autre part, nous pourrions également nous intéresser aux garanties statistiques en termes de contrôle du nombre de faux positifs pour chacune des méthodes proposées. De la même manière, nous proposons dans la méthode REMixed un algorithme d’estimation et de sélection conjointe. Les tests statistiques finaux, permettant un éventuel contrôle, sont réalisés sur un ensemble sélectionné par l’algorithme au préalable, posant des questions claires d’inférence post-sélection. Les estimateurs obtenus conditionnellement au modèle retenu tendent à être biaisés, et les intervalles de confiance sont souvent trop optimistes. Des approches récentes d’inférence sélective ont été proposées pour les modèles linéaires mixtes (Claeskens *et al.*, 2021), et certaines méthodes comme la division d’échantillon (*sample splitting*) restent des alternatives robustes (Rasines et Young, 2022). Toutefois, aucune solution satisfaisante

n'existe encore dans le contexte des modèles mécanistes de grande dimension. D'un point de vue appliqué également, il est important de souligner que si les applications présentées dans ce manuscrit se sont concentrées sur la réponse immunitaire humorale, les méthodes développées ne s'y limitent pas. Elles sont en effet applicables à un vaste éventail de domaines où les modèles mécanistes jouent un rôle central, qu'il s'agisse de pharmacologie, d'écologie, ou d'autres aspects de la vaccinologie. Dans ce dernier cadre, elles pourraient par exemple être mobilisées pour explorer la réponse immunitaire innée ou à médiation cellulaire, offrant ainsi un potentiel d'extension bien au-delà des applications illustrées ici.

En conclusion, cette thèse a proposé deux contributions méthodologiques originales pour l'analyse de données complexes à l'aide de modèles mécanistes. Au-delà de leurs différences, ces deux approches partagent une ambition commune : promouvoir et améliorer l'utilisation des modèles mécanistes basés sur des EDOs dans un contexte de données de grande dimension. L'application en vaccinologie a permis d'illustrer leur intérêt, en identifiant à la fois des signatures génétiques associées à la variabilité inter-individuelle et des dynamiques cellulaires inobservables. Ces contributions s'inscrivent ainsi dans un mouvement méthodologique plus large visant à rapprocher modélisation mécaniste et données transcriptomiques, et constituent une première étape vers des modèles plus intégratifs et interprétables en santé publique.

Matériels Supplémentaires

Table des matières

A	Matériels Supplémentaires associés au Chapitre 3	172
A.1	Simulation generation process	172
A.1.1	Pharmacokinetic model	172
A.1.2	Humoral Immune Response	173
A.1.3	Covariates Generation	175
A.2	Additional analysis for the real-data application	178
A.2.1	Initial model	178
A.2.2	Final Model	184
A.3	Additional graphs for results	189
A.4	Figures additionnelles concernant l'analyse sur les données PREVAC-UP	191
B	Matériels Supplémentaires associés au Chapitre 4	193
B.1	Matériels Supplémentaires associés à la Section 4.3	193
B.1.1	NLMEM log-likelihood	194
B.1.2	Updating the biomarker contributions	195
B.1.3	Simulation Setup	196
B.1.4	Evaluation metrics	199
B.1.5	Additional Tables and Figure	200
B.1.6	Application to COVID-19 mRNA vaccination data	211
B.2	Matériels Supplémentaires associés à la Section 4.4	217
B.2.1	Regularization step details	217
B.2.2	Structure of the project	220
B.2.3	Mlxtran model file for demo monolix project	222
B.2.4	Summary file for initialization task	222
B.2.5	Summary file for grid search	224
B.2.6	Summary file for REMix algorithm	231
B.2.7	Summary file for final tests and estimation	237

A Matériels Supplémentaires associés au Chapitre 3

A.1 Simulation generation process

For all simulated datasets, participant age, when considered, was centered following standard preprocessing practices before model fitting.

A.1.1 Pharmacokinetic model

For the growth model, we used the datasets provided by Naveau *et al.* (2023), available at https://github.com/Marion-Naveau/Supp_Information_SAEMVS/tree/main. The pharmacokinetic model is defined as :

$$y = \frac{D\varphi_1}{V\varphi_1 - \varphi_2} \left(e^{-\varphi_2 t/V} - e^{-\varphi_1 t} \right),$$

where the population parameters include the dose D , the volume V , and the two rate parameters φ_1 and φ_2 . Inter-individual variability is introduced on φ_1 and φ_2 , such that for each individual $i \leq N$:

$$\begin{cases} \varphi_{1i} &= \varphi_{1pop} + \beta_{11}C_1 + \beta_{12}C_2 + \beta_{13}C_3 + \eta_i^1, \\ \varphi_{2i} &= \varphi_{2pop} + \beta_{23}C_3 + \beta_{24}C_4 + \beta_{25}C_5 + \eta_i^2, \end{cases}$$

with $\boldsymbol{\eta}_i = (\eta_i^1, \eta_i^2) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Omega})$.

The error model is constant, yielding the following observation equation :

$$Y_{ij} = \frac{D\varphi_{1i}}{V\varphi_{1i} - \varphi_{2i}} \left(e^{-\varphi_{2i} t/V} - e^{-\varphi_{1i} t} \right) + \varepsilon_{ij},$$

where $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Each of the $N = 200$ individuals is observed at 12 time points :

$$\{0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40\}.$$

The simulation values used are summarized in Table 5.1.

Parameter	Definition	Mean value
Fixed Effects		
D	Dose	100
V	Volume	30
φ_{1pop}	Absorption rate (pop)	6
β_{11}	MF per unit increase in C_1	3
β_{12}	MF per unit increase in C_2	2
β_{13}	MF per unit increase in C_3	1
φ_{2pop}	Elimination rate (pop)	8
β_{23}	MF per unit increase in C_3	3
β_{24}	MF per unit increase in C_4	2
β_{25}	MF per unit increase in C_5	1
Random Effects		
Ω	Covariance matrix for (φ_1, φ_2)	$\begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$
Error Model		
σ^2	Error variance	10^{-3}

TABLE 5.1 – Simulation parameters for the pharmacokinetic model based on Naveau *et al.* (2023) (MF : Multiplicative factor).

A.1.2 Humoral Immune Response

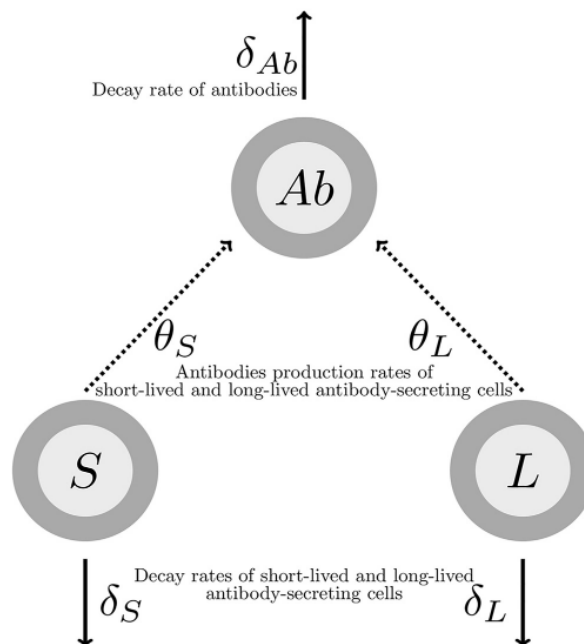


FIGURE 5.1 – Structure of the mechanistic model describing the humoral immune response (extract from Pasin *et al.* (2019)).

To simulate antibody dynamics, we adopted a non-linear mixed effects model based on the work of Pasin *et al.* (2019), previously used to characterize vaccine-induced humoral responses

(Figure 5.1). The model integrates short-lived (S) and long-lived (L) antibody-secreting cells (ASCs), each contributing to antibody production with distinct kinetics.

- Structural Model :

$$\forall i \leq N, j \leq n_i, \begin{cases} \frac{d}{dt} Ab_i(t_{ij}) &= \varphi_{S_i} e^{-\delta_S t_{ij}} + \varphi_{L_i} e^{-\delta_L t_{ij}} - \delta_{Ab_i} Ab_i(t_{ij}) \\ Ab_i(t_{i0} = 0) &= Ab_{0i} \end{cases}$$

$$\begin{cases} \log(\varphi_{S_i}) &= \log(\varphi_{S_{pop}}) + \log(FC\Delta AGE = +1an) \times (AGE_i - \overline{AGE}) + \eta_i^S \\ \log(\varphi_{L_i}) &= \log(\varphi_{L_{pop}}) + \beta_{G_1} \times G_{1i} + \eta_i^L \\ \log(\delta_{Ab_i}) &= \log(\delta_{Ab_{pop}}) + \beta_{G_2} \times G_{2i} + \eta_i^{Ab} \end{cases} \quad (\text{MOD AB STR})$$

- Statistical Model :

$$\forall i \leq N, \begin{cases} \eta_i^S &\overset{iid}{\sim} \mathcal{N}(0, \omega_S^2) \\ \eta_i^L &\overset{iid}{\sim} \mathcal{N}(0, \omega_L^2) \\ \eta_i^{Ab} &\overset{iid}{\sim} \mathcal{N}(0, \omega_{Ab}^2) \end{cases} \quad (\text{MOD AB STAT})$$

- Observation Model :

$$\forall i \leq N, j \leq n_i, \begin{cases} Y_{ij} &= \log_{10}(Ab_i(t_{ij})) + \varepsilon_{ij} \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}(0, \Sigma = \sigma_{Ab}^2 I_{n_i}) \end{cases} \quad (\text{MOD AB OBS})$$

The parameter values were based on estimates from the EBOVAC trial (Eurosurveillance editorial team, 2014; Alexandre *et al.*, 2023; Pasin *et al.*, 2019) and are provided in Table 5.2. Antibody concentrations were simulated at time points $t_j \in \{0, 7, 21, 123, 180, 300\}$.

To create high-dimensional covariate settings, we introduced 200 covariates per individual, including 197 noise variables and 3 informative ones (AGE , G_1 , G_2). Two covariate generation scenarios were considered : one with standard Gaussian covariates, and one with covariates drawn from randomly selected heterogeneous distributions. Details of the covariate simulation process are provided in Section A.1.3.

Parameter	Definition	Value
Fixed Effects		
<i>Influx rates</i>		
φ_S	ASC-S antibody production (Eu/mL/day)	3057 at age 35
FC_{AGE}	Fold change per year of age	0.934
φ_L	ASC-L antibody production (Eu/mL/day)	16.6 when $G_1 = 0$
β_{G_1}	Effect of G_1 on φ_L	1.2
<i>Decay rates</i>		
δ_{Ab}	Antibody decay rate (1/day)	0.0251 when $G_2 = 0$
β_{G_2}	Effect of G_2 on δ_{Ab}	0.8
δ_S	ASC-S decay rate	0.23
δ_L	ASC-L decay rate	$3.16 \cdot 10^{-4}$
Random Effects		
ω_S	SD for φ_S	0.92
ω_L	SD for φ_L	0.85
ω_{Ab}	SD for δ_{Ab}	0.30
Observation Error		
σ_{Ab}	SD of error on log antibody levels	0.11

TABLE 5.2 – Parameter values for the humoral immune response model based on the EBOVAC trial (Alexandre *et al.*, 2023; Pasin *et al.*, 2019). (*FC* : fold change ; *SD* : standard deviation)

A.1.3 Covariates Generation

To simulate realistic dependencies among covariates, we introduced correlation structures reflecting dependencies observed in biomedical datasets. The theoretical correlation matrix used to generate the datasets in the vaccinology framework with 200 covariates is shown in Figure 5.2.

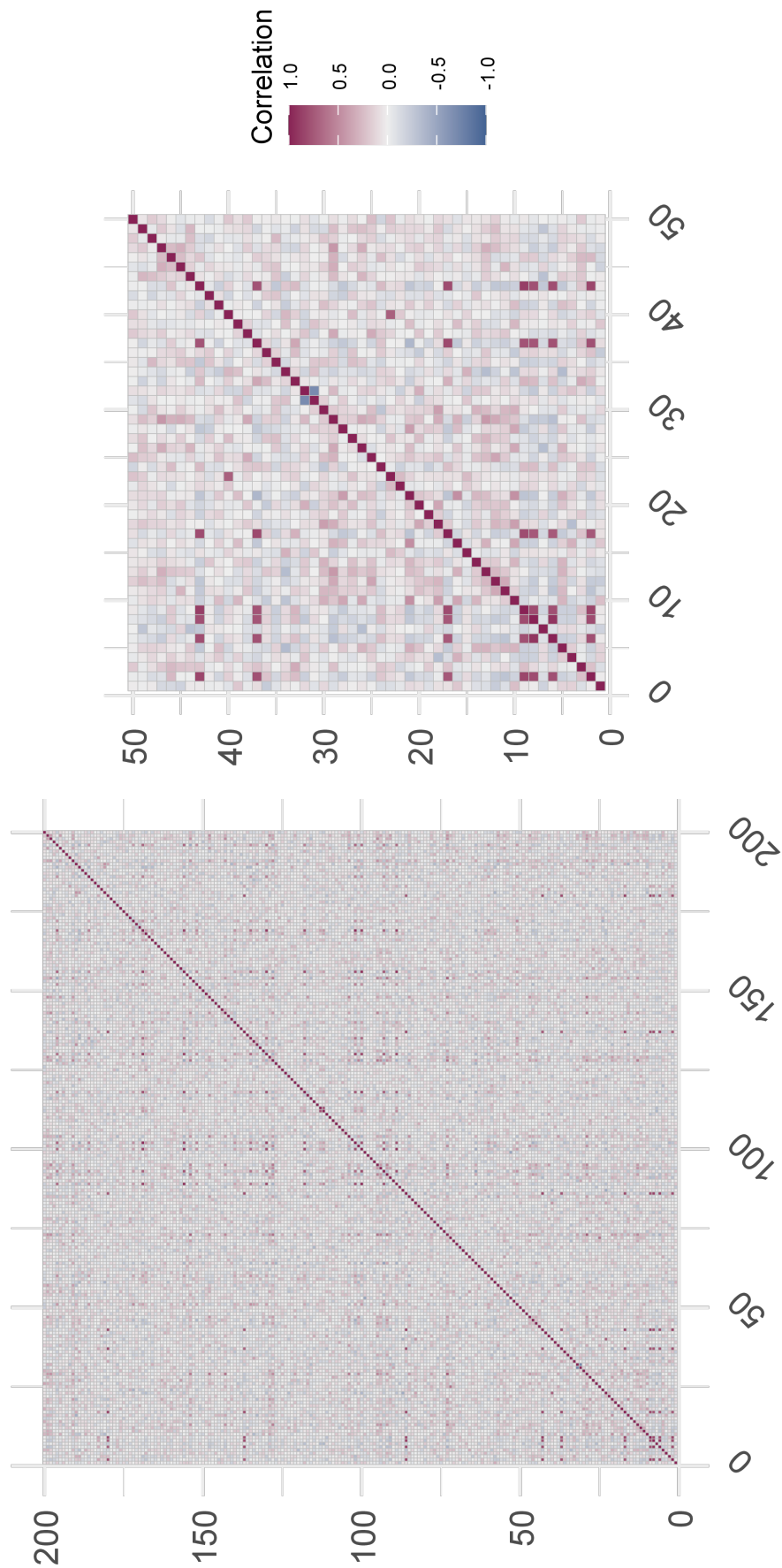


FIGURE 5.2 – Correlation matrix used to generate the 100 datasets in the vaccinology simulation framework with 200 covariates (left : full matrix ; right : zoom on the first 50 covariates).

In the first scenario, we generated all covariates from a multivariate normal distribution with the specified correlation structure using the `mvtnorm` package (Genz et Bretz, 2009) in R. In the second scenario, each covariate was drawn from a randomly selected distribution among a predefined set (Gaussian, Gamma, Poisson, Uniform). To automate this process, we implemented a function called `randomCovariate`, available on GitHub at <https://github.com/aurianegbt/LassoSAMBA>, which returns the name of a distribution and its associated parameters. The parameters are then sampled to define the distribution, and covariates are generated accordingly. Table 5.3 summarizes the rules used to define the parameters of each distribution.

To generate covariates with controlled correlations under the mixed-distribution scenario, we used the `simstudy` package (Goldfeld et Wujciak-Jens, 2020). It enables conditional simulation of variables using a flexible table-based specification.

The data definition table includes the following columns :

- **varname** : name of the variable to simulate.
- **formula** : value or R formula (defining the mean or transformation).
- **variance** : variance or another scale parameter.
- **dist** : name of the distribution (e.g., “gamma”, “norm”).
- **link** : link function between the formula and distribution parameters (identity, log, or logit).

Each variable is conditionally generated using the values or distributions defined above, and the resulting dataset is aligned with the specified correlation matrix using `genCorGen()` from `simstudy`. This approach ensures diverse, realistic high-dimensional covariate structures while maintaining reproducibility.

Suffix for the distribution	Element(s)	Generation
Gamma distribution		
“gamma”	“shape”	$\mathcal{P}(10)$
	“scale”	$\mathcal{E}(1/2)$
Normal Distribution		
“norm”	“mean”	$\mathcal{N}(0, 1)$
	“sd”	$\mathcal{E}(0.1)$
Poisson distribution		
“pois”	“lambda”	$\mathcal{E}(0.1)$
Uniform distribution		
“unif”		$x = \mathcal{N}(0_2, I_2)$
	“min”	$\min(x)$

“max”	max(x)
-------	--------

TABLE 5.3 – Process of generating various distributions using the implemented `randomCovariate` function.

A.2 Additional analysis for the real-data application

A.2.1 Initial model

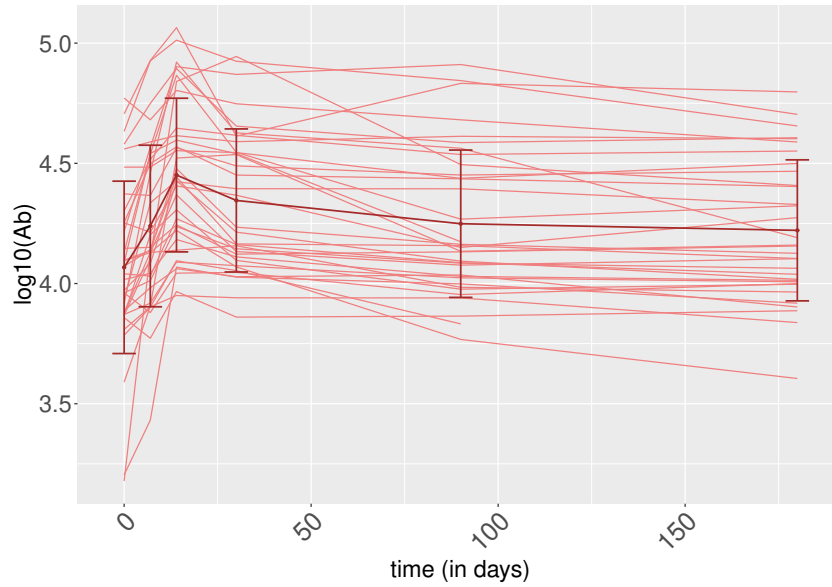


FIGURE 5.3 – Antibody titre trajectories for the 35 individuals receiving the ZOSTAVAX vaccine.

To describe the antibody kinetics observed in Figure 5.3, we fitted the humoral immune response model introduced by [Pasin *et al.* \(2019\)](#). The model structure is as follows :

►► Structural Model :

$$\forall i \leq N, j \leq n_i, \begin{cases} \frac{d}{dt} Ab_i(t_{ij}) &= \varphi_{S_i} e^{-\delta_{S_i} t_{ij}} + \varphi_{L_i} e^{-\delta_{L_i} t_{ij}} - \delta_{Ab} Ab_i(t_{ij}) \\ Ab_i(t_{i0} = 0) &= Ab_{0i} \end{cases}$$

$$\begin{cases} \log(\varphi_{S_i}) &= \log(\varphi_{S_{pop}}) + \eta_i^S \\ \log(\varphi_{L_i}) &= \log(\varphi_{L_{pop}}) + \eta_i^L \\ \log(\delta_{S_i}) &= \log(\delta_{S_{pop}}) + \eta_i^\delta \end{cases} \quad (5.1)$$

►► Statistical Model :

$$\forall i \leq N, \quad \begin{cases} \eta_i^S & \overset{iid}{\sim} \mathcal{N}(0, \omega_{\varphi_S}^2) \\ \eta_i^L & \overset{iid}{\sim} \mathcal{N}(0, \omega_{\varphi_L}^2) \\ \eta_i^\delta & \overset{iid}{\sim} \mathcal{N}(0, \omega_{\delta_S}^2) \end{cases} \quad (5.2)$$

►► Observation Model :

$$\forall i \leq N, j \leq n_i, \quad \begin{cases} Y_{ij} & = \log_{10}(Ab_i(t_{ij})) + \varepsilon_{ij} \\ (\varepsilon_{ij})_{j \leq n_i} & \overset{iid}{\sim} \mathcal{N}(0, \Sigma = \sigma_{Ab}^2) \end{cases} \quad (5.3)$$

To resolve identifiability issues, we fixed the decay rates δ_L (half-life for long-lived ASCs) and δ_{Ab} (half-life for antibodies), based on likelihood profile analyses shown in Figures 5.4 and 5.5.

Figure 5.4 shows that for long-lived ASCs, the log-likelihood stabilizes beyond a half-life of 5 years, indicating a lack of identifiability in the data. We therefore fixed $\delta_L = 0.00019 \text{ day}^{-1}$, corresponding to a 10-year half-life, to avoid identifiability inversion between ASC-S and ASC-L components, given their symmetrical roles in the model.

Regarding the antibody half-life, Figure 5.5 displays the mean values and 95% confidence intervals for fixed antibody half-life values and L-cell half-life values exceeding 5 years. The log-likelihood reaches a maximum around 11 days of half-life. We thus fixed $\delta_{Ab} = 0.063 \text{ day}^{-1}$. The remaining model parameters were estimated via the SAEM algorithm using Monolix, and are presented in Table 5.4.

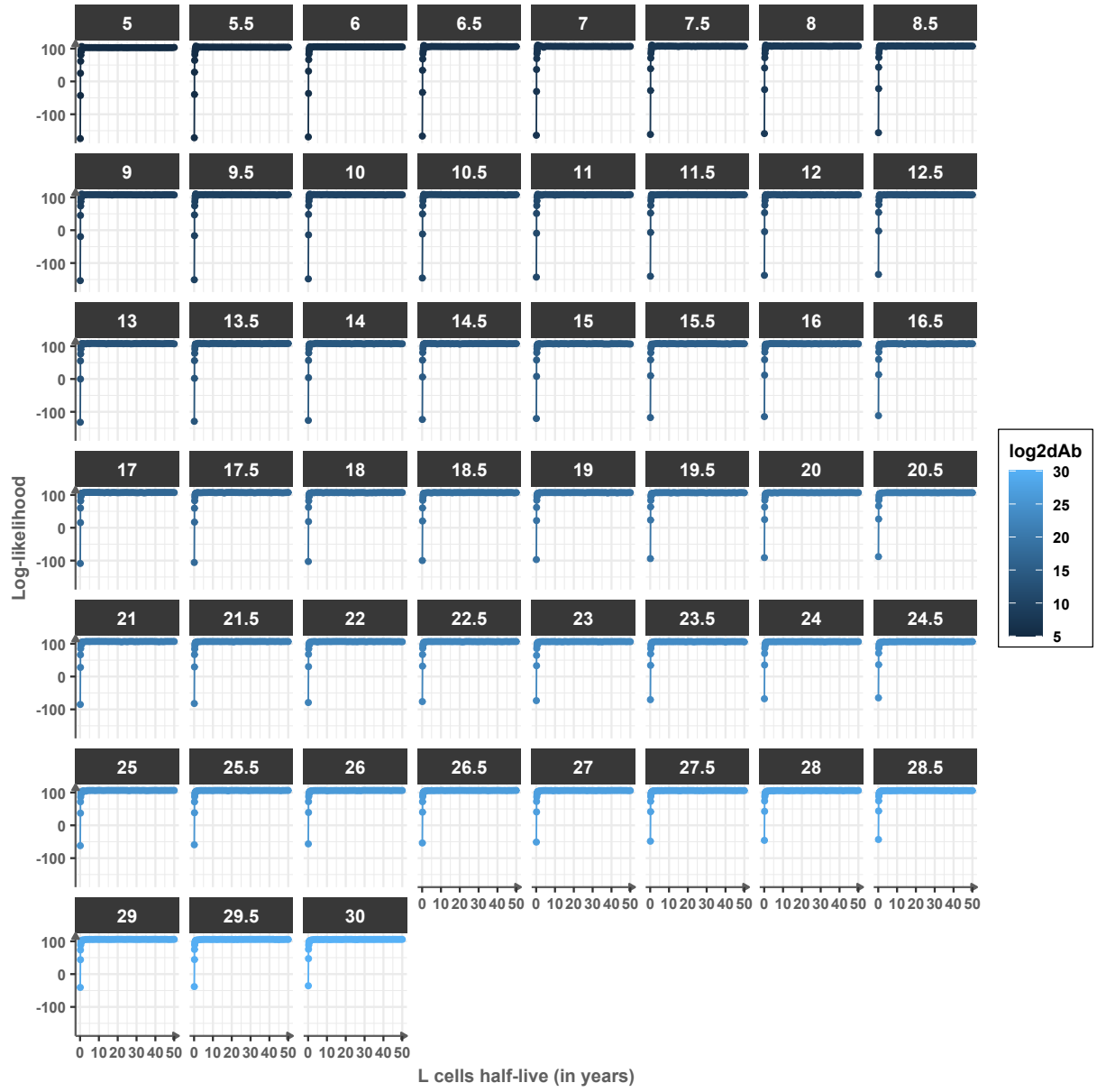


FIGURE 5.4 – Log-likelihood profile as a function of the L-cell half-life $\log(2)/\delta_L$, with antibody half-life fixed; each facet corresponds to an antibody half-life $\log(2)/\delta_{Ab}$.

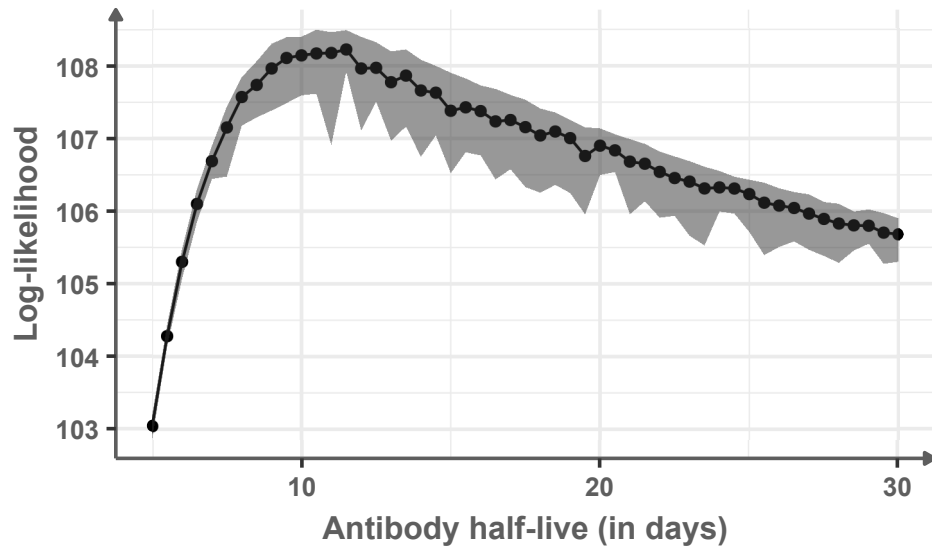


FIGURE 5.5 – Log-likelihood profile as a function of antibody half-life $\log(2)/\delta_{Ab}$. The line and ribbon indicate the mean and 95% confidence intervals over L-cell half-lives greater than 5 years.

Parameter	Definition	Mean value (standard error)
Fixed Effects		
<u>Influx</u>		
$\varphi_{S_{pop}}$	Antibodies influx of ASC S (Eu/mL/day)	907.35 (243.66)
$\varphi_{L_{pop}}$	Antibodies influx of ASC L (Eu/mL/day)	1069.24 (127.66)
<u>Decay rates</u>		
δ_{Ab}	Decay rates of antibodies (per day)	0.063
$\delta_{S_{pop}}$	Decay rate of ASC S (per day)	0.058 (0.017)
δ_L	Decay rate of ASC L (per day)	1.9×10^{-4}
Random Effects		
ω_{φ_S}	Standard deviation of random effects for φ_S	1.16 (0.22)
ω_{φ_L}	Standard deviation of random effects for φ_L	0.67 (0.084)
ω_{δ_S}	Standard deviation of random effects for δ_S	0.50 (0.20)
Error Model		
σ_{Ab}	Standard deviation of error model	0.095 (0.0056)

TABLE 5.4 – Estimation of the initial model parameters, without any covariates, for the ZOSTAVAX data application using the SAEM algorithm ($n = 35$).

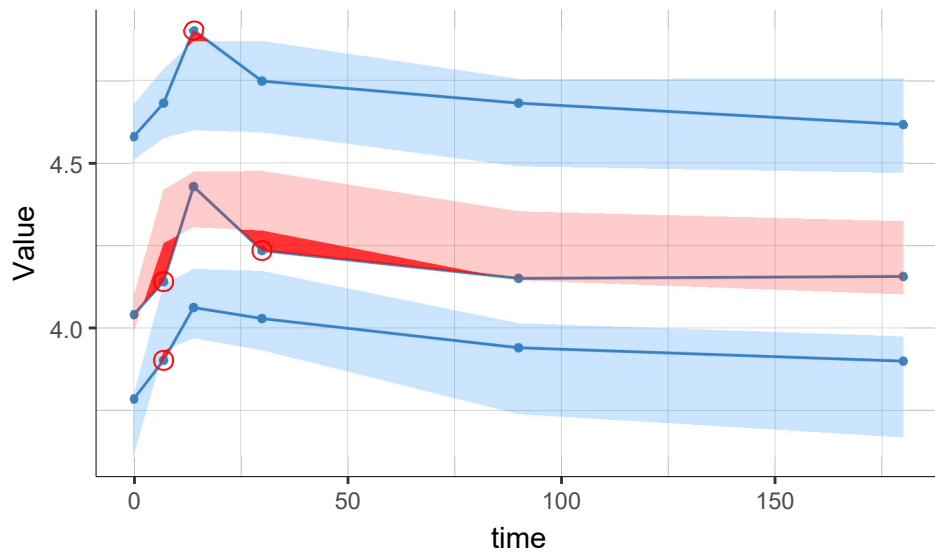


FIGURE 5.6 – Visual predictive check (VPC) of the initial model, without any covariates, for the ZOSTAVAX data application fitted using the SAEM algorithm ($n = 35$). The blue lines indicate the 10th, 50th, and 90th empirical percentiles; ribbons indicate the 90% prediction intervals for each percentile (10th and 90th in blue, 50th in pink). Deviations where empirical percentiles fall outside the prediction intervals are highlighted with red dots and shaded areas.

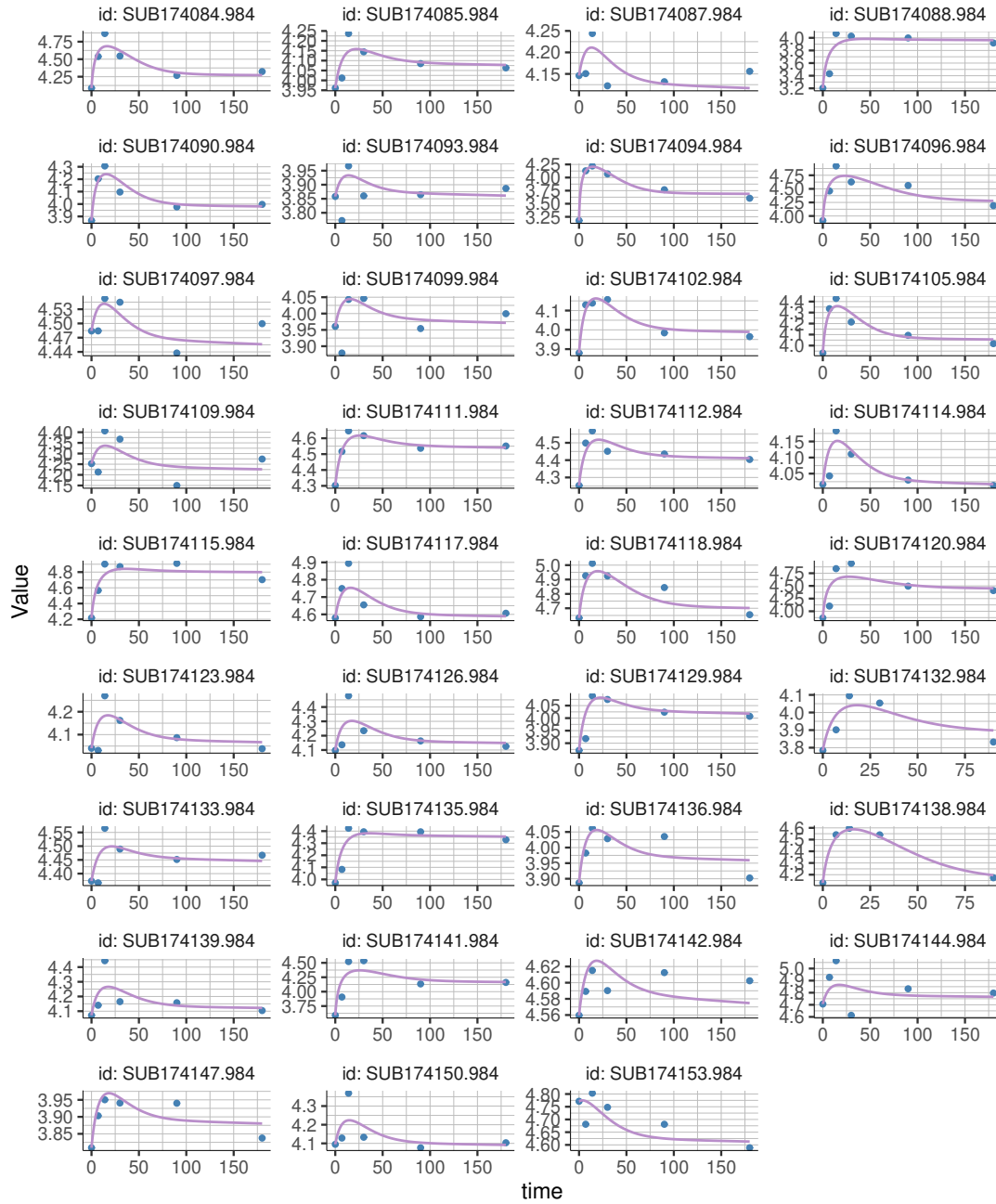


FIGURE 5.7 – Individual fits of the initial model, without any covariates, for the ZOSTAVAX data application fitted using the SAEM algorithm ($n = 35$). For each subject (facet), blue dots indicate observations and the purple line shows model predictions.

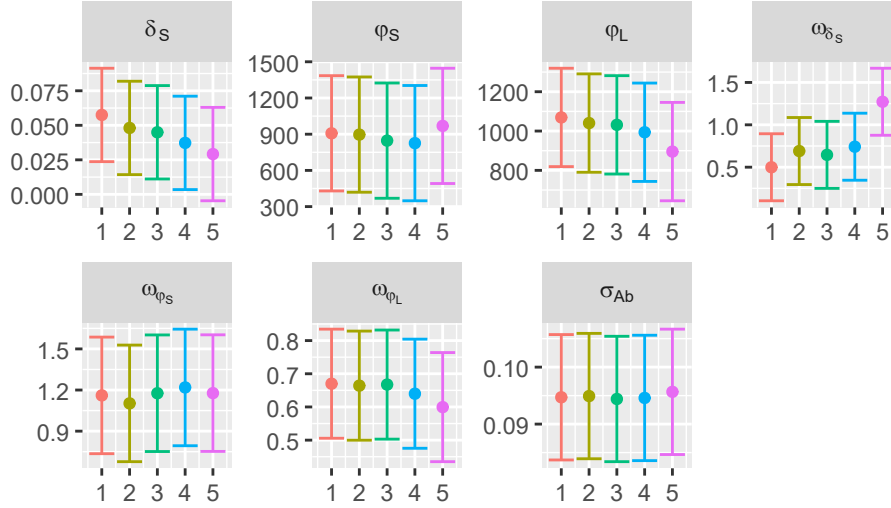


FIGURE 5.8 – Convergence diagnostics for the initial model, without any covariates, for the ZOSTAVAX data application : population parameter estimates across 5 independent Monolix runs of the SAEM algorithm with 95% confidence intervals.

A.2.2 Final Model

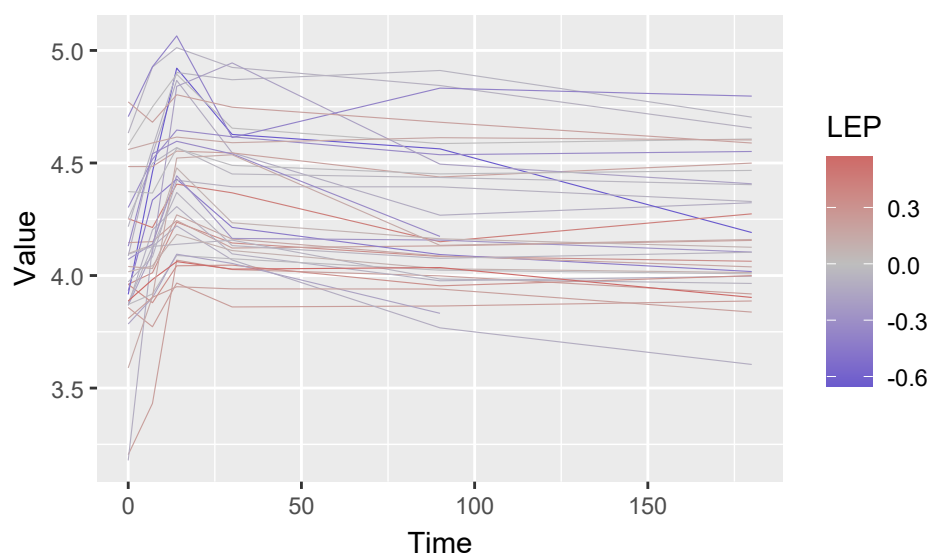
The final model was obtained using the Lasso-SAMBA procedure, which selected two genes : **LEP**, associated with φ_S , and **KIFC1**, associated with φ_L :

$$\log(\varphi_{Si}) = \log(\varphi_{Spop}) + \beta_{LEP} \times X_{LEPi} + \eta_i^S,$$

and

$$\log(\varphi_{Li}) = \log(\varphi_{Lpop}) + \beta_{KIFC1} \times X_{KIFC1i} + \eta_i^L.$$

Parameters were re-estimated via SAEM in Monolix and are summarized in Table 5.5. Gene expression data were centered before estimation. Figure 5.9 shows the antibody data stratified by gene expression levels.



Parameter	Definition	Mean value (standard error)
Fixed Effects		
<u>Influx</u>		
$\varphi_{S_{pop}}$	Antibodies influx of ASC S (Eu/mL/day)	945.43 (193.78)
β_{LEP}		-3.40 (0.71)
$\varphi_{L_{pop}}$	Antibodies influx of ASC L (Eu/mL/day)	1007.21 (105.63)
β_{KIFC1}		-2.05 (0.43)
<u>Decay rates</u>		
δ_{Ab}	Decay rates of antibodies (per day)	0.063
$\delta_{S_{pop}}$	Decay rate of ASC S (per day)	0.047 (0.018)
δ_L	Decay rate of ASC L (per day)	1.9×10^{-4}
Random Effects		
ω_{φ_S}	Standard deviation of random effects for φ_S	0.83 (0.15)
ω_{φ_L}	Standard deviation of random effects for φ_L	0.66 (0.089)
ω_{δ_S}	Standard deviation of random effects for δ_S	0.72 (0.27)
Error Model		
σ_{Ab}	Standard deviation of error model	0.094 (0.0056)

TABLE 5.5 – Estimation of the final model parameters for the ZOSTAVAX data application using the Lasso-SAMBA method ($n = 35$).

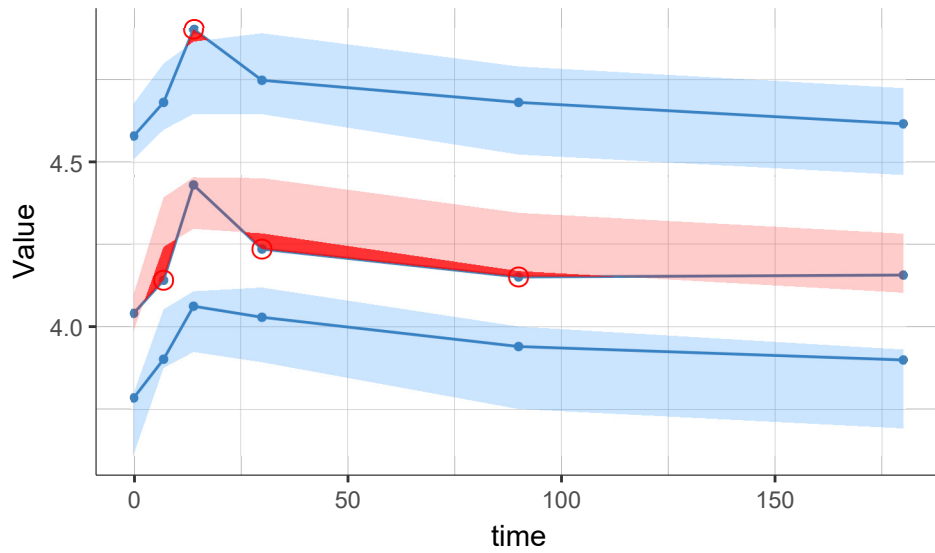


FIGURE 5.10 – Visual predictive check (VPC) of the final model built with the Lasso-SAMBA method for the ZOSTAVAX data application fitted using the SAEM algorithm ($n = 35$). The blue lines indicate the 10th, 50th, and 90th empirical percentiles; ribbons indicate the 90% prediction intervals for each percentile (10th and 90th in blue, 50th in pink). Deviations where empirical percentiles fall outside the prediction intervals are highlighted with red dots and shaded areas.

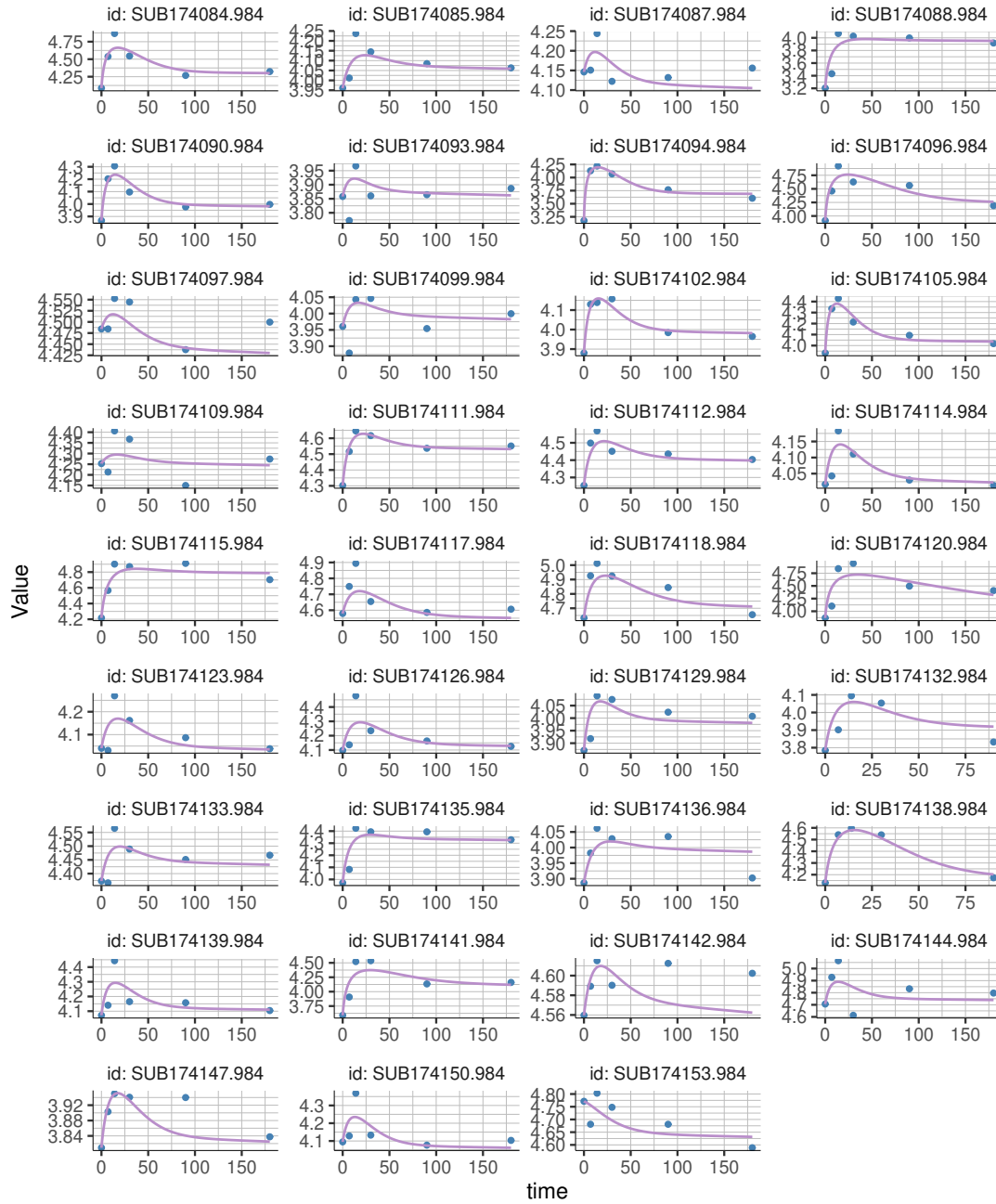


FIGURE 5.11 – Individual fits of the final model built with the Lasso-SAMBA method for the ZOSTAVAX data application fitted using the SAEM algorithm ($n = 35$). For each subject (facet), blue dots indicate observations and the purple line shows model predictions.

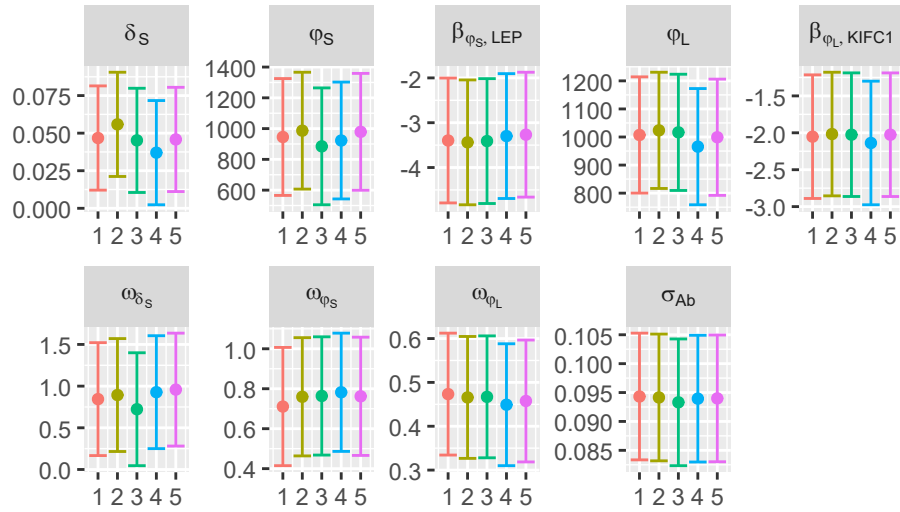


FIGURE 5.12 – Convergence diagnostics for the final model built with the Lasso-SAMBA method for the ZOSTAVAX data application : population parameter estimates across 5 independent Monolix runs of the SAEM algorithm with 95% confidence intervals.

A.3 Additional graphs for results

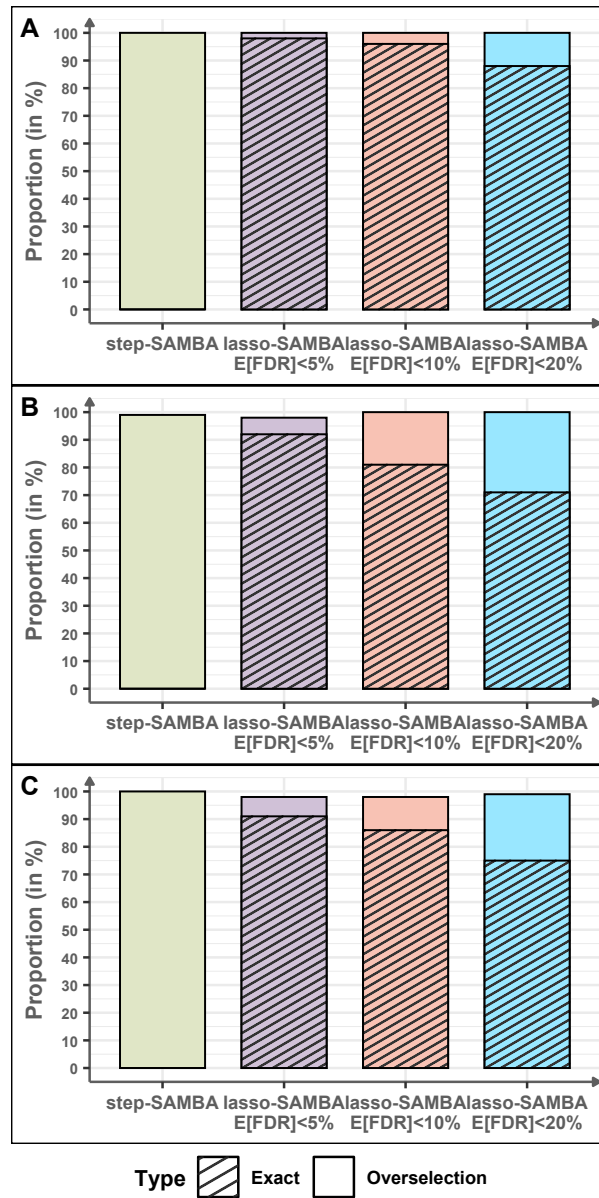


FIGURE 5.13 – Proportion of exact models (no false positives or false negatives) and models that strictly include the exact model (no false negatives but with false positives) for step-SAMBA and Lasso-SAMBA under different control thresholds. (A) Pharmacokinetic model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

Rate	PK model binomial covariates	Vaccinology model Gaussian covariates	Vaccinology model non-Gaussian covariates
FDR (%) :			
- step-SAMBA	87.5 [78.8;91.7]	72.7 [50.0;83.8]	70.0 [25.0;81.2]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	0.0 [0.0;0.0]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	0.0 [0.0;14.3]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	0.0 [0.0;14.3]	0.0 [0.0;40.0]	0.0 [0.0;25.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
FNR (%) :			
- step-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
F1-score (%) :			
- step-SAMBA	22.2 [15.4;34.9]	42.9 [27.9;66.7]	46.2 [31.6;85.7]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	100.0 [100.0;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	100.0 [92.3;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- Lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	100.0 [92.3;100.0]	100.0 [75.0;100.0]	100.0 [85.7;100.0]
- SAEMVS	100.0 [0.0;100.0]	ND	ND

TABLE 5.6 – False Discovery Rate (FDR), False Negative Rate (FNR), and F_1 -score computed for each simulation framework, comparing the original step-SAMBA method and the Lasso-SAMBA method under different error control thresholds. (ND : not defined).

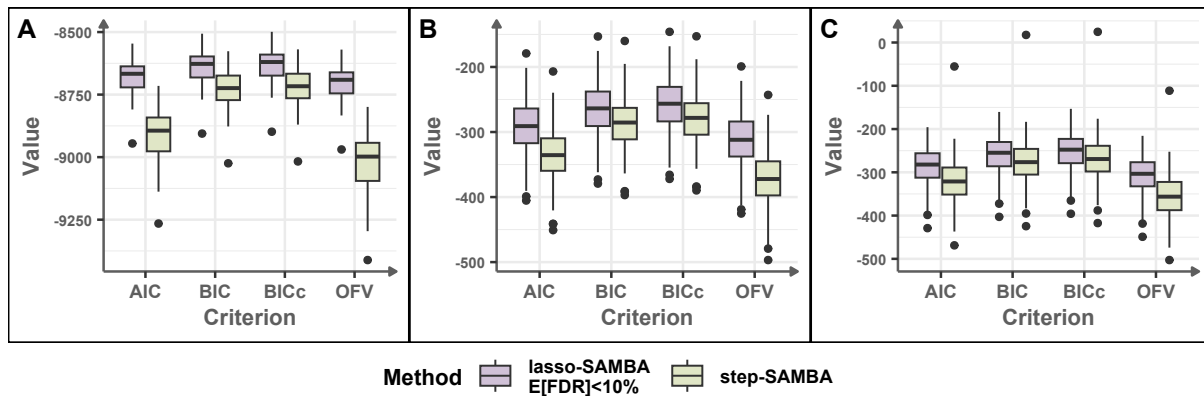


FIGURE 5.14 – Distribution of final model selection scores (AIC, BIC, BICc, OFV) across replicates in each framework, comparing step-SAMBA and Lasso-SAMBA with the error control threshold set at 10%. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

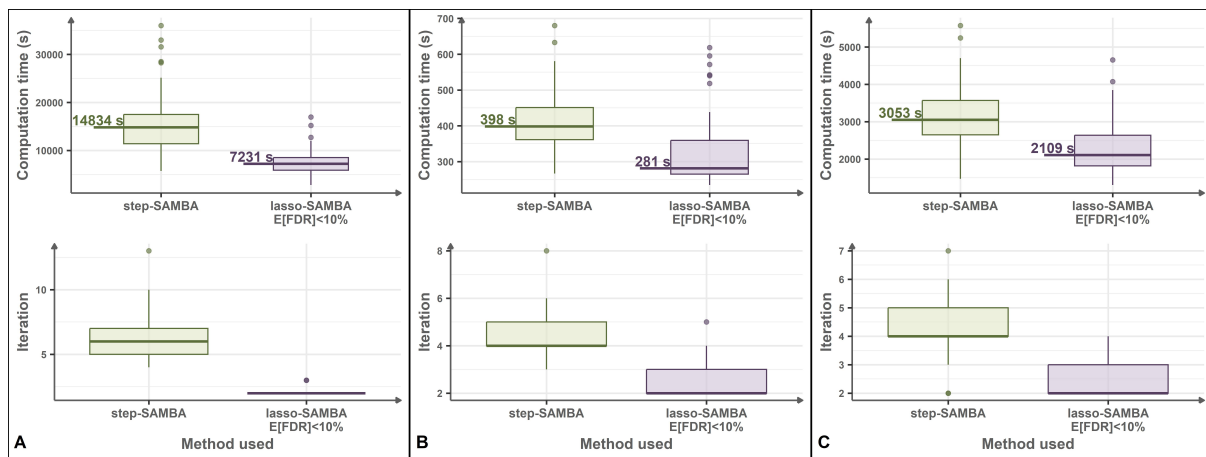


FIGURE 5.15 – Comparison of computation time distributions across simulation frameworks, for step-SAMBA and Lasso-SAMBA with the error control threshold set at 10%. (A) Pharmacokinetics model with categorical covariates ; (B) Vaccinology framework with Gaussian-correlated covariates ; (C) Vaccinology framework with randomly drawn covariates.

A.4 Figures additionnelles concernant l'analyse sur les données PREVAC-UP

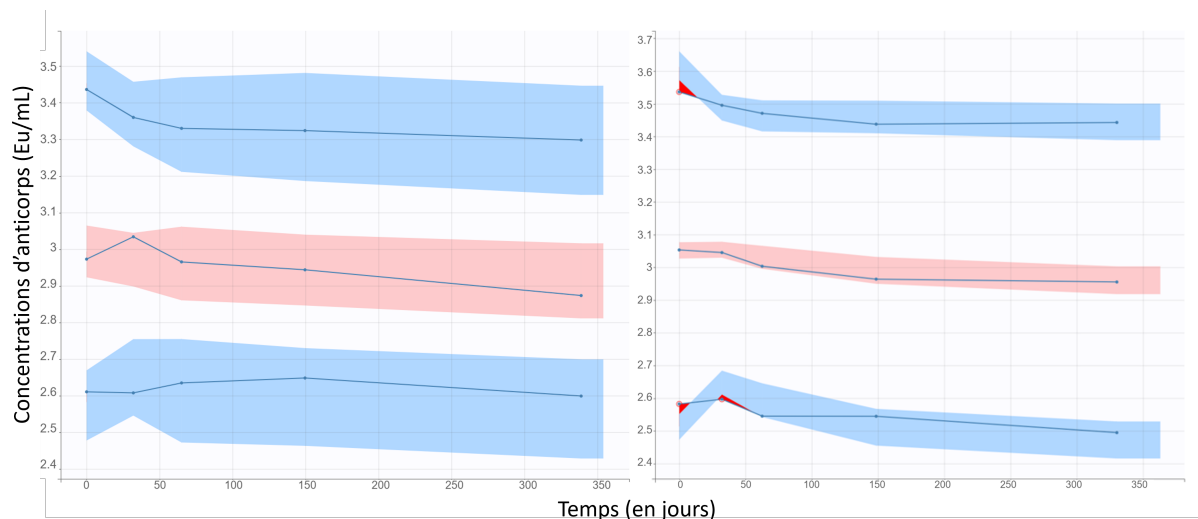
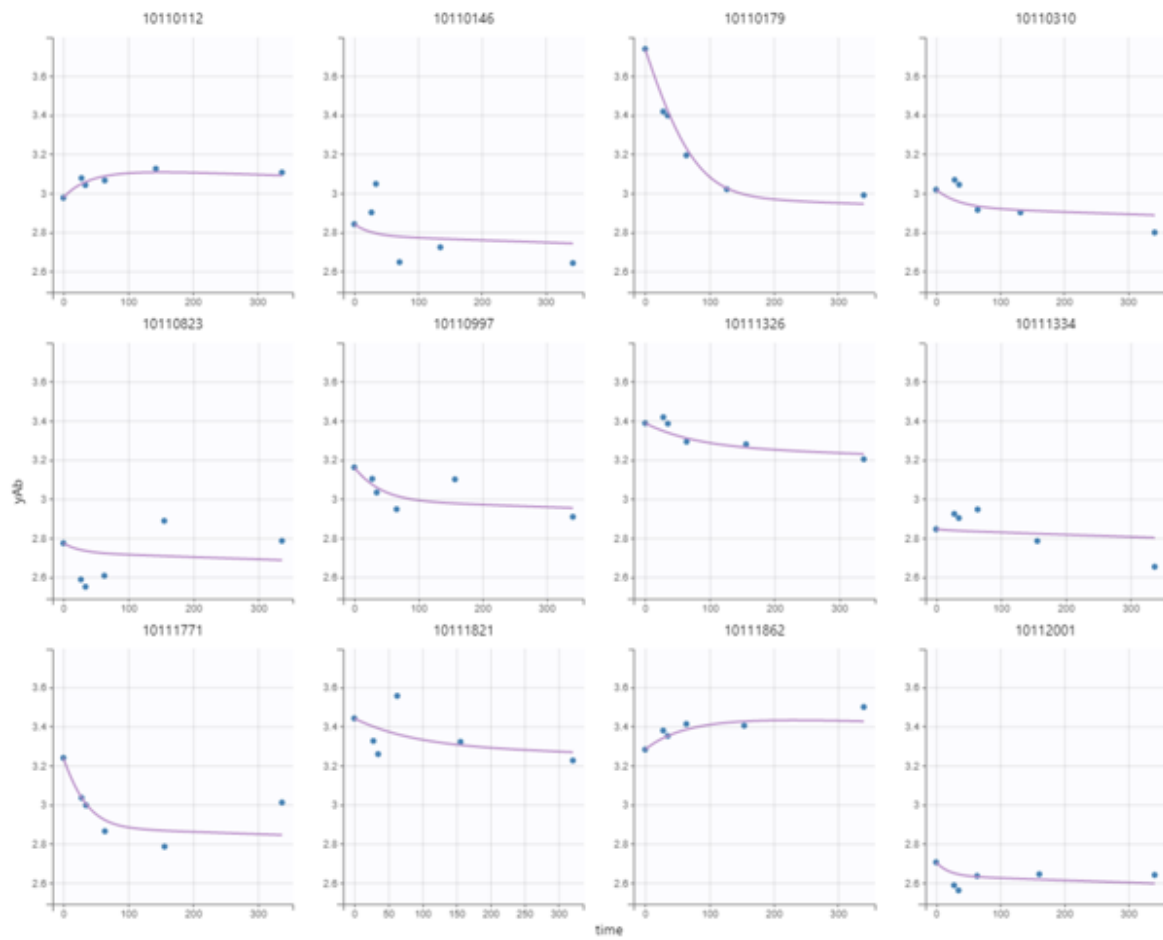


FIGURE 5.16 – VPC du modèle initial de la réponse humorale, sans effets de covariables, ajusté aux données des individus ayant reçu une dose de rVSV-ZEBOV dans l'essai clinique PREVAC-UP, avec l'algorithme SAEM ($n=395$) à gauche et du modèle final, incluant le gène FAM216A comme covariable influençant la décroissance des anticorps, ajusté aux données des 26 individus de la sous-étude immunologique. Les lignes bleues représentent les 10e, 50e et 90e percentiles empiriques des observations ; les rubans indiquent les intervalles de prédiction à 90% pour chaque percentile (10e et 90e en bleu, 50e en rose). Les points et zones rouges signalent des valeurs empiriques situées en dehors des intervalles de prédiction.



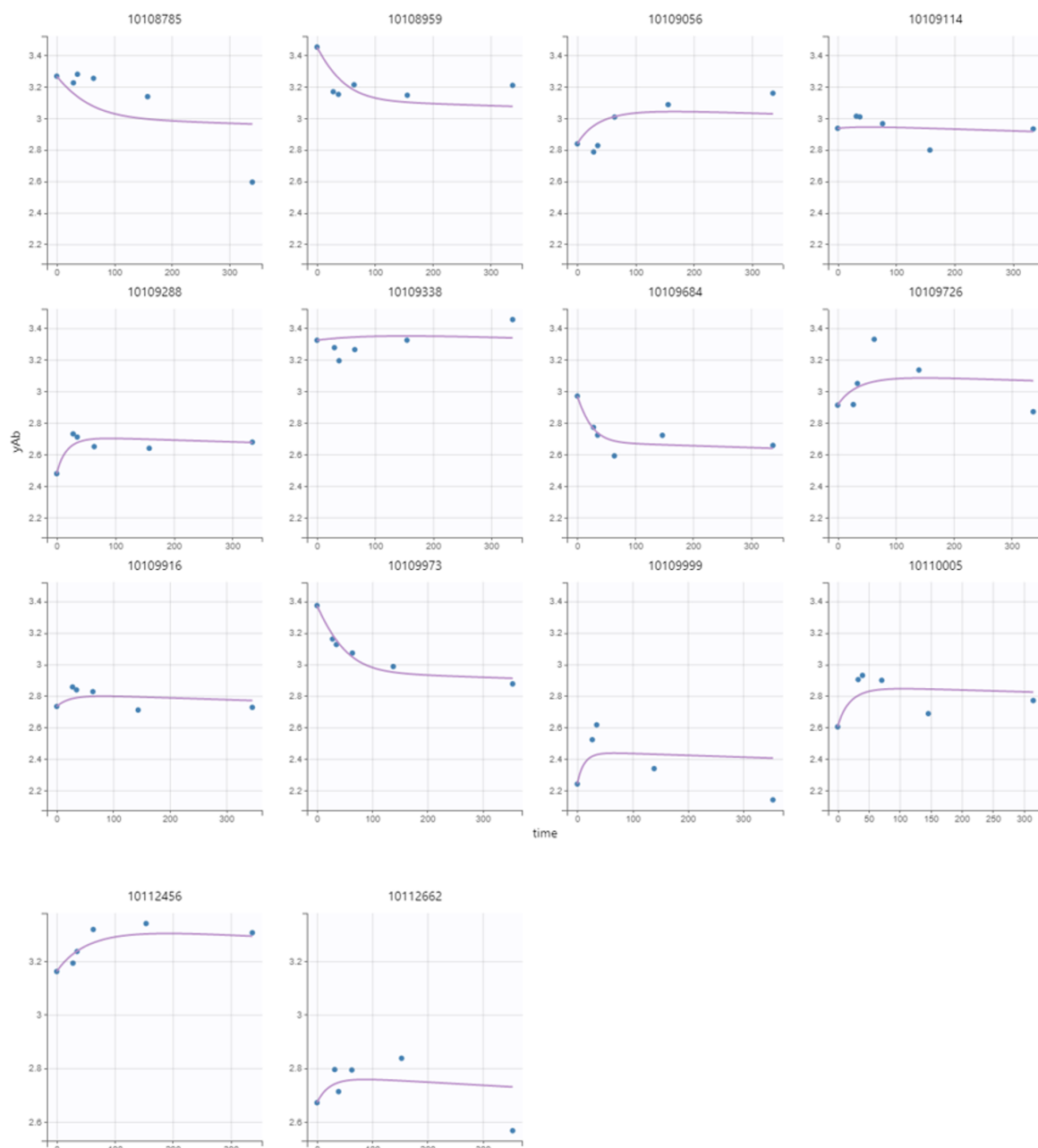


FIGURE 5.17 – Ajustements individuels du modèle final incluant le gène FAM216A, obtenus par SAEM sur les données PREVAC-UP ($n=26$). Chaque panneau correspond à un sujet : les points bleus représentent les mesures observées ($\log_{10}$ d’anticorps), et la ligne violette correspond aux trajectoires prédites par le modèle.

B Matériels Supplémentaires associés au Chapitre 4

B.1 Matériels Supplémentaires associés à la Section 4.3

B.1.1 NLMEM log-likelihood

For individual $i \leq N$, the contribution to the likelihood is

$$\mathcal{L}_i(\alpha, \theta) = f(\mathbf{Y}_i, \mathbf{Z}_i \mid \alpha, \theta) = \int_{\boldsymbol{\eta}_i} f(\mathbf{Y}_i, \mathbf{Z}_i \mid \alpha, \theta, \boldsymbol{\eta}_i) \times f(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i.$$

with $\mathbf{Y}_i = (\mathbf{Y}_{pi})_{p \leq P} = (Y_{pij})_{p \leq P, j \leq n_{ip}}$ and $\mathbf{Z}_i = (\mathbf{Z}_{ki})_{k \leq K} = (Z_{kij})_{k \leq K, j \leq n_{ik}}$ the vectors of observations defined in Equations (1) and (2) respectively.

The latent compartment $(S_{pi})_{p \leq P}$ and R_i , and by extensions their observations $\mathbf{Y}_i$ and $\mathbf{Z}_i$, are independent conditionally to the individual random effects so the log-likelihood can be rewritten :

$$\begin{aligned} \mathcal{L}_i(\alpha, \theta) &= \int_{\boldsymbol{\eta}_i} \prod_{p \leq P} (f(\mathbf{Y}_{pi} \mid \alpha, \theta, \boldsymbol{\eta}_i)) \times \prod_{k \leq K} (f(\mathbf{Z}_{ki} \mid \alpha, \theta, \boldsymbol{\eta}_i)) \times f(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i \\ &= \int_{\boldsymbol{\eta}_i} \prod_{p \leq P} \prod_{j \leq n_{pi}} f(Y_{pij} \mid \alpha, \theta, \boldsymbol{\eta}_i) \times \prod_{k \leq K} \prod_{j \leq n_{ki}} f(Z_{kij} \mid \alpha, \theta, \boldsymbol{\eta}_i) \times f(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i \\ &= \int_{\boldsymbol{\eta}_i} \left\{ \prod_{p \leq P} \prod_{j \leq n_{pi}} \frac{1}{s_p \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{pij} - g_p(S_{pi}(t_{pij}))}{s_p} \right)^2 \right] \right\} \\ &\quad \times \left\{ \prod_{k \leq K} \prod_{j \leq n_{ki}} \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))}{\sigma_k} \right)^2 \right] \right\} \\ &\quad \times \frac{1}{(2\pi)^{\frac{m}{2}} \det \boldsymbol{\Omega}^{\frac{1}{2}}} \exp \left(-\frac{1}{2} \boldsymbol{\eta}_i^T \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}_i \right) d\boldsymbol{\eta}_i, \end{aligned}$$

We will concentrate on the regularisation parameters α , so we denote the part of our individual likelihood not depending on α by $f_Y(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}_i)$, and the one depending on by $f_Z(\mathbf{Z}_i \mid \alpha, \boldsymbol{\theta}, \boldsymbol{\eta}_i)$. This gives :

$$\begin{aligned} f_Y(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}_i) &= \prod_{p \leq P} \prod_{j \leq n_{pi}} \frac{1}{s_p \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{pij} - g_p(S_{pi}(t_{pij}))}{s_p} \right)^2 \right], \\ f_Z(\mathbf{Z}_i \mid \boldsymbol{\theta}, \alpha, \boldsymbol{\eta}_i) &= \prod_{k \leq K} \prod_{j \leq n_{ki}} \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Z_{kij} - \mu_k - \alpha_k s_k(R_i(t_{kij}))}{\sigma_k} \right)^2 \right], \end{aligned}$$

and the density of random effects $\boldsymbol{\eta}_i$:

$$f(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) = f(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{m}{2}} \det \boldsymbol{\Omega}^{\frac{1}{2}}} \exp \left(-\frac{1}{2} \boldsymbol{\eta}_i^T \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}_i \right),$$

and so

$$\mathcal{L}_i(\alpha, \theta) = \int_{\boldsymbol{\eta}_i} f_Y(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}_i) f_Z(\mathbf{Z}_i \mid \alpha, \boldsymbol{\theta}, \boldsymbol{\eta}_i) f(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}) d\boldsymbol{\eta}_i.$$

Now, as individuals are independent, we can then define the log-likelihood LL of the considered model as :

$$\begin{aligned} LL(\alpha, \theta) &= \sum_{i \leq N} LL_i(\alpha, \theta) = \sum_{i \leq N} \log(\mathcal{L}_i(\alpha, \theta)) \\ &= \sum_{i \leq N} \log \left(\int_{\eta_i} f_Y(Y_i | \theta, \eta_i) f_Z(Z_i | \alpha, \theta, \eta_i) f(\eta_i | \theta) d\eta_i \right). \end{aligned}$$

Ultimately, only the first factor depends on the regularisation parameters α and biomarkers.

B.1.2 Updating the biomarker contributions

For the second coordinate, with fixed $\hat{\theta}$, the problem consists of solving

$$\max_{\alpha \in \mathbb{R}^K} (LL(\alpha, \hat{\theta}) - \lambda \|\alpha\|_1).$$

This corresponds to a Lasso optimization problem that we solve using a second order Taylor development.

For a fixed $\hat{\theta}$ and $\hat{\alpha}$, we have

$$LL(\alpha, \hat{\theta}) \approx LL(\hat{\alpha}, \hat{\theta}) + (\alpha - \hat{\alpha})^T \nabla_{\alpha} LL(\hat{\alpha}, \hat{\theta}) + \frac{1}{2} (\alpha - \hat{\alpha})^T \mathbf{H} (\alpha - \hat{\alpha}),$$

with $\mathbf{H} = \nabla_{\alpha}^2 LL(\hat{\alpha}, \hat{\theta}) = -\hat{\mathbf{X}} = -(\hat{x}_{lc})_{l,c \leq K}$ the Hessian matrix, supposed define negative. The equation then can be written as

$$\partial_{\alpha_k} LL(\hat{\alpha}, \hat{\theta}) - \sum_{l \neq k} \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_c \hat{x}_{kc} \hat{\alpha}_c - \lambda \text{sign}(\alpha_k) = 0.$$

We want to find α_k so that $LL(\alpha, \hat{\theta}) - \lambda \|\alpha\|_1$ reaches a maximum for this value. This corresponds to solving in α_k the equation

$$\begin{aligned} \partial_{\alpha_k} LL(\alpha, \hat{\theta}) - \lambda \text{sign}(\alpha_k) &= 0, \\ \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda \text{sign}(\alpha_k) &= 0. \end{aligned}$$

If $\alpha_k > 0$:

$$\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda = 0,$$

$$\alpha_k = \frac{\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda}{x_{kk}}.$$

So the condition $\alpha_k > 0$ implies that $\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c > \lambda$.

If $\alpha_k < 0$:

$$\begin{aligned} \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c + \lambda &= 0, \\ \alpha_k &= \frac{\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c + \lambda}{x_{kk}}. \end{aligned}$$

So the condition $\alpha_k < 0$ implies that $\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c < -\lambda$.

As a consequence, the condition $\left| \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c \right| < \lambda$ implies that α_k is set to 0.

We denote $\left(\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \hat{\alpha}_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c \right) = \left(\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} + \hat{x}_{kk} \hat{\alpha}_k \right) = A$, that approximate the previous expression. To summarize, we update the regularisation parameters α_k , for $k \leq K$ by :

$$\alpha_k = \begin{cases} \frac{A + \lambda}{x_{kk}} & \text{if } A < -\lambda, \\ \frac{A - \lambda}{x_{kk}} & \text{if } A > \lambda, \\ 0 & \text{otherwise.} \end{cases} \quad (5.4)$$

B.1.3 Simulation Setup

The antibody-secreting cell (ASC) compartment S_i and antibody compartment Ab_i dynamics for individual $i \leq 25$ are governed by the system of differential equations :

$$\begin{cases} \frac{dS_i}{dt}(t) &= -\delta_{Si}S_i(t) \\ \frac{dAb_i}{dt}(t) &= \varphi_{Si}S_i(t) - \delta_{Abi}Ab_i(t) \end{cases}$$

Individual parameters φ_{Si} , δ_{Si} , and δ_{Abi} follow a log-normal distribution incorporating random effects :

$$\begin{cases} \log(\varphi_{Si}) &= \log(\varphi_{S_{pop}}) + \eta_i^{\varphi_S} \\ \log(\delta_{Si}) &= \log(\delta_{S_{pop}}) + \eta_i^{\delta_S} \\ \log(\delta_{Abi}) &= \log(\delta_{Ab_{pop}}) + \eta_i^{\delta_{Ab}} \end{cases}$$

with $(\eta_i^{\varphi_S})_i \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\varphi_S}^2)$, $(\eta_i^{\delta_S})_i \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\delta_S}^2)$, and $(\eta_i^{\delta_{Ab}})_i \stackrel{iid}{\sim} \mathcal{N}(0, \omega_{\delta_{Ab}}^2)$.

Observations for antibodies and biomarkers are modeled as follows :

$$\begin{cases} Y_{ij} &= \log_{10}(Ab_i(t_j)) + \epsilon_{ij}, & \forall j \leq 6 \\ Z_{kij} &= \mu_k + \alpha_k \times S_i(j) + \varepsilon_{kij}, & \forall j \leq 21, k \leq K_1 \\ Z_{kij} &= \mu_k + \varepsilon_{kij}, & \forall j \leq 21, K_1 < k \leq K = K_1 + K_2 \end{cases}$$

with measurement errors $(\epsilon_{ij}) \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{Ab}^2)$ and $(\varepsilon_{kij}) \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$.

Parameters for biomarker models are sampled as follows :

- $(\mu_k)_{k \leq K} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$,
- $(\alpha_{1k})_{k \leq K_1} \stackrel{iid}{\sim} \mathcal{U}([-1, 1])$,
- $(\sigma_k)_{k \leq K} \stackrel{iid}{\sim} \mathcal{N}(0.3, 0.05^2)$.

Population parameter values for the structural biological model are provided in Table 5.7 and 5.8.

Parameter	Definition	Value
Fixed Effects		
<i>Influx rates</i>		
φ_S	ASC-S antibody production (Eu/mL/day)	611
<i>Decay rates</i>		
δ_{Ab}	Antibody decay rate (1/day)	0.025
δ_S	ASC-S decay rate	0.23
Random Effects		
ω_{φ_S}	SD for φ_S	0.60
ω_{δ_S}	SD for δ_S	0.1
$\omega_{\delta_{Ab}}$	SD for δ_{Ab}	0.30
Biomarkers related parameters		
$\alpha_k, k \leq 5$		0.65, -0.49, 1.06, -0.72, 0.33
$\alpha_k, k > 5$		0
$\mu_k, k \leq 5$		-0.61, 0.04, -1.8, 2.4, 1.09
$\mu_k, k > 5$		[-2.270; 0.8]
Observation Error		
σ_{Ab}	SD of error on log antibody levels	0.11
$\sigma_k, k \leq 5$	SD of error on biomarkers levels	0.310, 0.310, 0.350, 0.350, 0.260
$\sigma_k, k > 5$		[0.23; 0.42]

TABLE 5.7 – Parameter values for the humoral immune response model based on the EBOVAC trial [Alexandre et al. \(2023\)](#); [Pasin et al. \(2019\)](#) for the 20-gene scenario.

Parameter	Definition	Value
Fixed Effects		
<i>Influx rates</i>		
φ_S	ASC-S antibody production (Eu/mL/day)	611 at age 35
<i>Decay rates</i>		
δ_{Ab}	Antibody decay rate (1/day)	0.025 when $G_2 = 0$
δ_S	ASC-S decay rate	0.23
Random Effects		
ω_{φ_S}	SD for φ_S	0.60
ω_{δ_S}	SD for δ_S	0.1
$\omega_{\delta_{Ab}}$	SD for δ_{Ab}	0.30
Biomarkers related parameters		
$\alpha_k, k \leq 10$		1.06, -0.72, 0.33, -0.28, 1.55, -0.24, -1.05, -1.79, 0.76, -0.35
$\alpha_k, k > 10$		0
$\mu_k, k \leq 5$		-0.17, -0.45, 0.5, -2.27, -2.02, 0.19, 0.27, 0.98, 0.91, 1.47
$\mu_k, k > 5$		[-2.03; 2.46]
Observation Error		
σ_{Ab}	SD of error on log antibody levels	0.11
$\sigma_k, k \leq 5$	SD of error on biomarkers levels	0.23, 0.38, 0.31, 0.28, 0.30, 0.30, 0.98, 0.23, 0.31, 0.38
$\sigma_k, k > 5$		[0.17; 0.37]

TABLE 5.8 – Parameter values for the humoral immune response model based on the EBOVAC trial [Alexandre et al. \(2023\)](#); [Pasin et al. \(2019\)](#) for the 50-gene scenario

B.1.4 Evaluation metrics

For a true parameter θ , for R replicates, we consider the following metrics :

- **Mean relative bias** : $\frac{1}{R} \sum_{r=1}^R \left(\frac{\hat{\theta}^{(r)} - \theta}{\theta} \right)$,
where $\hat{\theta}^{(r)}$ is the final estimate from replicate r .
- **Empirical standard deviation** : $\sqrt{\frac{1}{R-1} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \bar{\hat{\theta}} \right)^2}$,
with $\bar{\hat{\theta}} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)}$.
- **Estimated standard error** : $\frac{1}{R} \sum_{r=1}^R \widehat{\text{se}}(\theta)^{(r)}$,
with $\widehat{\text{se}}(\theta)^{(r)}$ the estimated standard error of replicate r .

- **Coverage probability of 95% confidence interval :** $\frac{1}{R} \sum_{r=1}^R \mathbb{I} \left(\theta \in \left[\hat{\theta}^{(r)} \pm q_{97.5} \cdot \widehat{\text{se}}(\theta)^{(r)} \right] \right)$.
with $q_{97.5}$ the 97.5% quantile of the standard Gaussian distribution.

B.1.5 Additional Tables and Figure

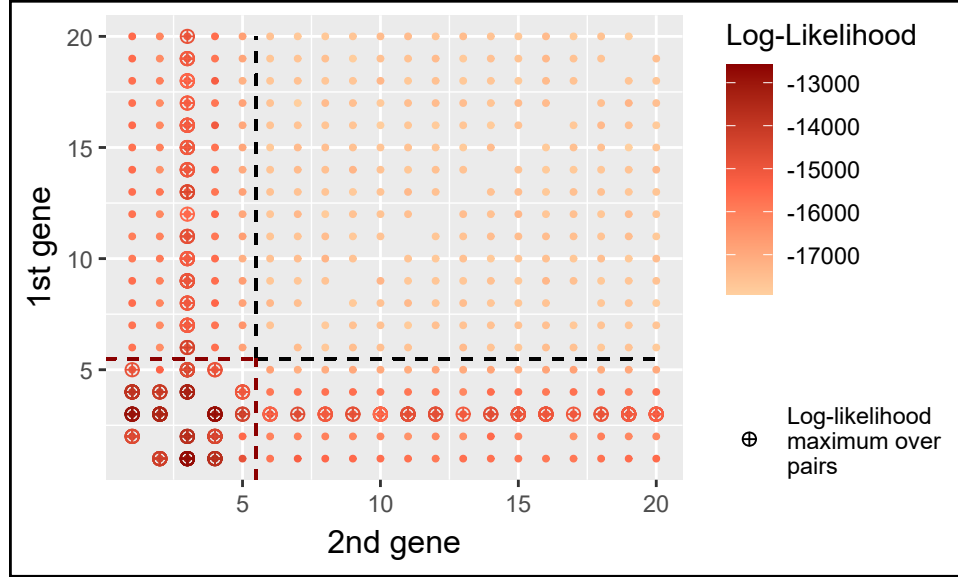


FIGURE 5.18 – Maximum log-likelihood of the mechanistic model with 20 genes but only considering a pair of them as relevant throughout the 200 replicates. Best pairs selected over all 10 pairs for each replicates tested is marked by a rounded cross. The dashed lines delimit the set of significant genes, range from 1 to 5 on each axis.

Parameter estimation accuracy for population and biomarker-specific parameters for the 20-gene scenario with additional tests

Parameter	Target Value	Mean estimate	Relative Bias (%)	Empirical SD	Estimated SD	95% Coverage
δ_{Ab}	0.025	0.025	0.50	0.0017	0.0015	92.0
δ_S	0.230	0.229	-0.24	0.0070	0.0063	91.5
$\varphi_S \times 10^{-2}$	6.110	6.150	0.62	0.0078	0.0075	90.0
$\omega_{\delta_{Ab}}$	0.300	0.288	-3.92	0.0395	0.0419	92.5
ω_{δ_S}	0.100	0.098	-2.11	0.0184	0.0164	92.5
ω_{φ_S}	0.600	0.574	-4.39	0.1030	0.0903	87.5
α_1	0.650	0.649	-0.18	0.0131	0.0128	93.5
α_2	-0.490	-0.489	-0.23	0.0116	0.0126	96.5
α_3	1.060	1.060	-0.15	0.0150	0.0151	92.5
α_4	-0.720	-0.718	-0.21	0.0140	0.0145	96.0
α_5	0.330	0.329	-0.20	0.0157	0.0147	91.0
μ_1	-0.610	-0.610	-0.08	0.0228	0.0196	90.5
μ_2	0.040	0.041	2.74	0.0186	0.0188	97.5
μ_3	-1.800	-1.800	0.10	0.0253	0.0243	95.5
μ_4	2.400	2.400	-0.02	0.0235	0.0222	95.5
μ_5	1.090	1.090	-0.05	0.0212	0.0214	95.5
σ_{Ab}	0.110	0.110	-0.13	0.0081	0.0077	93.0
α_1	0.310	0.309	-0.42	0.0087	0.0096	96.5
α_2	0.310	0.309	-0.16	0.0093	0.0096	96.0
α_3	0.350	0.348	-0.65	0.0110	0.0109	94.5
α_4	0.350	0.351	0.16	0.0119	0.0109	94.5
α_5	0.370	0.368	-0.42	0.0114	0.0114	97.5

TABLE 5.9 – Parameter estimation accuracy for population and biomarker-specific parameters using REMixed algorithm with additionnal test on α_1 parameters over 200 replicates using samples of 25 subjects and 20 genes.

Gene	Parameter	Target Value	Mean Estimate	Empirical SD	Estimated SD	Selection Frequency (%)
G6	μ_6	-1.660	-1.660	0.0250	0.0148	7.0
	α_6	0.000	-0.003	0.0242	0.0104	
	α_6	0.260	0.261	0.0089	0.0081	
G7	μ_7	-0.160	-0.168	0.0267	0.0167	5.5
	α_7	0.000	0.004	0.0287	0.0116	
	α_7	0.300	0.293	0.0119	0.0090	
G8	μ_8	-0.210	-0.218	0.0318	0.0159	6.0
	α_8	0.000	0.005	0.0259	0.0111	
	α_8	0.280	0.279	0.0052	0.0086	
G9	μ_9	-0.960	-0.963	0.0275	0.0171	6.5
	α_9	0.000	-0.002	0.0312	0.0119	
	α_9	0.300	0.299	0.0075	0.0092	
G10	μ_{10}	-0.520	-0.522	0.0278	0.0206	3.5
	α_{10}	0.000	0.005	0.0348	0.0144	
	α_{10}	0.360	0.362	0.0156	0.0112	
G11	μ_{11}	0.620	0.620	0.0205	0.0141	3.5
	α_{11}	0.000	-0.003	0.0232	0.0097	
	α_{11}	0.250	0.247	0.0056	0.0076	
G12	μ_{12}	0.800	0.801	0.0391	0.0240	3.0
	α_{12}	0.000	0.014	0.0483	0.0168	
	α_{12}	0.420	0.422	0.0125	0.0130	
G13	μ_{13}	0.070	0.075	0.0324	0.0165	5.5
	α_{13}	0.000	-0.002	0.0281	0.0114	
	α_{13}	0.290	0.288	0.0082	0.0089	
G14	μ_{14}	0.130	0.133	0.0229	0.0130	10.5
	α_{14}	0.000	-0.003	0.0229	0.0091	
	α_{14}	0.230	0.228	0.0068	0.0070	
G15	μ_{15}	-0.180	-0.180	0.0301	0.0141	3.5
	α_{15}	0.000	-0.001	0.0243	0.0097	
	α_{15}	0.250	0.246	0.0068	0.0076	
G16	μ_{16}	-0.170	-0.185	0.0277	0.0174	7.5
	α_{16}	0.000	0.016	0.0268	0.0121	
	α_{16}	0.310	0.306	0.0083	0.0094	
G17	μ_{17}	-0.450	-0.457	0.0358	0.0161	4.5
	α_{17}	0.000	0.002	0.0270	0.0112	
	α_{17}	0.290	0.282	0.0082	0.0087	
G18	μ_{18}	0.500	0.493	0.0314	0.0156	5.0
	α_{18}	0.000	0.007	0.0262	0.0110	
	α_{18}	0.280	0.276	0.0069	0.0085	
G19	μ_{19}	-2.270	-2.269	0.0360	0.0172	5.0
	α_{19}	0.000	-0.001	0.0316	0.0119	
	α_{19}	0.310	0.301	0.0053	0.0093	
G20	μ_{20}	-2.020	-2.023	0.0236	0.0153	3.5
	α_{20}	0.000	0.010	0.0250	0.0107	
	α_{20}	0.270	0.269	0.0045	0.0083	

TABLE 5.10 – Parameter estimation accuracy for the irrelevant biomarker (Gaussian Random noise) using REMixed algorithm with additional Wald tests over 200 replicates using samples of 25 subjects and 20 genes.

Parameter	Target Value	Mean estimate	Relative Bias (%)	Empirical SD	Estimated SD	95% Coverage
δ_{Ab}	0.025	0.025	0.47	0.0017	0.0015	92.5
δ_S	0.230	0.230	0.07	0.0053	0.0049	92.0
$\varphi_S \times 10^{-2}$	6.110	6.170	0.93	0.0079	0.0075	94.0
$\omega_{\delta_{Ab}}$	0.300	0.290	-3.47	0.0387	0.0421	93.0
ω_{δ_S}	0.100	0.100	-0.32	0.0151	0.0145	89.5
ω_{φ_S}	0.600	0.580	-3.27	0.0981	0.0911	92.5
α_1	1.060	1.060	0.02	0.0099	0.0095	93.5
α_2	-0.720	-0.719	-0.08	0.0152	0.0151	95.5
α_3	0.330	0.331	0.25	0.0123	0.0123	94.5
α_4	-0.280	-0.279	-0.18	0.0104	0.0111	97.5
α_5	1.550	1.550	0.00	0.0115	0.0124	97.0
α_6	-0.240	-0.240	0.01	0.0106	0.0118	97.0
α_7	-1.050	-1.050	0.00	0.0159	0.0152	95.5
α_8	-1.790	-1.790	0.00	0.0109	0.0101	92.5
α_9	0.760	0.760	0.04	0.0117	0.0124	96.0
α_{10}	-0.350	-0.348	-0.44	0.0153	0.0151	94.0
μ_1	-0.170	-0.169	-0.35	0.0133	0.0143	96.5
μ_2	-0.450	-0.453	0.73	0.0235	0.0219	94.5
μ_3	0.500	0.501	0.14	0.0162	0.0176	96.0
μ_4	-2.270	-2.270	0.09	0.0164	0.0160	94.5
μ_5	-2.020	-2.020	-0.09	0.0194	0.0190	96.0
σ_{Ab}	0.110	0.110	0.08	0.0069	0.0077	96.0
σ_1	0.230	0.230	-0.08	0.0073	0.0071	93.5
σ_2	0.380	0.379	-0.21	0.0114	0.0117	95.5
σ_3	0.310	0.309	-0.38	0.0094	0.0095	94.5
σ_4	0.280	0.280	-0.08	0.0080	0.0086	97.5
σ_5	0.300	0.299	-0.35	0.0087	0.0093	96.0
σ_6	0.300	0.298	-0.68	0.0096	0.0092	93.0
σ_7	0.380	0.379	-0.33	0.0122	0.0117	95.5
σ_8	0.230	0.230	-0.04	0.0072	0.0072	95.0
σ_9	0.310	0.310	-0.14	0.0098	0.0096	95.5
σ_{10}	0.380	0.380	-0.13	0.0106	0.0117	98.0

TABLE 5.11 – Parameter estimation accuracy for population and biomarker-specific parameters using REMixed algorithm (without additional test) over 200 replicates using samples of 25 subjects and 50 genes.

Parameter estimation accuracy for population and biomarker-specific parameters for the 50-gene scenario

Parameter	Target Value	Mean estimate	Relative Bias (%)	Empirical SD	Estimated SD	95% Coverage
δ_{Ab}	0.025	0.025	0.46	0.0017	0.0015	92.0
δ_S	0.230	0.230	0.03	0.0053	0.0049	93.5
$\varphi_S \times 10^{-2}$	6.110	6.160	0.75	0.0079	0.0075	93.5
$\omega_{\delta_{Ab}}$	0.300	0.290	-3.46	0.0386	0.0421	93.0
ω_{δ_S}	0.100	0.100	-0.31	0.0152	0.0145	89.5
ω_{φ_S}	0.600	0.580	-3.38	0.0987	0.0908	91.0
α_1	1.060	1.060	0.01	0.0098	0.0095	94.5
α_2	-0.720	-0.719	-0.10	0.0152	0.0151	95.5
α_3	0.330	0.331	0.24	0.0123	0.0123	94.5
α_4	-0.280	-0.279	-0.19	0.0104	0.0111	97.5
α_5	1.550	1.550	-0.01	0.0115	0.0124	97.0
α_6	-0.240	-0.240	0.02	0.0106	0.0118	96.5
α_7	-1.050	-1.050	0.05	0.0149	0.0152	96.0
α_8	-1.790	-1.790	-0.01	0.0109	0.0101	93.5
α_9	0.760	0.760	0.02	0.0118	0.0124	96.0
α_{10}	-0.350	-0.348	-0.45	0.0153	0.0151	94.5
μ_1	-0.170	-0.170	-0.16	0.0134	0.0143	96.0
μ_2	-0.450	-0.453	0.69	0.0234	0.0219	95.5
μ_3	0.500	0.501	0.13	0.0163	0.0176	95.5
μ_4	-2.270	-2.270	0.09	0.0164	0.0160	94.0
μ_5	-2.020	-2.020	-0.07	0.0191	0.0190	94.5
σ_{Ab}	0.110	0.110	0.12	0.0069	0.0078	96.0
σ_1	0.230	0.230	-0.08	0.0073	0.0071	93.0
σ_2	0.380	0.379	-0.21	0.0115	0.0117	96.0
σ_3	0.310	0.309	-0.38	0.0094	0.0095	94.5
σ_4	0.280	0.280	-0.08	0.0080	0.0086	97.5
σ_5	0.300	0.299	-0.35	0.0087	0.0093	96.0
σ_6	0.300	0.298	-0.68	0.0096	0.0092	93.0
σ_7	0.380	0.379	-0.39	0.0116	0.0117	96.0
σ_8	0.230	0.230	-0.03	0.0071	0.0072	95.5
σ_9	0.310	0.310	-0.13	0.0098	0.0096	95.5
σ_{10}	0.380	0.380	-0.13	0.0106	0.0117	98.0

TABLE 5.12 – Parameter estimation accuracy for population and biomarker-specific parameters using REMixed algorithm with additional test over 200 replicates using samples of 25 subjects and 50 genes.

Parameter estimation accuracy for non-significant biomarker-specific parameters when selected

Gene	Parameter	Target Value	Mean estimate	Empirical SD	Estimated SD	Number of Selection
G11	μ_{11}	-0.870	-0.861	0.0316	0.0192	7
	α_{11}	0.000	-0.009	0.0283	0.0139	
	σ_{11}	0.330	0.328	0.0112	0.0101	
G12	μ_{12}	0.030	0.032	0.0236	0.0169	17
	α_{12}	0.000	-0.004	0.0263	0.0118	
	σ_{12}	0.300	0.297	0.0061	0.0092	
G13	μ_{13}	-0.400	-0.400	0.0337	0.0193	5
	α_{13}	0.000	0.008	0.0335	0.0136	
	σ_{13}	0.340	0.341	0.0144	0.0105	
G14	μ_{14}	0.030	-0.000	0.0194	0.0198	9
	α_{14}	0.000	0.020	0.0250	0.0139	
	σ_{14}	0.350	0.347	0.0120	0.0107	
G15	μ_{15}	1.120	1.110	0.0222	0.0148	29
	α_{15}	0.000	0.006	0.0193	0.0104	
	σ_{15}	0.260	0.260	0.0068	0.0080	
G16	μ_{16}	-0.990	-0.985	0.0151	0.0123	36
	α_{16}	0.000	-0.004	0.0146	0.0086	
	σ_{16}	0.220	0.217	0.0064	0.0067	
G17	μ_{17}	2.460	2.470	0.0265	0.0156	17
	α_{17}	0.000	-0.011	0.0219	0.0109	
	σ_{17}	0.280	0.275	0.0069	0.0085	
G18	μ_{18}	-0.200	-0.226	0.0288	0.0209	7
	α_{18}	0.000	0.028	0.0286	0.0146	
	σ_{18}	0.370	0.368	0.0126	0.0113	
G19	μ_{19}	-1.500	-1.510	0.0239	0.0158	16
	α_{19}	0.000	0.006	0.0209	0.0111	
	σ_{19}	0.280	0.280	0.0081	0.0086	
G20	μ_{20}	-1.050	-1.040	0.0268	0.0169	14
	α_{20}	0.000	-0.010	0.0244	0.0117	
	σ_{20}	0.300	0.296	0.0074	0.0092	
G21	μ_{21}	0.280	0.282	0.0214	0.0188	13
	α_{21}	0.000	0.005	0.0279	0.0134	
	σ_{21}	0.330	0.323	0.0088	0.0100	
G22	μ_{22}	-0.170	-0.167	0.0387	0.0184	11
	α_{22}	0.000	0.000	0.0323	0.0120	
	σ_{22}	0.340	0.331	0.0075	0.0102	
G23	μ_{23}	-0.320	-0.341	0.0114	0.0201	3
	α_{23}	0.000	0.038	0.0014	0.0141	

	σ_{23}	0.350	0.355	0.0110	0.0110	
G24	μ_{24}	0.220	0.220	0.0299	0.0180	
	α_{24}	0.000	0.005	0.0312	0.0126	12
	σ_{24}	0.320	0.317	0.0105	0.0098	
G25	μ_{25}	-0.670	-0.669	0.0176	0.0135	
	α_{25}	0.000	0.000	0.0190	0.0094	31
	σ_{25}	0.240	0.237	0.0064	0.0073	
G26	μ_{26}	-1.310	-1.310	0.0241	0.0160	
	α_{26}	0.000	0.001	0.0245	0.0111	20
	σ_{26}	0.280	0.281	0.0097	0.0087	
G27	μ_{27}	-0.520	-0.518	0.0227	0.0135	
	α_{27}	0.000	-0.003	0.0205	0.0094	34
	σ_{27}	0.240	0.238	0.0075	0.0073	
G28	μ_{28}	-0.800	-0.790	0.0308	0.0162	
	α_{28}	0.000	-0.006	0.0257	0.0114	17
	σ_{28}	0.290	0.286	0.0099	0.0088	
G29	μ_{29}	1.290	1.300	0.0353	0.0186	
	α_{29}	0.000	-0.006	0.0337	0.0129	8
	σ_{29}	0.330	0.326	0.0082	0.0101	
G30	μ_{30}	-0.110	-0.105	0.0287	0.0169	
	α_{30}	0.000	-0.003	0.0276	0.0118	12
	σ_{30}	0.300	0.297	0.0074	0.0092	
G31	μ_{31}	-0.930	-0.938	0.0173	0.0143	
	α_{31}	0.000	0.008	0.0186	0.0100	26
	σ_{31}	0.250	0.253	0.0077	0.0078	
G32	μ_{32}	0.290	0.288	0.0255	0.0163	
	α_{32}	0.000	0.005	0.0284	0.0114	17
	σ_{32}	0.290	0.287	0.0090	0.0089	
G33	μ_{33}	1.370	1.370	0.0252	0.0147	
	α_{33}	0.000	-0.003	0.0208	0.0103	15
	σ_{33}	0.260	0.259	0.0115	0.0080	
G34	μ_{34}	-0.430	-0.447	0.0432	0.0181	
	α_{34}	0.000	0.015	0.0279	0.0126	9
	σ_{34}	0.320	0.319	0.0081	0.0098	
G35	μ_{35}	-0.030	-0.032	0.0253	0.0176	
	α_{35}	0.000	0.009	0.0254	0.0123	14
	σ_{35}	0.310	0.309	0.0099	0.0095	
G36	μ_{36}	0.100	0.111	0.0207	0.0153	
	α_{36}	0.000	-0.007	0.0234	0.0107	24
	σ_{36}	0.270	0.268	0.0076	0.0083	
G37	μ_{37}	0.790	0.790	0.0156	0.0118	
	α_{37}	0.000	0.000	0.0146	0.0083	47

	σ_{37}	0.210	0.208	0.0073	0.0064	
G38	μ_{38}	-0.160	-0.161	0.0287	0.0169	
	α_{38}	0.000	-0.002	0.0281	0.0113	17
	σ_{38}	0.300	0.301	0.0071	0.0093	
G39	μ_{39}	1.300	1.310	0.0270	0.0164	
	α_{39}	0.000	-0.008	0.0245	0.0114	18
	σ_{39}	0.290	0.291	0.0095	0.0090	
G40	μ_{40}	-0.170	-0.200	0.0056	0.0193	
	α_{40}	0.000	0.029	0.0018	0.0133	3
	σ_{40}	0.340	0.338	0.0103	0.0104	
G41	μ_{41}	-0.670	-0.670	0.0179	0.0123	
	α_{41}	0.000	-0.002	0.0159	0.0086	38
	σ_{41}	0.220	0.216	0.0075	0.0067	
G42	μ_{42}	1.020	1.020	0.0199	0.0146	
	α_{42}	0.000	-0.004	0.0205	0.0102	25
	σ_{42}	0.260	0.257	0.0076	0.0079	
G43	μ_{43}	-2.030	-2.040	0.0237	0.0141	
	α_{43}	0.000	0.002	0.0210	0.0099	22
	σ_{43}	0.250	0.248	0.0084	0.0076	
G44	μ_{44}	-0.260	-0.258	0.0256	0.0140	
	α_{44}	0.000	0.001	0.0202	0.0098	20
	σ_{44}	0.250	0.247	0.0084	0.0076	
G45	μ_{45}	-0.220	-0.219	0.0129	0.0097	
	α_{45}	0.000	-0.000	0.0107	0.0067	67
	σ_{45}	0.170	0.170	0.0053	0.0052	
G46	μ_{46}	-1.410	-1.420	0.0200	0.0145	
	α_{46}	0.000	0.007	0.0229	0.0101	23
	σ_{46}	0.260	0.255	0.0054	0.0079	
G47	μ_{47}	-0.860	-0.865	0.0145	0.0113	
	α_{47}	0.000	0.003	0.0130	0.0078	49
	σ_{47}	0.200	0.199	0.0060	0.0062	
G48	μ_{48}	0.930	0.930	0.0221	0.0142	
	α_{48}	0.000	0.002	0.0187	0.0100	18
	σ_{48}	0.250	0.250	0.0095	0.0077	
G49	μ_{49}	-1.190	-1.190	0.0220	0.0142	
	α_{49}	0.000	0.000	0.0234	0.0099	19
	σ_{49}	0.250	0.249	0.0093	0.0077	
G50	μ_{50}	-1.270	-1.270	0.0284	0.0183	
	α_{50}	0.000	-0.002	0.0300	0.0133	17
	σ_{50}	0.310	0.310	0.0098	0.0096	

TABLE 5.13 – Parameter estimation accuracy for the irrelevant biomarker (Gaussian random noise) using REMixed algorithm (without additional test) over 200 replicates using samples of 25 subjects and 50 genes.

Gene	Parameter	Target Value	Mean estimate	Empirical SD	Estimated SD	Number of Selection
G11	μ_{11}	-0.870	-0.861	0.0347	0.0185	6
	α_{11}	0.000	-0.011	0.0309	0.0130	
	σ_{11}	0.330	0.327	0.0118	0.0101	
G12	μ_{12}	0.030	0.032	0.0230	0.0168	12
	α_{12}	0.000	-0.005	0.0287	0.0117	
	σ_{12}	0.300	0.295	0.0066	0.0091	
G13	μ_{13}	-0.400	-0.400	0.0338	0.0193	5
	α_{13}	0.000	0.008	0.0337	0.0136	
	σ_{13}	0.340	0.341	0.0144	0.0105	
G14	μ_{14}	0.030	-0.005	0.0239	0.0196	5
	α_{14}	0.000	0.037	0.0080	0.0136	
	σ_{14}	0.350	0.344	0.0103	0.0106	
G15	μ_{15}	1.120	1.110	0.0251	0.0149	11
	α_{15}	0.000	0.006	0.0240	0.0104	
	σ_{15}	0.260	0.262	0.0085	0.0081	
G16	μ_{16}	-0.990	-0.988	0.0176	0.0123	8
	α_{16}	0.000	-0.004	0.0214	0.0086	
	σ_{16}	0.220	0.216	0.0057	0.0067	
G17	μ_{17}	2.460	2.470	0.0281	0.0157	10
	α_{17}	0.000	-0.013	0.0241	0.0109	
	σ_{17}	0.280	0.275	0.0070	0.0085	
G18	μ_{18}	-0.200	-0.226	0.0289	0.0209	7
	α_{18}	0.000	0.028	0.0287	0.0146	
	σ_{18}	0.370	0.368	0.0126	0.0113	
G19	μ_{19}	-1.500	-1.500	0.0285	0.0160	7
	α_{19}	0.000	-0.003	0.0253	0.0112	
	σ_{19}	0.280	0.282	0.0067	0.0087	
G20	μ_{20}	-1.050	-1.030	0.0253	0.0170	10
	α_{20}	0.000	-0.017	0.0235	0.0119	
	σ_{20}	0.300	0.299	0.0064	0.0092	
G21	μ_{21}	0.280	0.281	0.0218	0.0184	12
	α_{21}	0.000	0.006	0.0293	0.0128	
	σ_{21}	0.330	0.323	0.0090	0.0100	
G22	μ_{22}	-0.170	-0.166	0.0408	0.0188	10
	α_{22}	0.000	0.001	0.0342	0.0131	

	σ_{22}	0.340	0.331	0.0079	0.0102	
G23	μ_{23}	-0.320	-0.342	0.0114	0.0201	
	α_{23}	0.000	0.038	0.0013	0.0141	3
	σ_{23}	0.350	0.355	0.0110	0.0110	
G24	μ_{24}	0.220	0.216	0.0288	0.0181	
	α_{24}	0.000	0.008	0.0316	0.0126	11
	σ_{24}	0.320	0.318	0.0100	0.0098	
G25	μ_{25}	-0.670	-0.667	0.0202	0.0133	
	α_{25}	0.000	-0.002	0.0228	0.0093	13
	σ_{25}	0.240	0.233	0.0052	0.0072	
G26	μ_{26}	-1.310	-1.310	0.0259	0.0159	
	α_{26}	0.000	0.003	0.0254	0.0111	15
	σ_{26}	0.280	0.279	0.0077	0.0086	
G27	μ_{27}	-0.520	-0.522	0.0265	0.0136	
	α_{27}	0.000	-0.001	0.0245	0.0095	17
	σ_{27}	0.240	0.239	0.0091	0.0074	
G28	μ_{28}	-0.800	-0.786	0.0341	0.0163	
	α_{28}	0.000	-0.011	0.0259	0.0114	13
	σ_{28}	0.290	0.288	0.0099	0.0089	
G29	μ_{29}	1.290	1.300	0.0355	0.0186	
	α_{29}	0.000	-0.006	0.0339	0.0129	8
	σ_{29}	0.330	0.326	0.0082	0.0101	
G30	μ_{30}	-0.110	-0.112	0.0296	0.0169	
	α_{30}	0.000	0.003	0.0295	0.0117	9
	σ_{30}	0.300	0.297	0.0084	0.0092	
G31	μ_{31}	-0.930	-0.937	0.0196	0.0143	
	α_{31}	0.000	0.007	0.0247	0.0100	8
	σ_{31}	0.250	0.251	0.0081	0.0077	
G32	μ_{32}	0.290	0.285	0.0264	0.0164	
	α_{32}	0.000	0.009	0.0286	0.0114	15
	σ_{32}	0.290	0.288	0.0091	0.0089	
G33	μ_{33}	1.370	1.370	0.0249	0.0146	
	α_{33}	0.000	0.001	0.0233	0.0102	8
	σ_{33}	0.260	0.256	0.0104	0.0079	
G34	μ_{34}	-0.430	-0.447	0.0433	0.0181	
	α_{34}	0.000	0.015	0.0281	0.0126	9
	σ_{34}	0.320	0.319	0.0081	0.0098	
G35	μ_{35}	-0.030	-0.040	0.0227	0.0177	
	α_{35}	0.000	0.017	0.0229	0.0123	10
	σ_{35}	0.310	0.311	0.0107	0.0096	
G36	μ_{36}	0.100	0.107	0.0257	0.0154	
	α_{36}	0.000	-0.001	0.0292	0.0108	12

	σ_{36}	0.270	0.271	0.0074	0.0084	
G37	μ_{37}	0.790	0.802	0.0177	0.0118	
	α_{37}	0.000	-0.011	0.0165	0.0082	9
	σ_{37}	0.210	0.206	0.0062	0.0064	
G38	μ_{38}	-0.160	-0.159	0.0312	0.0171	
	α_{38}	0.000	-0.004	0.0304	0.0120	14
	σ_{38}	0.300	0.301	0.0063	0.0093	
G39	μ_{39}	1.300	1.310	0.0301	0.0167	
	α_{39}	0.000	-0.012	0.0273	0.0117	10
	σ_{39}	0.290	0.294	0.0092	0.0091	
G40	μ_{40}	-0.170	-0.201	0.0054	0.0193	
	α_{40}	0.000	0.029	0.0015	0.0133	3
	σ_{40}	0.340	0.338	0.0103	0.0104	
G41	μ_{41}	-0.670	-0.672	0.0276	0.0125	
	α_{41}	0.000	0.000	0.0224	0.0087	10
	σ_{41}	0.220	0.219	0.0086	0.0068	
G42	μ_{42}	1.020	1.030	0.0255	0.0143	
	α_{42}	0.000	-0.013	0.0252	0.0101	7
	σ_{42}	0.260	0.253	0.0087	0.0078	
G43	μ_{43}	-2.030	-2.040	0.0263	0.0141	
	α_{43}	0.000	-0.000	0.0240	0.0098	12
	σ_{43}	0.250	0.248	0.0078	0.0077	
G44	μ_{44}	-0.260	-0.255	0.0334	0.0142	
	α_{44}	0.000	-0.000	0.0257	0.0099	8
	σ_{44}	0.250	0.250	0.0083	0.0077	
G45	μ_{45}	-0.220	-0.218	0.0170	0.0097	
	α_{45}	0.000	-0.000	0.0158	0.0067	12
	σ_{45}	0.170	0.170	0.0034	0.0052	
G46	μ_{46}	-1.410	-1.420	0.0196	0.0146	
	α_{46}	0.000	0.009	0.0253	0.0102	15
	σ_{46}	0.260	0.256	0.0051	0.0079	
G47	μ_{47}	-0.860	-0.869	0.0158	0.0114	
	α_{47}	0.000	0.010	0.0156	0.0079	10
	σ_{47}	0.200	0.199	0.0044	0.0061	
G48	μ_{48}	0.930	0.934	0.0229	0.0142	
	α_{48}	0.000	0.000	0.0229	0.0100	8
	σ_{48}	0.250	0.251	0.0054	0.0077	
G49	μ_{49}	-1.190	-1.190	0.0226	0.0142	
	α_{49}	0.000	0.000	0.0258	0.0099	13
	σ_{49}	0.250	0.250	0.0073	0.0077	
G50	μ_{50}	-1.270	-1.270	0.0310	0.0176	
	α_{50}	0.000	-0.002	0.0335	0.0123	14

σ_{50}	0.310	0.309	0.0091	0.0095
---------------	-------	-------	--------	--------

TABLE 5.14 – Parameter estimation accuracy for the irrelevant biomarker (Gaussian random noise) using REMixed algorithm with additional test over 200 replicates using samples of 25 subjects and 50 genes.

B.1.6 Application to COVID-19 mRNA vaccination data

For the real-data application, we ran the REMixed algorithm over a grid of 20 penalty parameters ranging from 2.5 to 250, with an estimated $\lambda_{\max}$ of 383. Convergence criteria were set to 0.01 for parameter stability and 0.5 for penalized log-likelihood stability. In addition, we used 10 points per dimension in the adaptive Gauss–Hermite quadrature.

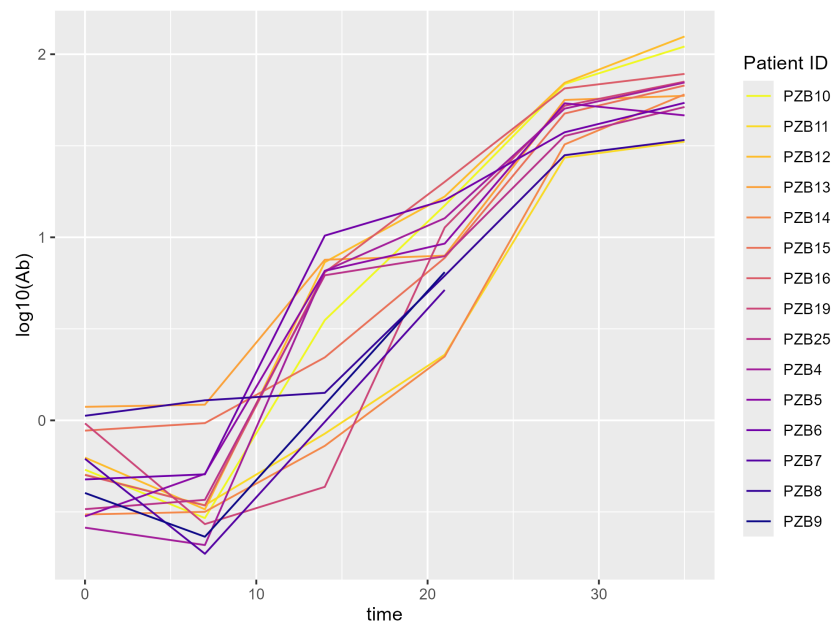
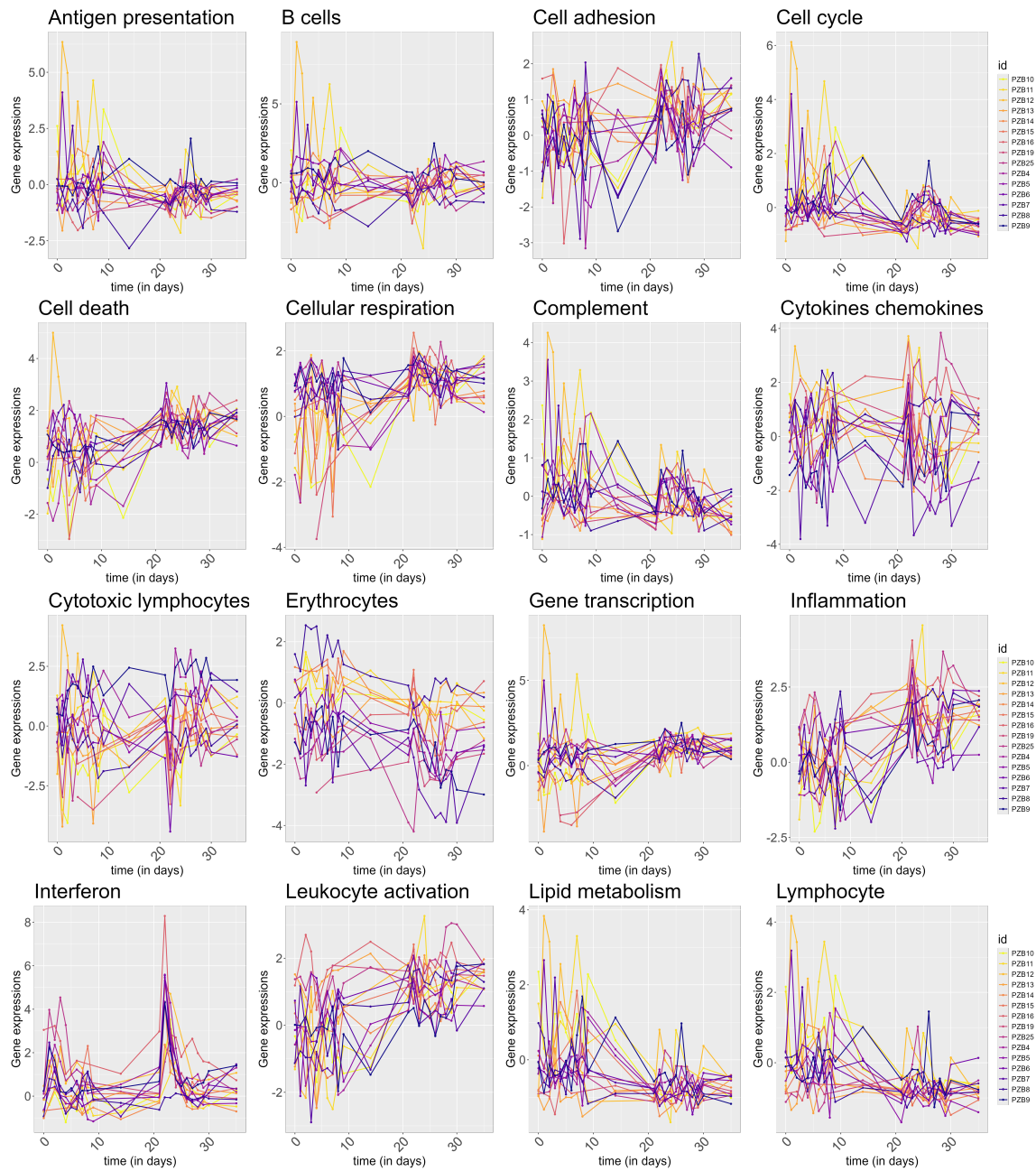


FIGURE 5.19 – Antibody titre trajectories for the 15 individuals receiving the COVID-19 mRNA vaccine (BNT162b2).



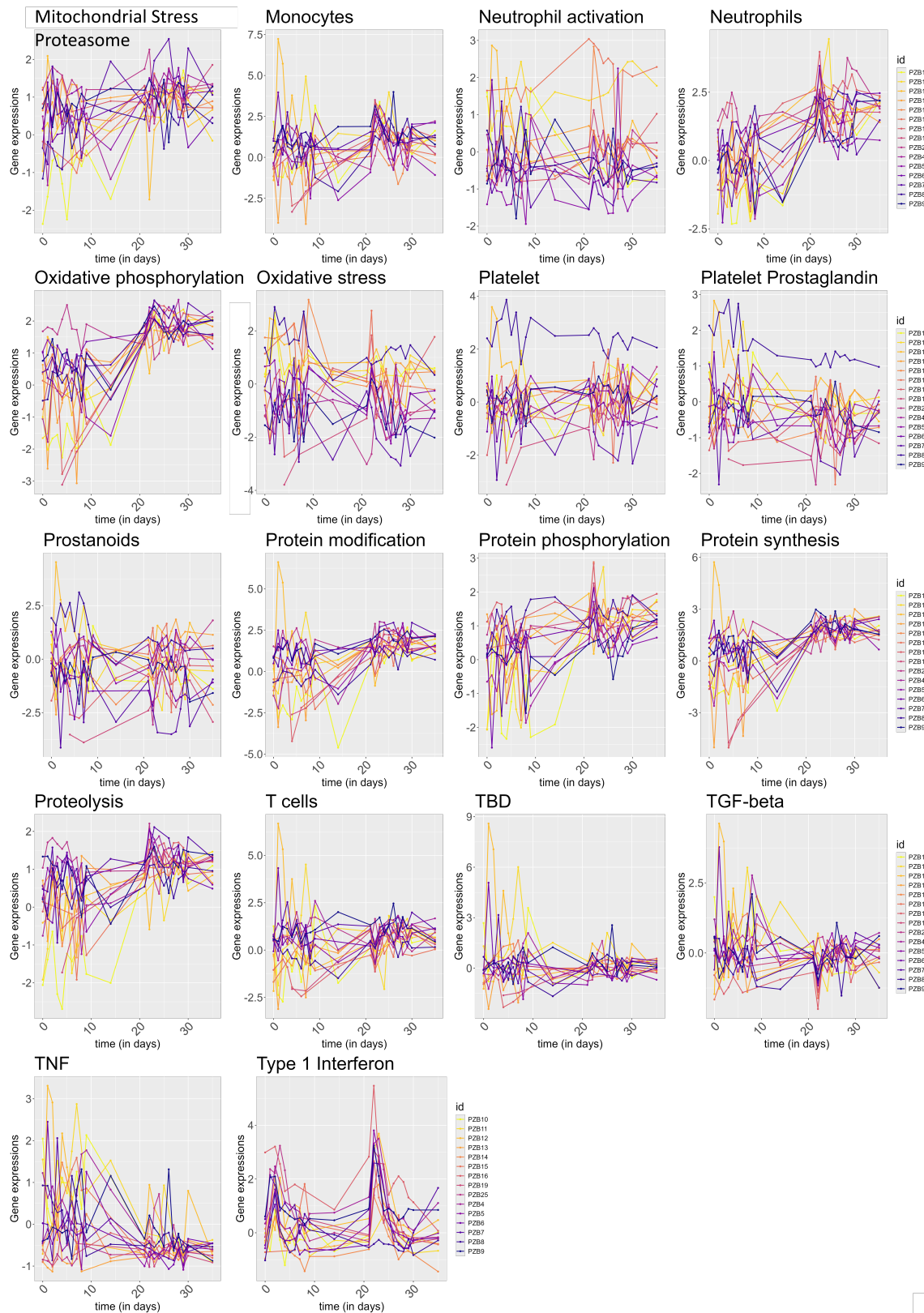


FIGURE 5.20 – Mean pathway representatives computed for each individual using the gene expression profile from [Rinchai *et al.* \(2022\)](#) and Classification by [Durinck *et al.* \(2005\)](#) and [Li *et al.* \(2014\)](#) to gather genes by pathways. In total 34 pathways are considered.

Parameter	REMixed estimate	SAEM estimate (SE)	SAEM after Wald tests (SE)
$\rho_0^{(2)}$	35.315	31.279 (9.119)	17.647 (6.533)
θ_0	0.889	0.894 (0.128)	0.913 (0.131)
ω_ρ	0.345	0.343 (0.081)	0.343 (0.083)
ω_θ	0.432	0.424 (0.135)	0.4 (0.135)
σ_{Ab}	0.252	0.252 (0.021)	0.253 (0.022)

TABLE 5.15 – NLME parameters estimates obtained (1) at the end of the REMixed algorithm with the selected penalty parameter $\lambda^* = 25$, (2) from a final SAEM algorithm used to compute standard errors and Wald tests (with standard errors), and (3) from a final SAEM algorithm after Wald tests (with standard errors).

Pathway	Parameter	REMixed estimate	SAEM estimate (SE)	SAEM after Wald tests (SE)
Cell cycle	μ_1	0.054	0.056 (0.064)	-0.003 (0.058)
	α_1	-0.013	-0.015 (0.008)	0 (ND)
	σ_1	0.964	0.964 (0.041)	0.973 (0.041)
TBD	μ_2	0.173	0.168 (0.075)	0.183 (0.068)
	α_2	0.002	0.004 (0.008)	0 (ND)
	σ_2	1.140	1.139 (0.048)	1.140 (0.048)
Prostanoids	μ_3	-0.191	-0.184 (0.085)	-0.237 (0.077)
	α_3	-0.010	-0.013 (0.010)	0 (ND)
	σ_3	1.281	1.281 (0.054)	1.285 (0.054)
Type 1	μ_4	0.159	0.159 (0.057)	0.146 (0.057)
Interferon	α_4	0.076	0.085 (0.024)	0.153 (0.057)
	σ_4	0.860	0.860 (0.037)	0.851 (0.037)
Cytotoxic lymphocytes	μ_5	0.101	0.108 (0.096)	0.019 (0.087)
	α_5	-0.018	-0.023 (0.012)	0 (ND)
	σ_5	1.448	1.448 (0.061)	1.460 (0.062)
Erythrocytes	μ_6	-0.597	-0.592 (0.081)	-0.648 (0.073)
	α_6	-0.011	-0.014 (0.010)	0 (ND)
	σ_6	1.224	1.224 (0.052)	1.230 (0.052)
Interferon	μ_7	0.216	0.216 (0.062)	0.200 (0.063)
	α_7	0.105	0.118 (0.033)	0.211 (0.077)
	σ_7	0.933	0.933 (0.041)	0.922 (0.040)
Protein synthesis	μ_8	0.800	0.794 (0.092)	0.792 (0.093)
	α_8	0.042	0.049 (0.016)	0.085 (0.034)
	σ_8	1.396	1.396 (0.059)	1.398 (0.059)
Platelet	μ_9	0.179	0.182 (0.073)	0.168 (0.066)
	α_9	-0.003	-0.004 (0.008)	0 (ND)
	σ_9	1.106	1.106 (0.047)	1.107 (0.047)
Neutrophil activation	μ_{10}	-0.027	-0.028 (0.065)	0.000 (0.059)
	α_{10}	0.006	0.007 (0.008)	0 (ND)
	σ_{10}	0.983	0.982 (0.041)	0.984 (0.042)

Protein modification	μ_{11}	0.857	0.850 (0.091)	0.849 (0.091)
	α_{11}	0.022	0.027 (0.012)	0.046 (0.024)
	σ_{11}	1.372	1.372 (0.058)	1.373 (0.058)
Monocytes	μ_{12}	0.504	0.500 (0.085)	0.493 (0.085)
	α_{12}	0.058	0.067 (0.020)	0.118 (0.045)
	σ_{12}	1.284	1.284 (0.054)	1.283 (0.054)
Gene transcription	μ_{13}	0.635	0.629 (0.084)	0.684 (0.076)
	α_{13}	0.011	0.014 (0.010)	0 (ND)
	σ_{13}	1.274	1.274 (0.054)	1.279 (0.054)
T cells	μ_{14}	0.544	0.549 (0.075)	0.526 (0.067)
	α_{14}	-0.004	-0.006 (0.008)	0 (ND)
	σ_{14}	1.130	1.130 (0.048)	1.131 (0.048)
B cells	μ_{15}	0.139	0.144 (0.090)	0.145 (0.091)
	α_{15}	-0.026	-0.030 (0.013)	-0.053 (0.025)
	σ_{15}	1.365	1.365 (0.058)	1.366 (0.058)
Protein phosphorylation	μ_{16}	0.489	0.488 (0.058)	0.488 (0.058)
	α_{16}	0.030	0.034 (0.011)	0.059 (0.023)
	σ_{16}	0.870	0.870 (0.037)	0.872 (0.037)
Inflammation	μ_{17}	0.562	0.560 (0.075)	0.559 (0.076)
	α_{17}	0.050	0.057 (0.018)	0.099 (0.037)
	σ_{17}	1.139	1.139 (0.048)	1.142 (0.048)
Oxidative stress	μ_{18}	-0.227	-0.234 (0.084)	-0.210 (0.076)
	α_{18}	0.004	0.006 (0.009)	0 (ND)
	σ_{18}	1.273	1.273 (0.054)	1.274 (0.054)
Oxidative phosphorylation	μ_{19}	0.762	0.76 (0.075)	0.758 (0.075)
	α_{19}	0.046	0.053 (0.016)	0.092 (0.035)
	σ_{19}	1.130	1.130 (0.048)	1.133 (0.048)
Cell death	μ_{20}	0.794	0.792 (0.068)	0.79 (0.068)
	α_{20}	0.033	0.038 (0.013)	0.066 (0.026)
	σ_{20}	1.024	1.024 (0.043)	1.025 (0.043)
Proteolysis	μ_{21}	0.561	0.56 (0.054)	0.559 (0.055)
	α_{21}	0.029	0.033 (0.011)	0.058 (0.023)
	σ_{21}	0.823	0.823 (0.035)	0.824 (0.035)
Cytokines chemokines	μ_{22}	0.014	0.001 (0.092)	0.086 (0.083)
	α_{22}	0.016	0.021 (0.011)	0 (ND)
	σ_{22}	1.386	1.386 (0.058)	1.396 (0.059)
Neutrophils	μ_{23}	0.660	0.657 (0.080)	0.657 (0.080)
	α_{23}	0.056	0.063 (0.019)	0.110 (0.040)
	σ_{23}	1.202	1.202 (0.051)	1.207 (0.051)
Cellular respiration	μ_{24}	0.547	0.546 (0.062)	0.544 (0.063)
	α_{24}	0.034	0.038 (0.012)	0.067 (0.026)

	σ_{24}	0.941	0.941 (0.040)	0.942 (0.040)
Leukocyte	μ_{25}	0.454	0.451 (0.073)	0.452 (0.073)
activation	α_{25}	0.029	0.034 (0.012)	0.058 (0.024)
	σ_{25}	1.098	1.098 (0.046)	1.101 (0.046)
Antigen	μ_{26}	-0.140	-0.137 (0.067)	-0.137 (0.068)
presentation	α_{26}	-0.017	-0.020 (0.009)	-0.035 (0.017)
	σ_{26}	1.021	1.021 (0.043)	1.022 (0.043)
Cell	μ_{27}	0.131	0.129 (0.058)	0.128 (0.059)
adhesion	α_{27}	0.029	0.033 (0.011)	0.059 (0.023)
	σ_{27}	0.881	0.881 (0.037)	0.882 (0.037)
Mitochondrial	μ_{28}	0.631	0.629 (0.051)	0.629 (0.052)
Stress/	α_{28}	0.013	0.015 (0.007)	0.026 (0.013)
Proteasome	σ_{28}	0.775	0.775 (0.033)	0.775 (0.033)
	μ_{29}	0.070	0.067 (0.052)	0.088 (0.047)
Complement	α_{29}	0.004	0.005 (0.006)	0 (ND)
	σ_{29}	0.786	0.786 (0.033)	0.787 (0.033)
	μ_{30}	-0.118	-0.116 (0.052)	-0.118 (0.052)
TNF	α_{30}	-0.013	-0.015 (0.007)	-0.025 (0.013)
	σ_{30}	0.788	0.788 (0.033)	0.789 (0.033)
	μ_{31}	-0.280	-0.278 (0.056)	-0.279 (0.057)
Lymphocyte	α_{31}	-0.013	-0.015 (0.007)	-0.025 (0.013)
	σ_{31}	0.852	0.852 (0.036)	0.853 (0.036)
Platelet/	μ_{32}	-0.109	-0.106 (0.058)	-0.157 (0.053)
Prostaglandin	α_{32}	-0.011	-0.013 (0.007)	0 (ND)
	σ_{32}	0.880	0.880 (0.037)	0.886 (0.037)
	μ_{33}	0.082	0.082 (0.054)	0.085 (0.054)
TGF-beta	α_{33}	-0.02	-0.023 (0.009)	-0.041 (0.017)
	σ_{33}	0.813	0.813 (0.034)	0.813 (0.034)
Lipid	μ_{34}	-0.278	-0.276 (0.055)	-0.277 (0.056)
metabolism	α_{34}	-0.019	-0.022 (0.008)	-0.038 (0.016)
	σ_{34}	0.836	0.836 (0.035)	0.837 (0.035)

TABLE 5.16 – Biomarkers parameters estimates obtained (1) at the end of the REMixed algorithm with the selected penalty parameter $\lambda^* = 25$, (2) from a final SAEM algorithm used to compute standard errors and Wald tests (with standard errors), and (3) from a final SAEM algorithm after Wald tests (with standard errors). ND : Non Defined.

B.2 Matériels Supplémentaires associés à la Section 4.4

B.2.1 Regularization step details

We denote the Hessian $\frac{\partial^2}{\partial \alpha^2} LL(\alpha, \hat{\theta})$ in $\alpha = \hat{\alpha}$ as $\frac{\partial^2}{\partial \alpha^2} LL(\alpha, \hat{\theta}) \Big|_{\hat{\alpha}} = -\hat{X} = -(\hat{x}_{lc})_{l,c \leq K}$. We suppose here that the opposite of the Hessian is definite positive. Then :

$$\begin{aligned} LL(\alpha, \hat{\theta}) &\approx LL(\hat{\alpha}, \hat{\theta}) + (\alpha - \hat{\alpha})^T \frac{\partial}{\partial \alpha} LL(\alpha, \hat{\theta}) \Big|_{\alpha=\hat{\alpha}} - \frac{1}{2} (\alpha - \hat{\alpha})^T \hat{X} (\alpha - \hat{\alpha}) \\ &= LL(\hat{\alpha}, \hat{\theta}) + \alpha^T \partial_{\alpha} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \hat{\alpha}^T \partial_{\alpha} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} \\ &\quad - \frac{1}{2} \left(\alpha^T \hat{X} \alpha + \hat{\alpha}^T \hat{X} \hat{\alpha} - \alpha^T \hat{X} \hat{\alpha} - \hat{\alpha}^T \hat{X} \alpha \right). \end{aligned}$$

We'll focus on the term depending on α , as we wish to derivate with respect to α_k for $k = 1, \dots, K$, and develop the matrix products, and isolate the term depending on a specific α_k :

$$\begin{aligned} \alpha^T \partial_{\alpha} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} &= \sum_{l=1}^K \alpha_l \partial_{\alpha_l} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} \\ &= \alpha_k \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} + \sum_{\substack{l=1 \\ \neq k}}^K \alpha_l \partial_{\alpha_l} LL(\alpha, \hat{\theta})|_{\hat{\alpha}}, \end{aligned}$$

with

$$\begin{aligned} \alpha^T \hat{X} \alpha &= \sum_{c=1}^K \sum_{l=1}^K \alpha_l \hat{x}_{lc} \alpha_c = \sum_{l,c \neq k} \alpha_l \hat{x}_{lc} \alpha_c + 2\alpha_k \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \alpha_k^2 \hat{x}_{kk}, \\ \alpha^T \hat{X} \hat{\alpha} &= \sum_{c=1}^K \hat{\alpha}_c \left(\sum_{\substack{l=1 \\ \neq k}}^K \alpha_l \hat{x}_{lc} + \alpha_k \hat{x}_{kc} \right) = \sum_{c=1}^K \sum_{\substack{l=1 \\ \neq k}}^K \alpha_l \hat{x}_{lc} \hat{\alpha}_c + \alpha_k \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c, \\ \hat{\alpha}^T \hat{X} \alpha &= \sum_{c=1}^K \sum_{l=1}^K \hat{\alpha}_l \hat{x}_{lc} \alpha_c = \sum_{\substack{c=1 \\ \neq k}}^K \sum_{l=1}^K \hat{\alpha}_l \hat{x}_{lc} \alpha_c + \alpha_k \sum_{l=1}^K \hat{\alpha}_l \hat{x}_{lk}. \end{aligned}$$

Hence

$$\begin{aligned}
LL(\alpha, \hat{\theta}) &\approx LL(\hat{\alpha}, \hat{\theta}) + \alpha_k \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} + \sum_{\substack{l=1 \\ \neq k}}^K \alpha_l \partial_{\alpha_l} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \hat{\alpha}^T \partial_{\alpha} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} \\
&- \frac{1}{2} \left(\hat{\alpha}^T \hat{X} \hat{\alpha} + \sum_{l, c \neq k} \alpha_l \hat{x}_{lc} \alpha_c + 2\alpha_k \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \alpha_k^2 \hat{x}_{kk} \right. \\
&\quad - \sum_{c=1}^K \sum_{\substack{l=1 \\ \neq k}}^K \alpha_l \hat{x}_{lc} \hat{\alpha}_c - \alpha_k \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c \\
&\quad \left. - \sum_{c=1}^K \sum_{\substack{l=1 \\ \neq k}}^K \hat{\alpha}_l \hat{x}_{lc} \alpha_c - \alpha_{k'} \sum_{l=1}^K \hat{\alpha}_l \hat{x}_{lk} \right).
\end{aligned}$$

We derivate the last formula with respect to α_k

$$\begin{aligned}
\partial_{\alpha_k} LL(\alpha, \hat{\theta}) &= \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \frac{1}{2} \left(2 \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + 2\alpha_k \hat{x}_{kk} - \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \sum_{k=1}^K \hat{\alpha}_l \underbrace{\hat{x}_{lk}}_{=\hat{x}_{kl}} \right) \\
&= \partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c.
\end{aligned}$$

We want to find a parameter α_k so that $LL(\alpha, \hat{\theta}) - \lambda \|\alpha\|_1$ reach a maximum for this value. We then want to solve in α_k the equation

$$\begin{aligned}
&\partial_{\alpha_k} LL(\alpha, \hat{\theta}) - \lambda \text{sign}(\alpha_k) = 0 \\
&\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda \text{sign}(\alpha_k) = 0.
\end{aligned}$$

If $\alpha_k > 0$:

$$\begin{aligned}
&\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda = 0 \\
&\alpha_k = \frac{\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c - \lambda}{x_{kk}}.
\end{aligned}$$

So the condition $\alpha_k > 0$ implies that $\partial_{\alpha_k} LL(\alpha, \hat{\theta})|_{\hat{\alpha}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c > \lambda$.

If $\alpha_k < 0$:

$$\begin{aligned} \partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l - \alpha_k \hat{x}_{kk} + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c + \lambda &= 0 \\ \alpha_k &= \frac{\partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c + \lambda}{x_{kk}}. \end{aligned}$$

So the condition $\alpha_k < 0$ implies that $\partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c < -\lambda$.

The condition for $\left| \partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \alpha_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c \right| < \lambda$, implies then that α_k is set to 0.

We then denote $\left(\partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} - \sum_{\substack{l=1 \\ \neq k}}^K \hat{x}_{lk} \hat{\alpha}_l + \sum_{c=1}^K \hat{x}_{kc} \hat{\alpha}_c \right) = \left(\partial_{\alpha_k} LL(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}})|_{\hat{\boldsymbol{\alpha}}} + \hat{x}_{kk} \hat{\alpha}_k \right) = A$ an approximation using $\hat{\boldsymbol{\alpha}}$ of the previously mentioned condition.

Solving for each coordinate α_k leads to the soft-thresholding rule :

$$\alpha_k = \begin{cases} \frac{A + \lambda}{\hat{x}_{kk}} & \text{if } A < -\lambda, \\ \frac{A - \lambda}{\hat{x}_{kk}} & \text{if } A > \lambda, \\ 0 & \text{otherwise.} \end{cases} \quad \text{with } A = \partial_{\alpha_k} LL(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) + \hat{x}_{kk} \hat{\alpha}_k \quad (5.5)$$

B.2.2 Structure of the project

```

/.
├── data.txt
├── model.txt
├── project.mlxproperties
└── project.mlxtran

```

FIGURE 5.21 – Structure of a Monolix project directory after creation, prior to running any estimation or simulation tasks, showing the core files required for model specification and configuration.

```

/.
├── data.txt
├── model.txt
├── project.mlxproperties
├── project.mlxtran
└── project/
    ├── ...
    └── summary.txt

```

FIGURE 5.22 – Structure of a Monolix project directory after task execution, illustrating the creation of a results subfolder containing output files such as `summary.txt` with estimation and diagnostic results.

```

/.
├── ...
├── project_upd.mlxtran
├── project_upd.mlxproperties
├── project_upd/...
├── project/
│   ├── ...
│   └── remix/
│       ├── Build_1/...
│       ├── ...
│       ├── project_init/...
│       ├── tmp_init/...
│       ├── Build_1.mlxtran
│       ├── Build_1.mlxproperties
│       ├── ...
│       ├── project_init.mlxtran
│       ├── project_init.mlxproperties
│       ├── summary.txt
│       ├── summary_1.txt
│       ├── ...
│       ├── summary_cv.txt
│       └── summary_init.txt

```

FIGURE 5.23 – Structure of a Monolix project directory after sequential execution of `initStrat`, `remix`, and `cv.remix`, with both `unlinkTemporaryProject` and `unlinkBuildProject` set to `FALSE`. The `remix` folder contains all temporary projects and summary files generated during initialization, cross-validation, and final estimation steps.

B.2.3 Mlxtran model file for demo monolix project

```
[LONGITUDINAL]
input = {delta_S,phi_S,delta_AB,alpha_11,alpha_12,alpha_13,alpha_14,alpha_15}

EQUATION:
t_0 = 0
AB_0 = 1000
S_0 = 5

ddt_S = - delta_S*S
ddt_AB = phi_S * S - delta_AB * AB
tAB = log10(AB)

G1= alpha_11 * S
G2= alpha_12 * S
G3= alpha_13 * S
G4= alpha_14 * S
G5= alpha_15 * S

OUTPUT:
output = {tAB,G1,G2,G3,G4,G5}
```

B.2.4 Summary file for initialization task

```
-----
Starting Initilization by group

Group formed :
(yG1,yG4) - (yG2,yG3) - (yG5)
-----

Starting with biomarkers : yG4, yG1

Computing SAEM update for population parameters...
Estimating log-likelihood...

> Estimated Population Parameters :

EstimatedValue
delta_S_pop      2.6e-01
phi_S_pop        6.3e+02
delta_AB_pop     2.5e-02
omega_delta_AB   2.9e-01
omega_delta_S    9.6e-02
omega_phi_S      6.3e-01
sigma_AB         1.1e-01
sigma_G1         2.8e-01
```

sigma_G2	3.4e-01
sigma_G3	3.9e-01
sigma_G4	3.6e-01
sigma_G5	3.2e-01

	EstimatedValue
alpha_11_pop	1.1e-01
alpha_12_pop	0e+00
alpha_13_pop	0e+00
alpha_14_pop	-1.1e-05
alpha_15_pop	0e+00

> Estimated logLikelihood :

LL	OFV	AIC	BIC	BICc
-3252.417	6504.830	6526.830	6555.490	6593.170

Starting with biomarkers : yG2, yG3

Computing SAEM update for population parameters...
Estimating log-likelihood...

> Estimated Population Parameters :

	EstimatedValue
delta_S_pop	3.4e-01
phi_S_pop	7.9e+02
delta_AB_pop	2.5e-02
omega_delta_AB	3e-01
omega_delta_S	2.9e-01
omega_phi_S	5.7e-01
sigma_AB	1.1e-01
sigma_G1	3e-01
sigma_G2	3.4e-01
sigma_G3	3.9e-01
sigma_G4	3.6e-01
sigma_G5	3.2e-01

	EstimatedValue
alpha_11_pop	0e+00
alpha_12_pop	7.8e-03
alpha_13_pop	-1.7e-05
alpha_14_pop	0e+00
alpha_15_pop	0e+00

> Estimated logLikelihood :

LL	OFV	AIC	BIC	BICc
-3502.774	7005.550	7027.550	7056.210	7093.880

Starting with biomarkers : yG5

Computing SAEM update for population parameters...

Estimating log-likelihood...

> Estimated Population Parameters :

	EstimatedValue
delta_S_pop	4e-01
phi_S_pop	9.4e+02
delta_AB_pop	2.5e-02
omega_delta_AB	3e-01
omega_delta_S	3.3e-01
omega_phi_S	5.7e-01
sigma_AB	1.2e-01
sigma_G1	3e-01
sigma_G2	3.4e-01
sigma_G3	3.9e-01
sigma_G4	3.6e-01
sigma_G5	3.2e-01

	EstimatedValue
alpha_11_pop	0e+00
alpha_12_pop	0e+00
alpha_13_pop	0e+00
alpha_14_pop	0e+00
alpha_15_pop	-4.2e-03

> Estimated logLikelihood :

LL	OFV	AIC	BIC	BICc
-3506.928	7013.860	7031.860	7055.300	7083.560

FINAL SET SELECTED : yG4, yG1.

B.2.5 Summary file for grid search

Starting Regulatization and Estimation Algorithm

Estimation of the R.E. distribution using the initial model ...

- - - < INITIAL PARAMETERS > - - -

	EstimatedValue
delta_S_pop	2.63e-01
phi_S_pop	6.28e+02

delta_AB_pop	2.48e-02
omega_delta_AB	2.94e-01
omega_delta_S	9.57e-02
omega_phi_S	6.29e-01
sigma_AB	1.14e-01
sigma_G1	2.76e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

	EstimatedValue
alpha_11_pop	1.07e-01
alpha_12_pop	0e+00
alpha_13_pop	0e+00
alpha_14_pop	-1.09e-05
alpha_15_pop	0e+00

Computing regularization path ...

Starting algorithm n°1/10 with lambda = 5.096...

- - - < CRITERION > - - -

LL : -3250.233

BIC : 6574.149

BICc : 6635

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	1.04e-01
alpha_12_pop	8.52e-03
alpha_13_pop	-7.22e-04
alpha_14_pop	0.00e+00
alpha_15_pop	-1.98e-03

	EstimatedValue
delta_S_pop	2.51e-01
phi_S_pop	6.02e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.96e-01
omega_delta_S	1.69e-01
omega_phi_S	6.14e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°2/10 with lambda = 10.169...

- - - < CRITERION > - - -

LL : -3250.009

BIC : 6573.7

BICc : 6635

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	1.04e-01
alpha_12_pop	8.37e-03
alpha_13_pop	-5.23e-04
alpha_14_pop	0.00e+00
alpha_15_pop	-1.85e-03
	EstimatedValue
delta_S_pop	2.49e-01
phi_S_pop	5.98e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.97e-01
omega_delta_S	1.47e-01
omega_phi_S	6.15e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°3/10 with lambda = 20.289...

- - - < CRITERION > - - -

LL : -3250.038

BIC : 6573.759

BICc : 6635

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	1.03e-01
alpha_12_pop	8.04e-03
alpha_13_pop	-2.14e-06
alpha_14_pop	0.00e+00
alpha_15_pop	-1.57e-03
	EstimatedValue
delta_S_pop	2.44e-01

phi_S_pop	5.87e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.98e-01
omega_delta_S	1.64e-01
omega_phi_S	6.06e-01
sigma_AB	1.15e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°4/10 with lambda = 40.483...

- - - < CRITERION > - - -

LL : -3249.804

BIC : 6568.685

BICc : 6625

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	1.02e-01
alpha_12_pop	7.52e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	-1.02e-03
	EstimatedValue
delta_S_pop	2.39e-01
phi_S_pop	5.74e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.97e-01
omega_delta_S	1.43e-01
omega_phi_S	6.13e-01
sigma_AB	1.15e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°5/10 with lambda = 80.774...

- - - < CRITERION > - - -

LL : -3250.241

BIC : 6564.954

BICc : 6617

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	1.01e-01
alpha_12_pop	6.38e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	2.46e-01
phi_S_pop	5.96e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.95e-01
omega_delta_S	1.58e-01
omega_phi_S	6.12e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°6/10 with lambda = 161.165...

- - - < CRITERION > - - -

LL : -3250.432
BIC : 6565.336
BICc : 6617

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	9.75e-02
alpha_12_pop	4.41e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	2.4e-01
phi_S_pop	5.79e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.97e-01
omega_delta_S	1.4e-01
omega_phi_S	6.22e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01

sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°7/10 with lambda = 321.566...

- - - < CRITERION > - - -

LL : -3251.821

BIC : 6568.115

BICc : 6620

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	9.35e-02
alpha_12_pop	2.54e-04
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	2.3e-01
phi_S_pop	5.59e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.95e-01
omega_delta_S	1.16e-01
omega_phi_S	6.32e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°8/10 with lambda = 641.609...

- - - < CRITERION > - - -

LL : -3256.438

BIC : 6572.744

BICc : 6620

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	8.42e-02
alpha_12_pop	0.00e+00
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00

alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	2.17e-01
phi_S_pop	5.27e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.95e-01
omega_delta_S	1.81e-01
omega_phi_S	6.18e-01
sigma_AB	1.14e-01
sigma_G1	2.76e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°9/10 with lambda = 1280.178...

```

- - - < CRITERION > - - -
LL : -3268.485
BIC : 6596.836
BICc : 6644

```

```

- - - < FINAL PARAMETERS > - - -

```

	EstimatedValue
alpha_11_pop	7.08e-02
alpha_12_pop	0.00e+00
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	2.07e-01
phi_S_pop	5.1e+02
delta_AB_pop	2.5e-02
omega_delta_AB	2.96e-01
omega_delta_S	1.21e-01
omega_phi_S	6.15e-01
sigma_AB	1.15e-01
sigma_G1	2.77e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Starting algorithm n°10/10 with lambda = 2554.292...

```

- - - < CRITERION > - - -

```

LL : -3307.982
 BIC : 6675.832
 BICc : 6723

- - - < FINAL PARAMETERS > - - -

	EstimatedValue
alpha_11_pop	4.94e-02
alpha_12_pop	0.00e+00
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00
	EstimatedValue
delta_S_pop	1.85e-01
phi_S_pop	4.68e+02
delta_AB_pop	2.52e-02
omega_delta_AB	2.93e-01
omega_delta_S	1.22e-01
omega_phi_S	6.22e-01
sigma_AB	1.15e-01
sigma_G1	2.82e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

B.2.6 Summary file for REMix algorithm

 Starting Regulatization and Estimation Algorithm

- - - < INITIAL PARAMETERS > - - -

	EstimatedValue
delta_S_pop	2.63e-01
phi_S_pop	6.28e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.94e-01
omega_delta_S	9.57e-02
omega_phi_S	6.29e-01
sigma_AB	1.14e-01
sigma_G1	2.76e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

```

                EstimatedValue
alpha_11_pop    1.07e-01
alpha_12_pop    0e+00
alpha_13_pop    0e+00
alpha_14_pop    -1.09e-05
alpha_15_pop    0e+00

Estimating the log-likelihood, and its derivates, using the initial model ...
      LL : -3252.263
      LL.pen : -3260.91
      time elapsed : 18.76s
-----

```

ITERATION 1

Computing taylor update for regularization parameters...

```

                EstimatedValue
alpha_11_pop    1.05e-01
alpha_12_pop    5.98e-03
alpha_13_pop    0.00e+00
alpha_14_pop    0.00e+00
alpha_15_pop    0.00e+00

```

Computing SAEM update for population parameters...

```

                EstimatedValue
delta_S_pop     2.47e-01
phi_S_pop       5.94e+02
delta_AB_pop    2.48e-02
omega_delta_AB  2.96e-01
omega_delta_S   1.3e-01
omega_phi_S     6.32e-01
sigma_AB        1.14e-01
sigma_G1        2.75e-01
sigma_G4        3.57e-01
sigma_G2        3.35e-01
sigma_G3        3.88e-01
sigma_G5        3.19e-01

```

Estimating penalised log-likelihood...

```

      LL : -3250.181
      LL.pen : -3259.167

```

```

Current parameter convergence criterion : 0.005
Current ll pen convergence criterion   : 1.743
time elapsed : 38.8s
-----

```

ITERATION 2

Computing taylor update for regularization parameters...

	EstimatedValue
alpha_11_pop	1.01e-01
alpha_12_pop	6.38e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00

Computing SAEM update for population parameters...

	EstimatedValue
delta_S_pop	2.46e-01
phi_S_pop	5.96e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.95e-01
omega_delta_S	1.58e-01
omega_phi_S	6.12e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL : -3250.241

LL.pen : -3258.934

Current parameter convergence criterion : 0.001

Current ll pen convergence criterion : 0.233

time elapsed : 39.17s

 ITERATION 3

Computing taylor update for regularization parameters...

	EstimatedValue
alpha_11_pop	1.01e-01
alpha_12_pop	6.43e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00

Computing SAEM update for population parameters...

	EstimatedValue
delta_S_pop	2.47e-01
phi_S_pop	5.91e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.99e-01
omega_delta_S	1.35e-01
omega_phi_S	6.22e-01

sigma_AB	1.13e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL :-3250.115

LL.pen :-3258.8

Current parameter convergence criterion : 0.001

Current ll pen convergence criterion : 0.134

time elapsed : 41.77s

 ITERATION 4

Computing taylor update for regularization parameters...

EstimatedValue

alpha_11_pop	1.01e-01
alpha_12_pop	6.40e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00

Computing SAEM update for population parameters...

EstimatedValue

delta_S_pop	2.51e-01
phi_S_pop	6e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.96e-01
omega_delta_S	2.08e-01
omega_phi_S	6.12e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL :-3251.027

LL.pen :-3259.722

Current parameter convergence criterion : 0.005

Current ll pen convergence criterion : 0.923

time elapsed : 41.73s

```

-----
                        ITERATION  5

Computing taylor update for regularization parameters...
      EstimatedValue
alpha_11_pop      1.02e-01
alpha_12_pop      6.41e-03
alpha_13_pop      0.00e+00
alpha_14_pop      0.00e+00
alpha_15_pop      0.00e+00

Computing SAEM update for population parameters...
      EstimatedValue
delta_S_pop      2.39e-01
phi_S_pop      5.75e+02
delta_AB_pop      2.49e-02
omega_delta_AB      2.97e-01
omega_delta_S      1.86e-01
omega_phi_S      6.27e-01
sigma_AB      1.13e-01
sigma_G1      2.75e-01
sigma_G4      3.57e-01
sigma_G2      3.35e-01
sigma_G3      3.88e-01
sigma_G5      3.19e-01

Estimating penalised log-likelihood...
      LL : -3250.337
      LL.pen : -3259.068

Current parameter convergence criterion : 0.003
Current ll pen convergence criterion : 0.654
time elapsed : 43.41s
-----

```

```

-----
                        ITERATION  6

Computing taylor update for regularization parameters...
      EstimatedValue
alpha_11_pop      9.91e-02
alpha_12_pop      6.59e-03
alpha_13_pop      0.00e+00
alpha_14_pop      0.00e+00
alpha_15_pop      0.00e+00

Computing SAEM update for population parameters...
      EstimatedValue
delta_S_pop      2.36e-01
phi_S_pop      5.76e+02

```

delta_AB_pop	2.49e-02
omega_delta_AB	2.96e-01
omega_delta_S	1.71e-01
omega_phi_S	6.23e-01
sigma_AB	1.13e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL : -3250.156

LL.pen : -3258.695

Current parameter convergence criterion : 0

Current ll pen convergence criterion : 0.374

time elapsed : 41.75s

 ITERATION 7

Computing taylor update for regularization parameters...

EstimatedValue

alpha_11_pop	9.87e-02
alpha_12_pop	6.63e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00

Computing SAEM update for population parameters...

EstimatedValue

delta_S_pop	2.32e-01
phi_S_pop	5.62e+02
delta_AB_pop	2.49e-02
omega_delta_AB	2.98e-01
omega_delta_S	1.24e-01
omega_phi_S	6.23e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL : -3249.774

LL.pen : -3258.28

Current parameter convergence criterion : 0.003
 Current ll pen convergence criterion : 0.415
 time elapsed : 43.35s

 ITERATION 8

Computing taylor update for regularization parameters...

	EstimatedValue
alpha_11_pop	9.76e-02
alpha_12_pop	6.68e-03
alpha_13_pop	0.00e+00
alpha_14_pop	0.00e+00
alpha_15_pop	0.00e+00

Computing SAEM update for population parameters...

	EstimatedValue
delta_S_pop	2.38e-01
phi_S_pop	5.71e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.98e-01
omega_delta_S	1.11e-01
omega_phi_S	6.24e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

Estimating penalised log-likelihood...

LL : -3249.909
 LL.pen : -3258.33

Current parameter convergence criterion : 0
 Current ll pen convergence criterion : 0.05

B.2.7 Summary file for final tests and estimation

 Starting Final Regulatization and Estimation Algorithm Step

- - - < INITIAL PARAMETERS > - - -

	EstimatedValue
delta_S_pop	2.38e-01

phi_S_pop	5.71e+02
delta_AB_pop	2.48e-02
omega_delta_AB	2.98e-01
omega_delta_S	1.11e-01
omega_phi_S	6.24e-01
sigma_AB	1.14e-01
sigma_G1	2.75e-01
sigma_G4	3.57e-01
sigma_G2	3.35e-01
sigma_G3	3.88e-01
sigma_G5	3.19e-01

	EstimatedValue
alpha_11_pop	9.76e-02
alpha_12_pop	6.68e-03
alpha_13_pop	0e+00
alpha_14_pop	0e+00
alpha_15_pop	0e+00

Computing final SAEM...

Estimating log-likelihood...

	EstimatedValue	se	CI_95
delta_S_pop	2.32e-01	8.5e-03	[2.16e-01;2.49e-01]
phi_S_pop	5.64e+02	4.03e+01	[4.85e+02;6.43e+02]
delta_AB_pop	2.49e-02	7.45e-04	[2.35e-02;2.64e-02]
alpha_11_pop	9.91e-02	4.72e-03	[8.99e-02;1.08e-01]
alpha_12_pop	9.1e-03	5.15e-03	[-9.87e-04;1.92e-02]
alpha_13_pop	0e+00		
alpha_14_pop	0e+00		
alpha_15_pop	0e+00		
omega_delta_S	1.44e-01	3.28e-02	[7.98e-02;2.08e-01]
omega_phi_S	6.27e-01	5.12e-02	[5.26e-01;7.27e-01]
omega_delta_AB	2.95e-01	2.14e-02	[2.53e-01;3.37e-01]
sigma_AB	1.14e-01	4.01e-03	[1.06e-01;1.22e-01]
sigma_G1	2.75e-01	4.26e-03	[2.67e-01;2.84e-01]
sigma_G2	3.35e-01	5.17e-03	[3.25e-01;3.45e-01]
sigma_G3	3.88e-01	5.98e-03	[3.76e-01;4e-01]
sigma_G4	3.57e-01	5.5e-03	[3.46e-01;3.68e-01]
sigma_G5	3.19e-01	4.93e-03	[3.1e-01;3.29e-01]

- - - < CRITERION > - - -

LL : -3249.806

BIC : 6564.085

BICc : 6615.889

Computing Wald test (with null hypothesis alpha1=0)...

	EstimatedValue	CI_95	stat.test	p.value
alpha_11_pop	9.91e-02	[8.99e-02;1.08e-01]	21.011	0 <0.05
alpha_12_pop	9.1e-03	[-9.87e-04;1.92e-02]	1.768	0.077

```

alpha_13_pop      0e+00
alpha_14_pop      0e+00
alpha_15_pop      0e+00
  Setting parameters to 0...
(re)Computing final SAEM...
Estimating log-likelihood...

```

```

- - - < FINAL PARAMETERS > - - -

```

	EstimatedValue	se	CI_95
delta_S_pop	2.43e-01	9.5e-03	[2.24e-01;2.62e-01]
phi_S_pop	5.86e+02	4.14e+01	[5.05e+02;6.67e+02]
delta_AB_pop	2.49e-02	7.47e-04	[2.34e-02;2.63e-02]
alpha_11_pop	1.02e-01	4.92e-03	[9.24e-02;1.12e-01]
alpha_12_pop	0e+00		
alpha_13_pop	0e+00		
alpha_14_pop	0e+00		
alpha_15_pop	0e+00		
omega_delta_S	1.66e-01	5.05e-02	[6.66e-02;2.65e-01]
omega_phi_S	6.14e-01	5.01e-02	[5.15e-01;7.12e-01]
omega_delta_AB	2.96e-01	2.16e-02	[2.53e-01;3.38e-01]
sigma_AB	1.14e-01	4.08e-03	[1.06e-01;1.22e-01]
sigma_G1	2.75e-01	4.26e-03	[2.67e-01;2.83e-01]
sigma_G2	3.35e-01	5.17e-03	[3.25e-01;3.45e-01]
sigma_G3	3.88e-01	5.98e-03	[3.76e-01;4e-01]
sigma_G4	3.57e-01	5.5e-03	[3.46e-01;3.68e-01]
sigma_G5	3.19e-01	4.93e-03	[3.1e-01;3.29e-01]

```

- - - < CRITERION > - - -

```

```

  LL : -3251.378
  BIC : 6562.624
  BICc : 6609.719

```


Bibliographie

- AALLEN, O. O., RØYSLAND, K., GRAN, J. M. et LEDERGERBER, B. (2012). Causality, mediation and time : a dynamic viewpoint. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 175(4):831–861.
- ABBAS, A. K., LICHTMAN, A. H. et PILLAI, S. (2021). *Cellular and molecular immunology*. Elsevier, tenth edition. édition.
- AKAIKE, H. (1973). *Information theory and an extension of the maximum likelihood principle*, pages 267–281. Akadémiai Kiadó.
- ALBERTS, B., JOHNSON, A., LEWIS, J., MORGAN, D., RAFF, M., ROBERTS, K. (2014). *Molecular Biology of the Cell*. Garland Science, 6th édition.
- ALEXANDRE, M. (2022). *Mechanistic modeling and optimization of vaccine response in infectious diseases : application to HIV, Ebola and SARS-CoV-2*. Thèse de doctorat, Université de Bordeaux.
- ALEXANDRE, M., PRAGUE, M., MCLEAN, C., BOCKSTAL, V., DOUGUIH, M., THIÉBAUT, R., *et al.* (2023). Prediction of long-term humoral response induced by the two-dose heterologous Ad26.ZEBOV, MVA-BN-Filo vaccine against Ebola. *npj Vaccines*, 8:174.
- ALGAMAL, Z. Y. et LEE, M. H. (2015). Adjusted Adaptive LASSO in high-dimensional poisson regression model. *Modern Applied Science*, 9.
- ANDRAUD, M., LEJEUNE, O., MUSORO, J. Z., OGUNJIMI, B., BEUTELS, P. et HENS, N. (2012). Living on three time scales : The dynamics of plasma cell and antibody populations illustrated for Hepatitis A virus. *PLoS Computational Biology*, 8:e1002418.
- ARVIN, A. M. (1996). Varicella-zoster virus. *Clinical Microbiology Reviews*, 9(3):361–381.
- AYRAL, G., SI ABDALLAH, J.-F., MAGNARD, C. et 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. H., KENNEDY, S. B., DOUMBIA, S., *et al.* (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*, 22:86.
- BAEY, C., MATHIEU, A., JULLIEN, A., TREVEZAS, S. et COURNÈDE, P.-H. (2018). Mixed-effects estimation in dynamic models of plant growth for the assessment of inter-individual variability. *Journal of Agricultural, Biological and Environmental Statistics*, 23:208–232.

- BALELLI, I., PASIN, C., PRAGUE, M., CRAUSTE, F., EFFELTERRE, T. V., BOCKSTAL, V., *et al.* (2020). A model for establishment, maintenance and reactivation of the immune response after vaccination against Ebola virus. *Journal of Theoretical Biology*, 495:110254.
- BEAL, S. L. et SHEINER, L. B. (1982). Estimating population kinetics. *Critical reviews in biomedical engineering*, 8:195–222.
- BELLU, G., SACCOMANI, M. P., AUDOLY, S. et D’ANGIÒ, L. (2007). DAISY : A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88:52–61.
- BENJAMINI, Y. et HOCHBERG, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1):289–300.
- BERTRAND, J. et BALDING, D. J. (2013). Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenetics and genomics*, 23(3):167–174.
- BERTRAND, J., DE IORIO, M. et BALDING, D. (2015). Integrating dynamic mixed-effect modelling and penalized regression to explore genetic association with pharmacokinetics. *Pharmacogenetics and genomics*, 25.
- BHATTACHARYA, S., DUNN, P., THOMAS, C., SCHAEFER, H., CHEN, J., HU, Z., *et al.* (2018a). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, 5:180015.
- BHATTACHARYA, S., DUNN, P., THOMAS, C., SCHAEFER, H., CHEN, J., HU, Z., *et al.* (2018b). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, 5:180015.
- BLOCKER, A. W. (2022). *fastGHQuad : Fast ‘Rcpp’ Implementation of Gauss-Hermite Quadrature*. R package version 1.0.1.
- BODINIER, B. (2024). *sharp : Stability-enhanced Approaches using Resampling Procedures*. R package version 1.4.6.
- BODINIER, B., FILIPPI, S., NØST, T. H., CHIQUET, J. et 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.
- BOGDAN, M., van den BERG, E., SABATTI, C., SU, W. et CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3).
- BONIN, C. R. B., FERNANDES, G. C., de MENEZES MARTINS, R., CAMACHO, L. A. B., TEIXEIRA-CARVALHO, A., da MOTA, L. M. H., *et al.* (2020). Validation of a yellow fever vaccine model using data from primary vaccination in children and adults, re-vaccination and dose-response in adults and studies with immunocompromised individuals. *BMC Bioinformatics*, 21:551.

- BONIN, C. R. B., FERNANDES, G. C., dos SANTOS, R. W. et LOBOSCO, M. (2017). Mathematical modeling based on ordinary differential equations : A promising approach to vaccinology. *Human Vaccines & Immunotherapeutics*, 13:484–489.
- BREIMAN, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- BURDEN, R. L. et FAIRES, J. D. (2000). *Numerical Analysis*. Brooks/Cole, 7th édition.
- BÜHLMANN, P. et van de GEER, S. (2011). *Variable selection with the Lasso*, pages 183–247. Springer Berlin Heidelberg, Berlin, Heidelberg.
- CAILLEBOTTE, A., KUHN, E. et LEMLER, S. (2024). Estimation and variable selection in a joint model of survival times and longitudinal outcomes with random effects.
- CHAPLIN, D. D. (2010). Overview of the immune response. *Journal of Allergy and Clinical Immunology*, 125:S3–S23.
- CLAESKENS, G., RELUGA, K. et SPERLICH, S. (2021). Post-selection inference for linear mixed model parameters using the conditional Akaike information criterion. Rapport technique. Working paper, 39 pages, 7 figures.
- CLAIRON, Q., PASIN, C., BALELLI, I., THIÉBAUT, R. et PRAGUE, M. (2023a). Parameter estimation in nonlinear mixed effect models based on ordinary differential equations : an optimal control approach. *Computational Statistics*, pages 1–31.
- CLAIRON, Q., PRAGUE, M., PLANAS, D., BRUEL, T., HOCQUELOUX, L., PRAZUCK, T., *et al.* (2023b). Modeling the kinetics of the neutralizing antibody response against SARS-CoV-2 variants after several administrations of Bnt162b2. *PLOS Computational Biology*, 19: e1011282.
- COMETS, E., LAVENU, A. et LAVIELLE, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix , an R implementation of the SAEM algorithm. *Journal of Statistical Software*, 80.
- CURTISS, C. F. et HIRSCHFELDER, J. O. (1952). Integration of stiff equations. *Proceedings of the National Academy of Sciences*, 38:235–243.
- DARI, A., JACQMIN, P., IWAKI, Y., NEYENS, M., GARS, M. L., SADOFF, J., *et al.* (2023). Mechanistic modeling projections of antibody persistence after homologous booster regimens of COVID-19 vaccine Ad26.COV2.S in humans. *CPT : Pharmacometrics & Systems Pharmacology*, 12:1485–1498.
- DAVIDIAN, M. et GALLANT, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics*, 20:529–556.
- DAVIDIAN, M. et GALLANT, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80:475–488.

- DELANY, I., RAPPUOLI, R. et GREGORIO, E. D. (2014). Vaccines for the 21st century. *EMBO Molecular Medicine*, 6:708–720.
- DELATTRE, M., LAVIELLE, M. et POURSAT, M.-A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, 8.
- DELYON, B., LAVIELLE, M. et MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27.
- DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 39:1–22.
- den HAAN, J. M., ARENS, R. et van ZELM, M. C. (2014). The activation of the adaptive immune system : Cross-talk between antigen-presenting cells, T cells and B cells. *Immunology Letters*, 162:103–112.
- DONG, R., GOODBRAKE, C., HARRINGTON, H. A. et POGUDIN, G. (2023). Differential elimination for dynamical models via projections with applications to structural identifiability. *SIAM Journal on Applied Algebra and Geometry*, 7:194–235.
- DURINCK, S., MOREAU, Y., KASPRZYK, A., DAVIS, S., DE MOOR, B., BRAZMA, A. (2005). BioMart and Bioconductor : a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440.
- EFRON, B., HASTIE, T., JOHNSTONE, I. et TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407 – 499.
- Eurosurveillance editorial TEAM, C. (2014). The innovative medicines initiative launches call on ebola and other filoviral haemorrhagic fevers. *Eurosurveillance*, 19(45).
- FAN, J. et LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.
- FENG, Y. et YU, Y. (2019). The restricted consistency property of leave-nv-out cross-validation for high-dimensional variable selection. *Statistica Sinica*.
- FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.
- FORT, G., OLLIER, E. et SAMSON, A. (2019). Stochastic proximal-gradient algorithms for penalized mixed models. *Statistics and Computing*, 29:231–253.
- FRIEDMAN, J., HASTIE, T. et TIBSHIRANI, R. (2010a). A note on the group lasso and a sparse group lasso.
- FRIEDMAN, J., HASTIE, T. et TIBSHIRANI, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22.

- FRIEL, N. et PETTITT, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 70(3):589–607.
- GABAUT, A., BERCU, A., HEJBLUM, B., THIÉBAUT, R., PROUST-LIMA, C. et PRAGUE, M. (2025a). Regularized Estimation for non-linear Mixed effect models in the presence of large-dimension biomarkers. En préparation pour être soumis.
- GABAUT, A., THIÉBAUT, R., PROUST-LIMA, C. et PRAGUE, M. (2025b). A stability-enhanced lasso approach for covariate selection in non-linear mixed effect model. *medRxiv*.
- GANSER, I., BUCKERIDGE, D. L., HEFFERNAN, J., PRAGUE, M. et THIÉBAUT, R. (2024). Estimating the population effectiveness of interventions against COVID-19 in France : A modelling study. *Epidemics*, 46:100744.
- GENOVESE, C. et WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 64(3):499–517.
- GENUER, R., POGGI, J.-M. et TULEAU-MALOT, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.
- GENZ, A. et BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.
- GIRAUD, C. (2021). *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC.
- GOLDFELD, K. et WUJCIAK-JENS, J. (2020). simstudy : Illuminating research methods through data generation. *Journal of Open Source Software*, 5(54):2763.
- GOLUB, G. H. et WELSCH, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23:221–230.
- GRÖMPING, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147.
- GU, M. G. et LI, S. (1998). A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data. *Canadian Journal of Statistics*, 26:567–582.
- HAEM, E., HARLING, K., AYATOLLAHI, S. M. T., ZARE, N. et KARLSSON, M. O. (2017). Adjusted adaptive lasso for covariate model-building in nonlinear mixed-effect pharmacokinetic models. *Journal of Pharmacokinetics and Pharmacodynamics*, 44:55–66.
- HAGAN, T., GERRITSEN, B., TOMALIN, L. E., FOURATI, S., MULÈ, M. P., CHAWLA, D. G., *et al.* (2022). Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses. *Nature Immunology*, 23:1788–1798.
- HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

- HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer New York.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97.
- HEINZE, G., WALLISCH, C. et DUNKLER, D. (2018). Variable selection—a review and recommendations for the practicing statistician. *Biometrical journal*, 60(3):431–449.
- HOEBE, K., JANSSEN, E. et BEUTLER, B. (2004). The interface between innate and adaptive immunity. *Nature Immunology*, 5:971–974.
- HOERL, A. E. et KENNARD, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12:55.
- IBARRA, M., LORIER, M. et TROCÓNIZ, I. F. (2021). Pharmacometrics : Population approach. In *The ADME Encyclopedia : A Comprehensive Guide on Biopharmacy and Pharmacokinetics*, pages 1–13. Springer.
- JONSSON, E. N. et KARLSSON, M. O. (1998). Automated covariate model building within NONMEM. *Pharmaceutical Research*, 15:1463–1468.
- KARLINE SOETAERT, THOMAS PETZOLDT et R. WOODROW SETZER (2010). Solving differential equations in R : Package deSolve. *Journal of Statistical Software*, 33(9):1–25.
- KLOEK, T. et van DIJK, H. K. (1978). Bayesian estimates of equation system parameters : An application of integration by Monte Carlo. *Econometrica*, 46:1.
- KONDĚLKOVÁ, K., VOKURKOVÁ, D., KREJSEK, J., BORSKÁ, L., FIALA, Z. et ANDRÝS, C. (2010). Regulatory T cells (Treg) and their roles in immune system with respect to immunopathological disorders. *Acta Medica (Hradec Kralove, Czech Republic)*, 53:73–77.
- KUHN, E. et LAVIELLE, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8:115–131.
- KUHN, E. et LAVIELLE, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, 49(4):1020–1038.
- LAIDLAW, B. J., CRAFT, J. E. et KAECH, S. M. (2016). The multifaceted role of CD4+ T cells in CD8+ T cell memory. *Nature Reviews Immunology*, 16:102–111.
- LAIRD, N. M. et WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963.
- LAVIELLE, M. (2014). *Mixed Effects Models for the Population Approach : Models, Tasks, Methods and Tools*. Chapman and Hall/CRC.
- LAVIELLE, M. et AARONS, L. (2016). What do we mean by identifiability in mixed effects models? *Journal of Pharmacokinetics and Pharmacodynamics*, 43:111–122.

- LAVIELLE, M. et MENTRÉ, F. (2007). Estimation of population pharmacokinetic parameters of saquinavir in HIV patients with the MONOLIX software. *Journal of pharmacokinetics and pharmacodynamics*, 34:229–249.
- LEHR, T., SCHAEFER, H.-G. et STAAB, A. (2010). Integration of high-throughput genotyping data into pharmacometric analyses using nonlinear mixed effects modeling. *Pharmacogenetics and genomics*, 20(7):442–450.
- LI, S., ROUPHAEL, N., DURAISINGHAM, S., ROMERO-STEINER, S., PRESNELL, S., DAVIS, C., *et al.* (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature immunology*, 15(2):195–204.
- LI, S., SULLIVAN, N. L., ROUPHAEL, N., YU, T., BANTON, S., MADDUR, M. S., *et al.* (2017). Metabolic phenotypes of response to vaccination in humans. *Cell*, 169:862–877.e17.
- LI, Y., WANG, S., SONG, P. X., WANG, N., ZHOU, L. et ZHU, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and its Interface*, 11:721–737.
- LIM, C. et YU, B. (2016). Estimation Stability With Cross-Validation (ESCV). *Journal of Computational and Graphical Statistics*, 25:464–492.
- LINDSTROM, M. J. et BATES, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014.
- LINDSTROM, M. J. et BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673.
- LIXOFT (2023). *lixoftConnectors : R connectors for Lixoft Suite (@Lixoft)*. Lixoft SAS, Antony, France. R package version 2023.1.
- LIXOFT (2024a). *lixoftConnectors : R connectors for Lixoft Suite (@Lixoft)*. R package version 2024.1.
- LIXOFT (2024b). Monolix Suite 2024R1. Software. <https://monolixsuite.slp-software.com/monolix/2024R1/>.
- LIXOFT (2024c). Simulx 2024R1. Software. <https://monolixsuite.slp-software.com/simulx/2024R1/>.
- LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 44:226–233.
- LOVE, M. I., HUBER, W. et ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.
- LOWE, R., SHIRLEY, N., BLEACKLEY, M., DOLAN, S. et SHAFEE, T. (2017). Transcriptomics technologies. *PLoS computational biology*, 13(5):e1005457.

- MALLOWS, C. L. (1973). Some comments on CP. *Technometrics*, 15:661.
- MARQUARDT, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11:431–441.
- MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.
- MEINSHAUSEN, N. et BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 72:417–473.
- MIHALJEVIC, F. (2023). *Rsmk : R Speaks 'Monolix'*. Lixoft SAS, Antony, France. R package version 2023.1.5.
- MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. et WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5:621–628.
- NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349.
- NAVEAU, M., KON KAM KING, G., RINCENT, R., SANSONNET, L. et DELATTRE, M. (2023). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm. *Statistics and Computing*, 34(1).
- NGUYEN, T., MOUKSASSI, M.-S., HOLFORD, N., AL-HUNITI, N., FREEDMAN, I., HOOKER, A. C., *et al.* (2017). Model evaluation of continuous data pharmacometric models : metrics and graphics. *CPT : pharmacometrics & systems pharmacology*, 6(2):87–109.
- NI, X., ZHANG, D. et ZHANG, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66(1):79–88.
- OBERMOSER, G., PRESNELL, S., DOMICO, K., XU, H., WANG, Y., ANGUIANO, E., *et al.* (2013). Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, 38:831–844.
- O'BRIEN, R. M., HUDSON, K. et STOCKARD, J. (2008). A mixed model estimation of age, period, and cohort effects. *Sociological Methods & Research*, 36:402–428.
- OLLIER, E. (2022). Fast selection of nonlinear mixed effect models using penalized likelihood. *Computational Statistics & Data Analysis*, 167:107373.
- PAGADALA, M., SEARS, T. J., WU, V. H., PÉREZ-GUIJARRO, E., KIM, H., CASTRO, A., *et al.* (2023). Germline modifiers of the tumor immune microenvironment implicate drivers of cancer risk and immunotherapy response. *Nature Communications*, 14:2744.
- PAN, J. et HUANG, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24:725–738.

- PARK, M. Y. et HASTIE, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 69(4):659–677.
- PASIN, C., BALELLI, I., EFFELTERRE, T. V., BOCKSTAL, V., SOLFOROSI, L., PRAGUE, M., *et al.* (2019). Dynamics of the humoral immune response to a prime-boost Ebola vaccine : Quantification and sources of variation. *Journal of Virology*, 93.
- PATRO, R., DUGGAL, G., LOVE, M. I., IRIZARRY, R. A. et KINGSFORD, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14:417–419.
- PERELSON, A. S. et RIBEIRO, R. M. (2018). Introduction to modeling viral infections and immunity. *Immunological reviews*, 285(1):5.
- PERROT-DOCKÈS, M., LÉVY-LEDUC, C., SANSONNET, L. et CHIQUET, J. (2018). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *Journal of Multivariate Analysis*, 166:78–97.
- PHILIPPS, V., BORIS, P. H., PRAGUE, M., COMMENGES, D. et PROUST-LIMA, C. (2021). Robust and efficient optimization using a Marquardt-Levenberg algorithm with R package `marqLevAlg`. *The R Journal*, 13:273.
- PINHEIRO, J. C. et BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12.
- PLOTKIN, S. (2014). History of vaccination. *Proceedings of the National Academy of Sciences*, 111:12283–12287.
- PRAGUE, M., COMMENGES, D., GRAN, J. M., LEDERGERBER, B., YOUNG, J., FURRER, H. (2017). Dynamic models for estimating the effect of HAART on CD4 in observational studies : Application to the Aquitaine Cohort and the Swiss HIV Cohort Study. *Biometrics*, 73(1):294–304.
- PRAGUE, M., GABAUT, A., BA, K., ALEXANDRE, M., HEJBLUM, B. et THIÉBAUT, R. (2025). Integrating gene expression from whole blood into mechanistic modeling of the immune system : illustration in a model of the antibody response to COVID vaccination. En préparation pour être soumis.
- PRAGUE, M. et 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.
- PULENDRAN, B. et AHMED, R. (2006). Translating innate immunity into immunological memory : Implications for vaccine development. *Cell*, 124:849–863.
- QIAN, W. et YANG, Y. (2013). Model selection via standard error adjusted adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 65:295–318.

- R CORE TEAM (2024). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RASINES, D. G. et YOUNG, G. A. (2022). Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614.
- RAUE, A., KREUTZ, C., MAIWALD, T., BACHMANN, J., SCHILLING, M., KLINGMÜLLER, U. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25:1923–1929.
- RIBBING, J., NYBERG, J., CASTER, O. et JONSSON, E. N. (2007). The lasso - a novel method for predictive covariate model building in nonlinear mixed effects models. *Journal of Pharmacokinetics and Pharmacodynamics*, 34:485–517.
- RIEDEL, S. (2005). Edward jenner and the history of smallpox and vaccination. *Baylor University Medical Center Proceedings*, 18:21–25.
- RINCHAI, D., DEOLA, S., ZOPPOLI, G., KABEER, B. S. A., TALEB, S., PAVLOVSKI, I., *et al.* (2022). High-temporal resolution profiling reveals distinct immune trajectories following the first and second doses of COVID-19 mRNA vaccines. *Science Advances*, 8.
- ROBERT, C. P. et CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer New York.
- ROBINSON, M. D., MCCARTHY, D. J. et SMYTH, G. K. (2010). edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140.
- SCHAPFL, M. A., LOMASTRO, G. M., BRAUN, V. Z., HIRAI, M., LEVINE, M. S., KIERMAIER, E., *et al.* (2024). Centrioles are frequently amplified in early B cell development but dispensable for humoral immunity. *Nature Communications*, 15(1):8890.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.
- SHAH, R. D. et SAMWORTH, R. J. (2013). Variable selection with error control : Another look at stability selection. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 75:55–80.
- SHEINER, L. B. et BEAL, S. L. (1980a). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-menten model : Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8:553–571.
- SHEINER, L. B. et BEAL, S. L. (1980b). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model : Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8:553–571.
- SHEINER, L. B., ROSENBERG, B. et MARATHE, V. V. (1977). Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *Journal of Pharmacokinetics and Biopharmaceutics*, 5:445–479.
- SHEINER, L. B., ROSENBERG, B. et MELMON, K. L. (1972). Modelling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research*, 5:441–459.

- STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:111–147.
- SUN, J. et BASU, S. (2024). Penalized joint models of high-dimensional longitudinal biomarkers and a survival outcome. *The Annals of Applied Statistics*, 18(2):1490 – 1505.
- SVENSSON, R. J. et JONSSON, E. N. (2022). Efficient and relevant stepwise covariate model building for pharmacometrics. *CPT : Pharmacometrics & Systems Pharmacology*, 11(9):1210–1222.
- THE HUMAN PROTEIN ATLAS (2025). FAM216A expression. Accessed : 2025-08-21.
- THIÉBAUT, R., LHOMME, E., HOCINI, H., PELLEGRIN, I., BOIZARD-MORACCHINI, A., DUVI-GNAUD, A., *et al.* (2025). Self-collected finger-prick blood for gene expression profiling : Unveiling early immune responses in mild COVID-19. *medRxiv*.
- THIÉBAUT, R., HEJBLUM, B. P., HOCINI, H., BONNABAU, H., SKINNER, J., MONTES, M., *et al.* (2019). Gene expression signatures associated with immune and virological responses to therapeutic vaccination with dendritic cells in hiv-infected individuals. *Frontiers in Immunology*, 10.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58:267–288.
- TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(3):273–282.
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. et KNIGHT, K. (2005). Sparsity and smoothness via the Fused Lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67:91–108.
- TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization 1.
- UHLÉN, M., FAGERBERG, L., HALLSTRÖM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., *et al.* (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.
- VERBEKE, J. et COOLS, R. (1995). The Newton-Raphson method. *International Journal of Mathematical Education in Science and Technology*, 26:177–193.
- VIVIER, E., TOMASELLO, E., BARATIN, M., WALZER, T. et UGOLINI, S. (2008). Functions of natural killer cells. *Nature Immunology*, 9:503–510.
- WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482.
- WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R. et LI, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10:380.

- WANG, Y. (2007). Derivation of various NONMEM estimation methods. *Journal of Pharmacokinetics and pharmacodynamics*, 34:575–593.
- WEBER, A. P. (2015). Discovering new biology through RNA-Seq. *Plant Physiology*, page pp.01081.2015.
- YU, T. et XIANG, L. (2024). An iterative two-step regularization approach for selection of fixed and random effects in generalized linear mixed models. *Japanese Journal of Statistics and Data Science*.
- YUAN, M. et LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 68:49–67.
- YUAN, M., ZHU, Z., YANG, Y., ZHAO, M., SASSER, K., HAMADEH, H., *et al.* (2021). Efficient algorithms for covariate analysis with dynamic data using nonlinear mixed-effects model. *Statistical Methods in Medical Research*, 30(1):233–243.
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38.
- ZHANG, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, 4(7).
- ZHAO, P., TANG, A. et TANG, N. (2014). Double penalized variable selection procedure for partially linear models with longitudinal data. *Acta Mathematica Sinica, English Series*, 30.
- ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.
- ZOU, H. et HASTIE, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67(2):301–320.

Méthodes de régularisation pour l'intégration de données de grande dimension dans les modèles mécanistes : application pour le développement de vaccins.

Résumé : La vaccination est hautement efficace pour prévenir les maladies infectieuses, mais la réponse immunitaire aux vaccins varie considérablement d'un individu à l'autre, et les mécanismes sous-jacents ne sont pas entièrement compris. La modélisation mathématique a le potentiel d'améliorer notre compréhension des réponses immunitaires induites par les vaccins. Cependant, cela nécessite de nouvelles méthodes pour traduire des observations complexes, bruitées, incomplètes et de grande dimension en modèles mécanistes. Cette thèse s'inscrit dans cette perspective et propose de nouvelles méthodes visant à intégrer des marqueurs de grande dimension, comme les données transcriptomiques, à des processus décrits par des modèles mécanistes basés sur des équations différentielles ordinaires, représentant les processus cellulaires sous-jacents impliqués dans la réponse immunitaire à la vaccination. L'hétérogénéité interindividuelle est modélisée à l'aide d'effets mixtes appliqués aux paramètres du modèle mécaniste. L'intégration de ces données repose sur des techniques de régularisation adaptées à la haute dimension.

Le premier axe de la thèse s'intéresse à l'intégration de marqueurs transcriptomiques mesurés à un temps fixe, utilisés comme variables explicatives de la variabilité interindividuelle dans les paramètres du modèle mécaniste. La méthode développée repose sur un algorithme itératif combinant régression pénalisée de type Lasso et estimation des paramètres par l'algorithme SAEM, afin de construire un modèle final parcimonieux et interprétable. Le second axe vise à intégrer des données transcriptomiques longitudinales. Il repose sur l'hypothèse que certains biomarqueurs - en particulier les profils d'expression génique - reflètent l'évolution de compartiments non observés du modèle mécaniste, tels que des populations cellulaires. Un algorithme de descente par coordonnées est proposé, alternant entre régularisation Lasso des biomarqueurs et estimation des paramètres de population via SAEM, dans le but de maximiser une vraisemblance jointe pénalisée. Cette approche permet d'effectuer simultanément la sélection des marqueurs associés aux compartiments et l'estimation des paramètres du modèle. Ces deux contributions ont donné lieu à des paquets R accessibles sur le CRAN ou GitHub. Les méthodes proposées sont appliquées à des données d'essais vaccinaux contre Ebola, la varicelle et le SARS-CoV-2, afin d'identifier quels marqueurs transcriptomiques décrivent le mieux le processus d'établissement de la réponse immunitaire après vaccination. La méthodologie développée peut être plus largement appliquée à d'autres processus biologiques complexes, y compris dans des pathologies comme le cancer, pour lesquels des modèles mécanistes et des données de grande dimension sont disponibles.

Mots-clés : Modèles mécanistes, Modèles à variables latentes, Grande dimension, Biomarqueurs, Vaccination

Regularization methods for high-dimensional data integration into mechanistic models : application for vaccine development.

Abstract: Vaccination is highly effective in preventing infectious diseases, but the immune response to vaccines varies considerably between individuals, and the underlying mechanisms are not fully understood. Mathematical modeling has the potential to improve our understanding of vaccine-induced immune responses. However, this requires new methods to translate complex, noisy, incomplete, and high-dimensional observations into mechanistic models. This thesis contributes to this by proposing new methods to integrate high-dimensional markers, such as transcriptomic data, into processes described by mechanistic models based on ordinary differential equations, which represent the underlying cellular mechanisms involved in the immune response to vaccination where inter-individual heterogeneity is accounted for using mixed effects applied to the parameters of the mechanistic model. The integration of these data relies on regularization techniques adapted to high-dimensional settings.

The first part of the thesis focuses on the integration of transcriptomic markers measured at a fixed time point, used as explanatory variables for the inter-individual variability in the mechanistic model parameters. The proposed method is based on an iterative algorithm combining Lasso-type penalized regression and parameter estimation via the SAEM algorithm, in order to build a final model that is both parsimonious and interpretable. The second part addresses the integration of longitudinal transcriptomic data. It is based on the assumption that certain biomarkers in particular gene expression profiles reflect the dynamics of unobserved compartments in the mechanistic model, such as cellular populations. A coordinate descent algorithm is proposed, alternating between Lasso regularization of the biomarkers and population parameter estimation via SAEM, aiming to maximize a penalized joint likelihood. This approach allows simultaneous biomarker selection and parameter estimation. These two contributions have led to R packages available on CRAN or GitHub. The proposed methods are applied to vaccine trial data for Ebola, varicella, and SARS-CoV-2, to identify which transcriptomic markers best describe the establishment of the immune response following vaccination. The methodology can be more broadly applied to other complex biological processes, including diseases such as cancer, for which mechanistic models and high-dimensional data are available.

Keywords: Mechanistic Models, Latent variable models, high-dimensional data, Biomarkers, Vaccination
