Master 2 : Probabilités et Modèles Aléatoires

MÉMOIRE DE STAGE

présenté et soutenu par

Florian Robert

le 13 septembre 2022

Estimations de modèles mécanistes par une approche variationnelle et séquentielle. Application au SARS-CoV-2.

Directeur de stage : **Mélanie Prague**
Contact université : **Irina Kourkova**

Équipe SISTM
Bordeaux Population Health
INSERM U1219, 146 rue Léo Saignat,
33076 Bordeaux CEDEX

La science pour la santé
From science to health

Description de la structure de stage

Le centre de recherche Bordeaux Population Health est co-animé par l'Université de Bordeaux et l'Institut National de la Santé Et de la Recherche Médicale (INSERM U1219). Il regroupe chercheurs, doctorants, post-doctorants et ingénieurs dans une même structure. Il est constitué de dix équipes de recherches qui étudient (i) la santé du cerveau tout au long de la vie ; (ii) les *data science* (intelligence artificielle, omiques, données longitudinales, données de pharmaco-épidémiologie, *et cetera*) ; (iii) les maladies infectieuses et leur prévention ; (iv) le vieillissement et la résilience ; (v) les déterminants environnementaux et sociaux de la santé.

L'équipe SISTM (*Statistics In Systems biology and Translationnal Medicine*) faisant partie du centre Bordeaux Population Health, de l'Institut National de la Santé Et de la Recherche Médicale (INSERM U1219) et de l'institut national de recherche en sciences et technologies du numérique (INRIA) m'accueille pour ce stage de Master 2. Cette équipe se consacre au développement de méthodes statistiques pour l'analyse intégrative des données à la médecine et à la biologie. Elle se concentre à l'étude de trois axes avec un double défi consistant à développer des méthodes pour traiter des données de grande dimension avec des échantillons de petite taille et à accélérer le développement de vaccins.

Ainsi dans l'Axe 1, l'information pertinente est extraite des données de grande dimension. Cette information est utilisée pour estimer les paramètres des modèles mécanistes dans l'Axe 2. Ces modèles mécanistes sont utilisés dans l'Axe 3 pour simuler les stratégies vaccinales optimales à évaluer dans les prochains essais cliniques.

Lors de ce stage, nous nous sommes consacrés à l'étude de l'Axe 2 par la construction de modèles mécanistes et l'estimation des paramètres par assimilation variationnelle des données. Nous avons effectué un second projet en collaboration avec Annabelle Collin de l'équipe MONC (Modélisation en ONCologie) de l'INRIA (institut national de recherche en sciences et technologies du numérique). Cette équipe cherche à construire des outils numériques basés sur des équations aux dérivées partielles et des méthodes statistiques pour mieux comprendre ou suivre l'évolution du cancer. Le second projet a pour but de poursuivre de précédents travaux fruits de la collaboration des équipes SISTM et MONC et traite l'assimilation séquentielle de données.

Remerciements

Mes premières pensées vont à ma directrice de stage Mélanie Prague pour son encadrement, sa pédagogie et sa patience. Je te remercie pour la confiance et le temps que tu m'as accordé, pour ce stage passionnant et pour tous les conseils que tu as pus me donner. Un grand merci pour toutes les connaissances que tu m'as partagées et les articles que tu m'as montrés.

Je souhaite remercier Marie Alexandre pour ses conseils, son temps et son aide précieuse. Merci pour les échanges que nous avons eu sur la construction des modèles ; merci de m'avoir partagé certains codes afin de gagner de précieuses heures.

Je souhaite ensuite remercier toute l'équipe SISTM et tout particulièrement les membres du bureau N311 pour leur accueil et leur intégration.

Je tiens à remercier Annabelle Collin pour son aide et sa patience. Merci de m'avoir partagé tes connaissances en analyse numérique avec une très grande pédagogie. Merci pour tout le temps que tu as passé à m'expliquer les codes que tu avais implémentés.

J'adresse mes remerciements à Irina Kourkova pour ses enseignements au cours de mon année de Master et son encadrement. Merci pour les conseils que vous avez pu me donner tout au long de cette année scolaire afin que je puisse réaliser mes objectifs.

Pour terminer, j'ai une pensée sentimentale pour Cécile qui n'a cessée de m'encourager dans les moments difficiles de cette année parisienne dont ce mémoire est le symbole d'un aboutissement. Merci pour ton amour et ta joie de vivre.

Table des matières

Introduction

En biologie et en médecine, nous souhaitons décrire et expliquer de nombreux phénomènes. Pour cela, les chercheurs peuvent utiliser deux sources d'études qu'il est judicieux de combiner : les modèles et les observations. D'une part, les modèles utilisés ne sont que des approximations plus ou moins réalistes des phénomènes biologiques qui ont réellement lieu. Cela repose sur des simplifications mathématiques et approximations numériques qui rendent les informations de ces modèles inexactes ou imparfaites. D'autre part, les observations obtenues suite à des expériences ou prélèvements sont sujettes à des erreurs de mesures. De plus, lors des études en laboratoires, tous les marqueurs expliquant la dynamique étudiée ne peuvent pas être ciblés. Il est alors important d'utiliser ces deux sources d'information, sélectionner la partie la plus fiable des observations et la propager dans le temps grâce aux modèles biologiques.

Il existe à ce jour deux grandes classes de méthodes d'assimilation de données qui correspondent à deux approches bien distinctes. Tout d'abord, l'assimilation *séquentielle* qui repose sur des considérations statistiques et qui procède par corrections successives de la prévision du modèle au fur et à mesure que des observations sont disponibles ; et l'assimilation *variationnelle*, qui consiste à ajuster au mieux une solution du modèle à toutes les observations disponibles tout au long de la période d'assimilation [2]. Dans un cadre purement linéaire, les techniques séquentielles et variationnelles sont très souvent équivalentes mais lorsque les modèles sont non linéaires – comme dans la plupart des phénomènes complexes – des différences apparaissent.

Ainsi, les méthodes d'assimilation de données consistent à résoudre un problème *inverse*. Un problème inverse est une situation dans laquelle on souhaite déterminer les causes d'un phénomène à partir des observations expérimentales de ses effets. La résolution d'un problème inverse passe en général une étape initiale de modélisation du phénomène, dite *problème direct* qui décrit comment les paramètres du modèles se traduisent en effets observables expérimentalement. Ensuite, à partir des mesures obtenues du phénomène réel, la démarche va consister à approximer au mieux les paramètres qui permettent de rendre compte de ces mesures. Cependant, la résolution numérique est rendue difficile par le fait que les observations à disposition ne suffisent pas à déterminer parfaitement tous les paramètres du modèle. Il est donc nécessaire d'ajouter des contraintes ou des hypothèses qui permettent de diminuer la taille du problème et le rendre résoluble.

Dans le chapitre 1, nous verrons une méthode d'assimilation variationnelle et une méthode séquentielle pour l'estimation des paramètres d'un système d'équations différentielles ordinaires (EDO).

Nous nous intéresserons à l'application de ces méthodes au cas des dynamiques de la charge virale et de la réponse immunitaire face au *SARS-CoV-2*. Nous décrirons dans le chapitre 2 les modèles et détaillerons les critères de sélection utilisés.

Le chapitre 3 sera consacré à l'étude de deux problèmes théoriques. Le premier consistera à l'estimation du taux de reproduction R_0 d'un agent pathogène. Le second abordera l'implémentation des schémas en temps pour la résolution des systèmes d'équations différentielles ordinaires.

Différentes expériences numériques ont été menées sur les modèles précédemment définis. Les différents résultats seront présentés dans le chapitre 4, qui précèdera une conclusion générale.

Chapitre 1

Les modèles mécanistes

Dans ce chapitre, nous présentons les notations employées et le bagage théorique nécessaire à la compréhension des chapitres suivants.

Avant toute chose, nous définirons les modèles non linéaires à effets mixtes qui seront notre objet central d'étude. Nous verrons ensuite deux classes de méthodes pour l'estimation des paramètres de ces modèles par assimilation de données.

La première méthode que nous étudierons sera une approche variationnelle reposant sur la maximisation de la vraisemblance. Cela consiste à chercher la valeur des paramètres la plus vraisemblable d'un système d'EDO à partir de connaissances disponibles.

La seconde méthode sera une approche séquentielle. Elle repose sur des études statistiques des états du système afin de trouver celui qui, statistiquement, correspond le mieux aux observations [2]. Nous nous restreindrons à la technique du filtrage de Kalman.

1.1 Les modèles non linéaires à effets mixtes

Dans le cas de données longitudinales, celles-ci sont collectées à des temps t_1, t_2, ..., t_n. Dans ce contexte, y_j représente la mesure d'une certaine quantité au temps t_j. Notre objectif est de trouver la meilleure modélisation de cette quantité au cours du temps à l'aide des données à notre disposition.
Pour cela, nous pouvons supposer qu'il existe une relation linéaire entre ces observations et d variables explicatives $(x_j^{(1)}, 1 \leq j \leq n)$, ..., $(x_j^{(d)}, 1 \leq j \leq n)$ de sorte que :

$$y_j = a_1 x_j^{(1)} + a_2 x_j^{(2)} + ... + a_d x_j^{(d)} + \epsilon_j, \tag{1.1}$$

où ϵ_j est une variable aléatoire représentant les erreurs résiduelles de notre modèle. Cependant, il n'est plus à démontrer que nous ne pouvons pas nous limiter aux modèles linéaires. Ainsi, nous allons étendre ce modèle aux relations non linéaires.

1.1.1 Les modèles non linéaires

Une manière d'étendre le modèle linéaire défini en (1.1) est d'écrire :

$$y_j = f(t_j; \phi) + \epsilon_j, \tag{1.2}$$

où ϕ est un vecteur de paramètres pour le modèle structurel f et $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$. Dire que ce modèle est non linéaire signifie que f n'est pas une fonction linéaire en les composantes de ϕ.

Dans les applications que nous rencontrerons, le modèle structurel f ne dépendra pas uniquement du temps et des paramètres ϕ mais sera décomposé en deux parties : un modèle mathématique et un modèle d'observation.

Le modèle mathématique sera défini par un système d'EDO qui dépend des paramètres ϕ. Ce système a pour but de décrire l'évolution de quantités dans les compartiments que nous étudierons (par exemple : la concentration d'un traitement dans le sang ou dans certains organes, la quantité de virus ou d'anticorps dans un milieu, *et cetera*). Comme les systèmes que nous considérerons ne seront pas linéaires, nous utiliserons la méthode de résolution numérique implémentée dans MONOLIX.

Le modèle d'observation sera une fonction qui dépend du temps, des paramètres ϕ et des valeurs de certains compartiments, obtenues par la résolution du système d'EDO.

Exemple 1
Le poids d'un rat peut être modélisé en utilisant le modèle non linéaire suivant :

$$y_j = w(t_j;\phi) + \epsilon_j, \tag{1.3}$$

où w et solution de l'EDO $\dot{W} = -k(W - a - b)$ et les paramètres du modèle sont $\phi = (a, b, k)$.

1.1.2 Les modèles non linéaires à effets mixtes

Les modèles présentés ci-dessus sont définis pour un seul individu. Si nous supposons maintenant que notre étude contient N individus, nous noterons y_{ij} la j^{ieme} observation de l'individu i prise à l'instant t_{ij}. Sachant que certains paramètres de notre modèle varient entre les individus, nous aimerions prendre en compte cette variabilité. Ceci nous amène à définir les modèles non linéaires à effets mixtes :

$$y_{ij} = f(t_{ij};\phi_i) + \epsilon_{ij}. \tag{1.4}$$

Le vecteur des paramètres ϕ_i de chaque individu peut se décomposer en *effets fixes* β et en *effets aléatoires* η_i de sorte à avoir $\phi_i = h(\beta, \eta_i)$ où η_i suit une loi normalement distribuée. Nous pouvons aussi ajouter aux paramètres ϕ_i des covariables en fonction des données que nous avons (par exemple : sexe, âge, poids, *et cetera*). Ainsi, nous aurons $\phi_i = h(\beta, \eta_i, c_i)$

Exemple 2
En reprenant l'exemple 1, nous pouvons ajouter des effets aléatoires sur nos paramètres individuels :

$$y_{ij} = w(t_{ij}, \phi_i) + \epsilon_{ij}, \tag{1.5}$$

où w et solution de l'EDO $\dot{W} = -k(W - a - b)$ et $\phi_i = (a_i, b_i, k_i)$ est défini par :

$$\begin{cases} a_i &= a_{pop} + \eta_{a,i} \\ \log(b_i) &= \log(b_{pop}) + \eta_{b,i} \\ k_i &= k_{pop} + \eta_{k,i} \end{cases} \tag{1.6}$$

Nous pouvons également ajouter un effet du sexe sur notre modèle. Pour cela, nous posons $c_i = 1$ si le rat est un mâle et $c_i = 0$ si le rat est une femelle. Nous pouvons, par exemple, faire agir le sexe sur les paramètres a et b de la manière suivante :

$$\begin{cases} a_i &= a_{pop} + \gamma c_i + \eta_{a,i} \\ \log(b_i) &= \log(b_{pop}) + \delta c_i + \eta_{b,i} \\ k_i &= k_{pop} + \eta_{k,i} \end{cases} \tag{1.7}$$

Ainsi, nous obtenons le vecteur d'effets fixes $\beta = (a_{pop}, b_{pop}, k_{pop}, \gamma, \delta)$

Il est aussi possible de supposer que les erreurs résiduelles varient en fonction des individus et au cours du temps. Nous pouvons modéliser cette variabilité par la fonction g qui peut dépendre de $\phi_i = h(\beta, \eta_i, c_i)$ et de paramètres additionnels ξ_i :

$$y_{ij} = f(t_{ij};\phi_i) + g(t_{ij};\phi_i,\xi_i).\epsilon_{ij} \tag{1.8}$$

Dans toute la suite, ϵ_{ij} suit une loi normale centrée réduite et le vecteur de paramètres individuels est défini par $\psi_i = (\phi_i, \xi_i) = k(\beta, \eta_i, c_i)$ où η_i suit une loi normale centrée. Nous écrirons nos modèles de la manière suivante :

$$\begin{cases} y_{ij} &= f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i)\epsilon_{ij} \\ \psi_i &= k(\beta, \eta_i, c_i) \end{cases} \tag{1.9}$$

Pour résumer, les modèles que nous étudions se décomposent en trois sous-modèles. Le modèle d'observation est représenté par la première équation du système précédent. Le modèle mathématique est représenté au travers de la fonction f qui est généralement définie par la solution d'un système d'EDO. La troisième partie, le modèle statistique, est représentée par la deuxième équation et définit les variabilités des paramètres entre les individus.

Lorsque nous étudions des phénomènes biologiques, certains paramètres sont parfois déterminés par des expériences *in vitro*. Notre objectif est de déterminer les autres paramètres en utilisant ces modèles mécanistes. La bonne estimation des paramètres nous permettra de décrire les phénomènes biologiques étudiés ou de mieux identifier leurs dynamiques.

1.2 Les algorithmes variationnels pour l'estimation de paramètres

Dans cette section, nous détaillons les principaux algorithmes et méthodes implémentés dans Monolix. Ce logiciel est très utilisé en pharmacométrie lors de l'utilisation de modèles non linéaires à effets mixtes. Nous utiliserons la version 2021R1 de ce logiciel pour l'estimation des modèles définis dans le chapitre 2. Pour plus de détails, le lecteur pourra se référer à l'ouvrage de Marc Lavielle [14] et à la documentation en ligne [15]. Nous prenons les notations suivantes :

- Lorsque la distribution des paramètres individuels ψ_i de l'individu i dépend de covariables c_i et de paramètres de la population θ, on écrit explicitement cette dépendance par : $p(\psi_i; c_i, \theta)$.
- On utilise le point virgule pour séparer les variables aléatoires et les variables non aléatoires et on utilise la virgule pour séparer les variables du même type. Par exemple, $p(y_i, \psi_i; c_i, \theta)$ représente la distribution conjointe de y_i et ψ_i qui dépend de c_i et θ.
- On utilise la barre verticale | pour définir les distributions conditionnelles.

Dans un premier temps, nous étudions l'implémentation de l'algorithme *Stochastic Approximation Expectation-Maximization* (SAEM) dans un cas particulier de modèles. Cet algorithme est implémenté dans Monolix pour estimer les paramètres de la population en maximisant la vraisemblance.

Dans un second temps, nous expliquons la méthode utilisée pour estimer les paramètres ayant une variabilité individuelle. Nous introduisons le phénomène de *shrinkage* qui peut intervenir lorsque nos données sont trop parcimonieuses.

Nous terminons en expliquant comment nous pouvons obtenir des intervalles de confiance pour les estimateurs des paramètres de la population grâce à une estimation de la matrice d'information de Fisher (FIM).

1.2.1 Estimation du maximum de vraisemblance des paramètres de la population

Estimer le maximum de la vraisemblance du vecteur de paramètres de la population θ revient au problème d'optimisation suivant :

$$\hat{\theta} = \arg\max_{\theta} p(y; \theta) = \arg\max_{\theta} \int p(y, \psi; \theta) d\psi \tag{1.10}$$

Pour résoudre ce problème d'estimation, nous allons utiliser l'algorithme SAEM implémenté dans MONOLIX. Cet algorithme est itératif et stochastique.

En effet, l'utilisateur doit donner les estimations de départ des paramètres aussi bonnes que possible afin d'espérer une meilleure convergence vers un maximum de la vraisemblance.

Comme l'algorithme est stochastique et que $p(y;\theta)$ ne sera en général pas concave, nous ne pouvons pas garantir une convergence vers le maximum global de vraisemblance. Cependant, il a été démontré que sous certaines conditions assez générales, il converge vers un maximum – global ou local – de la vraisemblance [9]. De plus, les trajectoires des paramètres au fur et à mesure des itérations dépendent de la suite des nombres aléatoires utilisée par l'algorithme. Ainsi, pour des raisons de reproductibilité, cette suite de nombres est fixée par défaut dans MONOLIX. Néanmoins, le fait d'utiliser différentes suites de nombres aléatoires peut permettre de vérifier la stabilité de notre solution.

Dans la suite de cette section, nous supposerons qu'il existe une transformation strictement monotone h telle que $z_i = h(\psi_i)$ soit un vecteur gaussien. Ainsi, on peut écrire la distribution conjointe de y_i et ψ_i comme :

$$p(y_i, z_i; \theta, c_i) = p(y_i|z_i)p(z_i; \theta, c_i) \tag{1.11}$$

où $z_i \sim \mathcal{N}(\mu(\beta, c_i), \Omega)$ et les paramètres de population sont $\theta = (\beta, \Omega)$.

L'implémentation de l'algorithme SAEM peut être compliquée lorsque nous regardons des modèles statistiques complexes tels que les modèles de mélange (en anglais : *mixed models*) ou les modèles de Markov cachés (en anglais : *hidden markov models*). Afin de comprendre comment l'algorithme est implémenté en pratique, nous allons nous limiter aux modèles pour des données longitudinales suivant :

$$\begin{cases} y_{ij} &= f(t_{ij}, z_i) + a\epsilon_{i,j} \\ z_i &\sim \mathcal{N}(\beta, \Omega) \end{cases} \tag{1.12}$$

où $(\epsilon_{i,j}, 1 \leq i \leq N, 1 \leq j \leq d)$ sont des variables aléatoires indépendantes de loi normale centrée réduite, $z_i = (z_{i,1}, z_{i,2}, ..., z_{i,d})^t$ est le vecteur dans $\mathbb{R}^d$ des paramètres individuels transformés, β le vecteur dans $\mathbb{R}^d$ des effets fixes et Ω la matrice de variance-covariance de taille $d \times d$ supposée définie positive. En notant $\theta = (\beta, \Omega, a)$, on peut décomposer le modèle de la manière suivante :

$$p(y_i, z_i; \theta) = p(y_i|z_i; a)p(z_i; \beta, \Omega) \tag{1.13}$$

Nous rentrons ainsi dans le cas où le modèle complet est de la forme

$$p(y, z; \theta) = \exp\Big(-\xi(\theta) + \tilde{S}(y, z) \cdot \phi(\theta)\Big) \tag{1.14}$$

où $\tilde{S}(y, z)$ est une statistique exhaustive du modèle avec ses valeurs dans l'ouvert $\mathcal{S}$ de $\mathbb{R}^m$. Dans ce cas, il existe une fonction $\tilde{\theta}$ telle que pour tout $s \in \mathcal{S}$:

$$\tilde{\theta}(s) := \arg\max_{\theta}(-\xi(\theta) + s \cdot \phi(\theta)) \tag{1.15}$$

La k$^{\text{ième}}$ itération de l'algorithme SAEM se décompose par les trois étapes suivantes :

- ♦ **Simulation :** Pour $i = 1, ..., N$, on simule $z_i^{(k)}$ suivant la distribution conditionnelle $p(z_i|y_i; \theta_{k-1})$ en utilisant l'algorithme de Metropolis-Hastings (voir section 1.2.2).
- ♦ **Approximation stochastique :** On met à jour s_{k-1} par :

$$s_k = s_{k-1} + \gamma_k(\tilde{S}(y, z^{(k)}) - s_{k-1}) \tag{1.16}$$

♦ **Maximisation :** On met à jour θ_{k-1} par :

$$\theta_k = \tilde{\theta}(s_k) \tag{1.17}$$

Le fait d'être dans ce cas particulier permet de simplifier les étapes d'approximation et de maximisation. Cependant, comme le lecteur pourra le vérifier par la suite, les modèles considérés dans le chapitre 2 font partie de cette famille de modèles.

L'algorithme implémenté dans MONOLIX est constitué de deux phases. L'objectif de la première est d'approcher un voisinage de la solution en peu d'itérations en fixant $\gamma_k = 1$. La seconde phase permet la convergence vers un maximum en fixant $\gamma_k = 1/k$. Le choix de ces valeurs pour γ_k permet de vérifier les conditions de convergence données dans les travaux de Bernard DELYON, Marc LAVIELLE et Eric MOULINES en 1999 [9].

Après un rappel théorique sur les statistiques exhaustives, nous donnerons les expressions de celles-ci et les étapes de l'algorithme SAEM dans le cas du modèle (1.12).

Définition 1
Une statistique $t = T(Y)$ est exhaustive pour un vecteur de paramètres θ si, conditionnellement à la statistique $t = T(Y)$, la distribution des données Y ne dépend pas du vecteur θ.

Il existe un théorème caractérisant les statistiques exhaustives.

Théorème 1 (Factorisation de Fisher-Neyman)
Notons la distribution des données Y par f_θ.
La statistique T est exhaustive pour θ si, et seulement si, il existe deux fonctions positives g_θ et h telles que $f_\theta(y) = h(y)g_\theta(T(y))$.
Autrement dit, la densité f peut être factorisée de sorte que l'un des facteur, h, ne dépende pas de θ et que l'autre facteur g_θ, qui dépend de θ, ne dépende de y qu'au travers de $T(y)$.

Exemple 3
Appliquons cet énoncé lorsque $Z_1, Z_2, ..., Z_N$ sont des vecteurs gaussiens indépendants tels que pour tout $i \in \{1, ..., N\}$, $Z_i \sim \mathcal{N}_d(\beta, \Omega)$ et $\theta = (\beta, \Omega)$. Nous notons $\bar{Z} = \frac{1}{N}\sum_{i=1}^N Z_i$

$$\begin{aligned} f_\theta(Z_1, ..., Z_N) &= \prod_{i=1}^N (2\pi)^{-d/2} (\det \Omega)^{-1/2} \exp\left(-\frac{1}{2}(Z_i - \beta)^t \Omega^{-1} (Z_i - \beta)\right) \\ &= \left[(2\pi)^{-d/2}(\det \Omega)^{-1/2}\right]^N \exp\left(-\frac{1}{2}\left(\sum_{i=1}^N (Z_i - \bar{Z})^t \Omega^{-1}(Z_i - \bar{Z}) + \sum_{i=1}^N (\bar{Z} - \beta)^t \Omega^{-1}(\bar{Z} - \beta) + \right.\right. \\ &\quad \left.\left. \sum_{i=1}^N (Z_i - \bar{Z})^t \Omega^{-1}(\bar{Z} - \beta) + \sum_{i=1}^N (\bar{Z} - \beta)^t \Omega^{-1}(Z_i - \bar{Z})\right)\right) \end{aligned}$$

Or, $\sum_{i=1}^N (Z_i - \bar{Z})^t \Omega^{-1}(\bar{Z} - \beta) = \sum_{i=1}^N (\bar{Z} - \beta)^t \Omega^{-1}(Z_i - \bar{Z}) = 0$. Donc :

$$\begin{aligned} f_\theta(Z_1, ..., Z_N) &= \left[(2\pi)^{-d/2}(\det \Omega)^{-1/2}\right]^N \exp\left(-\frac{1}{2}\sum_{i=1}^N (Z_i - \bar{Z})^t \Omega^{-1}(Z_i - \bar{Z})\right) \exp\left(-\frac{N}{2}(\bar{Z} - \beta)^t \Omega^{-1}(\bar{Z} - \beta)\right) \\ &= \left[(2\pi)^{-d/2}(\det \Omega)^{-1/2}\right]^N \exp\left(-\frac{1}{2} Tr\left(\sum_{i=1}^N (Z_i - \bar{Z})^t \Omega^{-1}(Z_i - \bar{Z})\right)\right) \exp\left(-\frac{N}{2}(\bar{Z} - \beta)^t \Omega^{-1}(\bar{Z} - \beta)\right) \end{aligned}$$

Or, on a :

$$Tr\left(\sum_{i=1}^{N}(Z_i-\bar{Z})^t\Omega^{-1}(Z_i-\bar{Z})\right) = Tr\left(\Omega^{-1}\sum_{i=1}^{N}(Z_i-\bar{Z})^t(Z_i-\bar{Z})\right)$$
$$= Tr\left(\Omega^{-1}\sum_{i=1}^{N}Z_iZ_i^t - N\bar{Z}\bar{Z}^t\right) \tag{1.18}$$

Ainsi, on a $f_\theta(Z_1,...,Z_N) = g_\theta(\tilde{S}(Z))$ où :

$$\tilde{S}(Z) = (\tilde{S}_1(Z),\tilde{S}_2(Z)) = \left(\sum_{i=1}^{N}Z_i, \sum_{i=1}^{N}Z_iZ_i^t\right) = \left(N\bar{Z}, \sum_{i=1}^{N}Z_iZ_i^t\right)$$

.
Donc, par le théorème de factorisation de Fisher-Neyman, $\tilde{S}(Z)$ est une statistique exhaustive pour (β,Ω).

De la même manière, on peut déterminer la statistique exhaustive pour a :

$$\begin{aligned} p(y|z_i;a) &= \prod_{i=1}^{N}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}a}\exp\left(-\frac{1}{2a^2}\left(y_{ij}-f(t_{ij},z_i)\right)^2\right) \\ &= (\sqrt{2\pi}a)^{-\sum_{i=1}^{N}n_i}\exp(-\frac{1}{2a^2}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(y_{ij}-f(t_{ij},z_i)\right)^2}_{=\widetilde{S}_3(y,z)}) \\ &= g_a(\widetilde{S}_3(y,z)) \end{aligned} \tag{1.19}$$

On conclut par le théorème de factorisation de Fisher-Neyman que $\widetilde{S}_3$ est bien une statistique exhaustive pour a.

Par ailleurs, comme $z_i \sim \mathcal{N}_d(\beta,\omega)$, les estimateurs du maximum de vraisemblance de (β,Ω) sont :

$$\widehat{\beta} = \bar{z} = \frac{1}{N}\sum_{i=1}^{N}z_i \quad \text{and} \quad \hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})(z_i-\bar{z})^t = \frac{1}{N}\sum_{i=1}^{N}z_iz_i^t - \bar{z}\bar{z}^t \tag{1.20}$$

et par passage au logarithme de l'expression (1.19), on obtient que la statistique suivante est un estimateur du maximum de vraisemblance de a :

$$\widehat{a^2} = \frac{1}{\sum_{i=1}^{N}n_i}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(y_{ij}-f(t_{ij},z_i)\right)^2 \tag{1.21}$$

Nous avons maintenant tous les outils nécessaires pour écrire la k$^{\text{ième}}$ itération de l'algorithme SAEM pour le modèle défini en (1.12).

- ♦ **Simulation :** Pour $i = 1,...,N$, on simule $z_i^{(k)}$ par m itérations de l'algorithme de Metropolis-Hastings suivant une distribution $p(z_i|y_i;\beta_{k-1},\Omega_{k-1})$.
- ♦ **Approximation Stochastique :** On actualise $s_k = (s_{k,1},s_{k,2},s_{k,3})$ par :

$$\begin{aligned} s_{k,1} &= s_{k-1,1} + \gamma_k\left(\textstyle\sum_{i=1}^{N}z_i^{(k)} - s_{k-1,1}\right) \\ s_{k,2} &= s_{k-1,2} + \gamma_k\left(\textstyle\sum_{i=1}^{N}z_i^{(k)}(z_i^{(k)})^t - s_{k-1,2}\right) \\ s_{k,3} &= s_{k-1,3} + \gamma_k\left(\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(y_{ij}-f(t_{ij},z_i)\right)^2 - s_{k-1,2}\right) \end{aligned} \tag{1.22}$$

♦ **Maximisation :** On actualise $(\beta_{k-1}, \Omega_{k-1}, a_{k-1})$ en utilisant les estimateurs du maximum de vraisemblance :

$$\begin{aligned} \beta_k &= \frac{s_{k,1}}{N} \\ \Omega_k &= \frac{s_{k,2}}{N} - \beta_k \beta_k^t \\ a_k^2 &= \frac{s_{k,3}}{\sum_{i=1}^N n_i} \end{aligned} \tag{1.23}$$

1.2.2 Estimation des paramètres individuels

Une fois que les paramètres de la population θ ont été estimés, notre objectif est de générer les valeurs de paramètres individuels en utilisant la loi conditionnelle :

$$p(\psi_i|y_i;\theta) = \frac{p(y_i,\psi_i;\theta)}{p(y_i)} \tag{1.24}$$

Cependant, il n'existe pas d'expression analytique de cette distribution lorsque le modèle n'est pas linéairement gaussien. Les méthodes de Monte Carlo par chaines de Markov construisent une chaine de Markov dont la distribution stationnaire est la distribution qui nous intéresse. Les états de la chaine après un grand nombre d'itérations sont ensuite utilisés pour créer un échantillon de la distribution d'intérêt.

L'algorithme de Métropolis-Hastings fait partie de cette famille de méthodes et est implémenté dans MONOLIX pour réaliser l'estimation des paramètres individuels. Il nous permet de simuler une suite $(\psi_i^{(l)}, l = 1, 2, ...)$ qui converge en loi vers la distribution $p(\psi_i|y_i;\theta)$. L'implémentation de l'algorithme et des explications plus détaillées se trouvent aux pages 255 à 258 de l'ouvrage de Marc LAVIELLE [14].

La suite $(\psi_i^{(l)}, l = 1, 2, ...)$ générée par l'algorithme de Metropolis-Hastings est utilisée pour calculer deux statistiques afin d'estimer les paramètres individuels ψ_i. La première permet l'approximation de la moyenne conditionnelle $\mathbb{E}(\psi_i|y_i;\widehat{\theta})$:

$$\widehat{\psi}_i^{\text{mean}} = \frac{1}{K}\sum_{k=1}^{K} \psi_i^{(k)} \tag{1.25}$$

La deuxième est le mode de la distribution conditionnelle (ou *Empirical Bayes Estimate* (EBE) dans la littérature) :

$$\widehat{\psi}_i^{\text{mode}} = \arg\max_{\psi_i} p(\psi_i|y_i;\widehat{\theta}) \tag{1.26}$$

Le choix entre ces deux estimateurs est arbitraire mais, par défaut, MONOLIX utilise le mode plutôt que la moyenne car il représente la valeur la plus probable de la distribution conditionnelle. Cependant, lorsque nous regarderons les graphiques diagnostiques pour sélectionner les modèles, nous nous exposerons à un phénomène de *shrinkage*.

En effet, lorsque nos données sont trop parcimonieuses, il est possible que le mode des distributions conditionnelles des individus soient trop proches du mode de la distribution de la population. Dans ce cas, les paramètres individuels sont concentrés autour du mode de la distribution de la population et ils représentent mal la variabilité individuelle. Pour éviter que les graphiques soient biaisés de cette manière et nous mènent à de fausses affirmations, nous pouvons simuler un échantillon des paramètres individuels pour chaque individu en utilisant la loi conditionnelle. Nous utilisons ensuite ces simulations pour construire nos graphiques diagnostiques et réduire les potentielles erreurs liées au manque d'information.

1.2.3 Estimation de la matrice d'information de Fisher

Lorsque nous estimons les paramètres de la population $\widehat{\theta}$, nous aimerions connaitre la variance de ces estimateurs afin de construire des intervalles de confiance. Pour cela, nous pouvons utiliser la FIM définie par :

$$I_y(\widehat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log(\mathcal{L}_y(\widehat{\theta})) \tag{1.27}$$

La matrice de variance-covariance de $\widehat{\theta}$ peut être estimée par l'inverse de la matrice d'information de Fisher. Pour estimer la matrice d'information de Fisher, deux méthodes sont implémentées dans MONOLIX.

La première ne peut être utilisée que dans le cas de données longitudinales et consiste à linéariser le modèle d'observation de l'individu i autour du vecteur de paramètres individuels estimé. L'objectif est d'approximer la distribution des observations y_i de l'individu i conditionnellement aux paramètres de cet individu par une loi normale.

La seconde est plus générale et utilise un algorithme de Monte Carlo par chaines de Markov. Contrairement à l'estimation des paramètres individuels, ce n'est pas l'algorithme de Metropolis-Hastings qui est utilisé ici. Pour plus d'informations à propos de ces deux méthodes implémentées, le lecteur pourra se référer aux pages 259 à 262 de l'ouvrage de Marc LAVIELLE [14].

Les erreurs standards (s.e.) pour chaque composante de $\widehat{\theta}$ représentent les écarts-types et sont obtenues en prenant la racine carrée des éléments diagonaux de la matrice de variance-covariance de $\widehat{\theta}$. Il est ensuite facile d'obtenir l'intervalle de confiance de niveau $1-\alpha$ pour le paramètre ψ_{pop} :

$$CI_{1-\alpha}(\psi_{pop}) = \left[\widehat{\psi_{pop}} + q_{\alpha/2}\widehat{s.e.}(\widehat{\psi_{pop}}), \widehat{\psi_{pop}} + q_{1-\alpha/2}\widehat{s.e.}(\widehat{\psi_{pop}})\right] \tag{1.28}$$

où q_α est le quantile d'ordre α de la loi normale centrée réduite.

Cependant, ces intervalles de confiance ne respectent pas les hypothèses que l'on a sur certains paramètres dans le cas de transformation (par exemple : $\psi_{pop} > 0$ lors d'une transformation logarithmique). Si h est une fonction croissante telle que $h(\psi_i) \sim \mathcal{N}(h(\psi_{pop}), \omega^2)$ et $\mu_{pop} = h(\psi_{pop})$. Nous pouvons alors construire un intervalle de confiance de niveau $1-\alpha$ pour μ_{pop} :

$$CI_{1-\alpha}(\mu_{pop}) = \left[\mu_{pop,\alpha/2}, \mu_{pop,1-\alpha/2}\right] \tag{1.29}$$

où $\mu_{pop,\alpha/2} = \widehat{\mu}_{pop} + q_{\alpha/2}\widehat{s.e.}(\widehat{\mu_{pop}})$ et $\mu_{pop,1-\alpha/2} = \widehat{\mu}_{pop} + q_{1-\alpha/2}\widehat{s.e.}(\widehat{\mu_{pop}})$. Puis, nous déduisons l'intervalle de confiance de niveau $1-\alpha$ pour ψ_{pop} :

$$CI_{1-\alpha}(\psi_{pop}) = \left[h^{-1}(\mu_{pop,\alpha/2}), h^{-1}(\mu_{pop,1-\alpha/2})\right] \tag{1.30}$$

Nous terminerons en remarquant que plus la s.e. d'un paramètre est grande par rapport à ce dernier, moins l'estimation du paramètre est fiable. Ainsi, MONOLIX estime aussi l'erreur standard relative (r.s.e.) par la s.e. divisée par la valeur du paramètre estimé.

1.3 Les algorithmes séquentiels pour l'estimation de paramètres

Avant de présenter une méthode d'assimilation séquentielles de données, nous reformulons l'écriture des modèles mécanistes. Par soucis de simplicité, nous adoptons les notations introduites dans l'article d'Annabelle COLLIN et al. [7]. Puis, nous nous concentrons sur la technique du filtrage de Kalman pour le cas des modèles linéaires.

1.3.1 Réécriture du modèle

Modèle déterministe

Supposons que nous avons une population de N individus que nous observons sur un intervalle de temps $[0;T]$. Pour tout $1 \leq i \leq N$, on note $x^i(t) \in \mathcal{X} \simeq \mathbb{R}^{N_x}$ le vecteur état qui contient toutes les variables dépendantes du temps, c'est-à-dire les variables décrites dans notre modèle mathématique. Pour tout $1 \leq i \leq N$, on note $\theta^i \in \mathcal{P} \simeq \mathbb{R}^{N_\theta}$ les paramètres du système qui sont fixés dans le temps. Nous modélisons les dynamiques par une fonction $f \in \mathcal{C}^1(\mathcal{X} \times \mathcal{P} \times [0;T], \mathcal{X})$ telle que pour tout $1 \leq i \leq N$:

$$\frac{d}{dt}x^i(t) = f(x^i(t), \theta^i, t) \tag{1.31}$$

La première différence avec l'approche variationnelle est que désormais nous considérerons les paramètres θ^i comme des variables d'état. Comme ces paramètres sont fixes dans le temps, nous aurons $\dot{\theta}^i = 0$. Ainsi, lorsque nous considèrerons l'approche séquentielle, nous noterons z^i l'état augmenté tel que pour tout $1 \leq i \leq N$:

$$z^i = \begin{pmatrix} x^i \\ \theta^i \end{pmatrix} \in \mathcal{Z} = \mathcal{X} \times \mathcal{P} \simeq \mathbb{R}^{N_z} \tag{1.32}$$

et les dynamiques par :

$$\begin{cases} \dfrac{d}{dt}z^i(t) &= a(z^i(t), t) \\ \\ z^i(0) &= \begin{pmatrix} x^i(0) \\ \theta^i \end{pmatrix} \end{cases} \tag{1.33}$$

où pour tout $1 \leq i \leq N$, $a(z^i(t), t) = \begin{pmatrix} f(x^i(t), \theta^i, t) \\ \mathbf{0}_{\mathbb{R}^{N_\theta}} \end{pmatrix}$

Dans la plupart des modèles, la résolution du système (1.33) se fait numériquement car la solution n'a pas de forme analytique. La solution numérique dépendra donc du schéma en temps utilisé. Pour simplifier, nous nous concentrons ici sur un schéma d'Euler explicite. Par la suite, nous verrons que ce schéma n'est pas adapté au modèle (3.7) que nous étudierons. Nous implémenterons d'autres schémas en temps dans le chapitre 3. En fonction du schéma en temps utilisé, il faudra donc changer l'expression (1.34) et celle de l'opérateur de transition $\psi_{n+1|n}$ défini ci-dessous.

Le schéma d'Euler explicite pour le système (1.33) s'écrit de la manière suivante :

$$z^i_{n+1} = z^i_n + \delta T_n a(z^i_n, T_n) \qquad \text{avec } 0 \leq n \leq N_T - 1 \tag{1.34}$$

où le pas $\delta T_n = T_{n+1} - T_n$ doit être suffisamment petit pour permettre la convergence vers la solution et $0 = T_0 \leq T_1 \leq ... \leq T_{N_T} = T$.

Ainsi, on peut définir l'opérateur de transition par $\psi_{n+1|n}(z^i_n) = z^i_n + \delta T_n a(z^i_n, T_n)$ de sorte que :

$$\begin{cases} z^i_{n+1} &= \psi_{n+1|n}(z^i_n), \quad 0 \leq n \leq N_T - 1, \quad 1 \leq i \leq N \\ \\ z^i_0 &= \begin{pmatrix} x^i(0) \\ \theta^i \end{pmatrix} \end{cases} \tag{1.35}$$

Modèle stochastique

Maintenant que la partie déterministe a été reformulée, nous pouvons introduire des effets stochastiques dans notre modèle. Lors de l'approche variationnelle, les effets aléatoires portent uniquement sur certains paramètres afin de modéliser la variabilité individuelle. Dans notre approche séquentielle, nous mettons à la fois des effets aléatoires sur les paramètres d'intérêts θ^i et sur les états x^i. Ceci fait un deuxième point de différence quant à la formulation du problème entre les deux approches étudiées.

Pour cela, quelque soit $0 \leq n \leq N_T - 1$ et $1 \leq i \leq N$, nous introduisons un opérateur $B_n \in \mathcal{L}(\mathbb{R}^{N_z}, \mathbb{R}^{N_z})$ et un vecteur aléatoire $\nu_n^i \sim \mathcal{N}(0, Q_n^i)$ indépendants les uns des autres de sorte que :

$$\left\{ \begin{array}{rcl} z_{n+1}^i & = & \psi_{n+1|n}(z_n^i) + B_n v_n^i, \quad 0 \leq n \leq N_T - 1, \quad 1 \leq i \leq N \\ z_0^i & = & z_0 + \xi^i \end{array} \right. \tag{1.36}$$

où ξ^i représente l'effet aléatoire gaussien sur les valeurs initiales de l'état augmenté et z_0 est déterministe. Nous pouvons alors décomposer ξ^i de la manière suivante : $\xi^i = \xi^{pop} + \widetilde{\xi^i}$ où $\widetilde{\xi^i} \sim \mathcal{N}(0, P_0)$ pour $1 \leq i \leq N$.

Modèle d'observation

En pratique, nous n'observons pas directement les états x^i mais nous avons des observations ou mesures échantillonnées dans le temps $y^i \in \mathcal{Y} \subset \mathbb{R}^{N_{obs}}$. Ces observations peuvent être retrouvées grâce aux états par le modèle d'observation. En effet, nous définissons des opérateurs d'observations h_k dépendant du temps $t_k \in [0, T]$ tel que :

$$y_k^i = h_k^i(x^i(t_k^i)) + \mathcal{X}_k^i, \quad 1 \leq i \leq N, \quad 0 \leq k \leq N_{T_{obs}} \tag{1.37}$$

où $\mathcal{X}_k^i \sim \mathcal{N}(0, W_k^i)$ représente les erreurs de mesures (ou erreur résiduelles, biais de mesure) et les temps de mesure t_k^i sont supposés identiques pour tous les individus.

Nous pouvons réécrire ce modèle d'observation en utilisant les notations des états augmentés par :

$$y_k^i = c_k^i(z^i(t_k^i)) + \mathcal{X}_k^i, \quad 1 \leq i \leq N, \quad 0 \leq k \leq N_{T_{obs}} \tag{1.38}$$

où

$$\begin{array}{rcl} c_k^i : \mathcal{X} \times \mathcal{P} & \longrightarrow & \mathcal{Y} \\ z = (x, \theta) & \longmapsto & h_k^i(x) \end{array} \tag{1.39}$$

Approche populationnelle

Comme nous souhaitons avoir un point de vue populationnel pour nos estimations, nous allons ajouter quelques notations supplémentaires.

On regroupe les effets aléatoires des valeurs initiales de l'état augmenté $\boldsymbol{\xi} = (\xi^1, \ldots, \xi^N)^T \in (\mathcal{X} \times \mathcal{P})^N$ et les effets aléatoires $\boldsymbol{\nu_n} = (\nu_n^1, \ldots, \nu_n^N)^T$ pour tout $0 \leq n \leq N_T - 1$ qui ont pour matrice de covariance :

$$\boldsymbol{Q}_n = \begin{pmatrix} Q_n^1 & & 0 \\ & \ddots & \\ 0 & & Q_n^N \end{pmatrix} \tag{1.40}$$

On fait de même avec les états individuels $\boldsymbol{z} = (z^1, \ldots, z^N)^T$, ce qui permet de réécrire le modèle dynamique de la manière suivante :

$$\left\{ \begin{array}{rcl} \boldsymbol{z}_{n+1} & = & \boldsymbol{\psi}_{n+1|n}(\boldsymbol{z}_n) + \boldsymbol{B}_n \boldsymbol{\nu}_n \\ \boldsymbol{z}_0 & = & 1_N \otimes z_0 + \boldsymbol{\xi} \end{array} \right. \tag{1.41}$$

où $\otimes$ représente le produit de Kronecker, $1_N = (1, \ldots, 1)^T \in \mathbb{R}^N$ et

$$\boldsymbol{B}_n = Id_N \otimes B_n \quad \text{et} \quad \boldsymbol{\psi}_{n+1|n} = (\psi_{n+1|n}(z_n^1), \ldots, \psi_{n+1|n}(z_n^N))^T \tag{1.42}$$

Les observations sont aussi regroupées en $\boldsymbol{y}_k = (y_k^1, \ldots, y_k^N)^T$ de sorte que pour tout $1 \leq k \leq N_{T_{obs}}$:

$$\boldsymbol{y}_k = \boldsymbol{c}_k(\boldsymbol{z}_{n_k}) + \mathcal{X}_k \tag{1.43}$$

où $\boldsymbol{c}_k(\boldsymbol{z}_{n_k}) = (\boldsymbol{c}_k^1(\boldsymbol{z}_{n_k}^1), \ldots, \boldsymbol{c}_k^N(\boldsymbol{z}_{n_k}^N))^T$ et les erreurs résiduelles (ou erreurs de mesures) $\mathcal{X}_k$ ont la matrice de covariance :

$$\boldsymbol{W}_k = \begin{pmatrix} W_k^1 & & 0 \\ & \ddots & \\ 0 & & W_k^N \end{pmatrix} \tag{1.44}$$

1.3.2 Estimation des paramètres par les filtres de Kalman

Les filtres de Kalman consistent à fournir à chaque nouvelle observation une nouvelle estimation de la variable d'état. Cette estimation se fait en deux phases distinctes appelées *prédiction* et *mise à jour*. La phase de prédiction utilise l'état estimé de l'instant précédent pour produire une estimation de l'état courant. Pour la mise à jour, les observations de l'instant courant sont utilisées pour corriger l'état prédit dans le but d'obtenir une estimation plus précise.

Nous nous concentrons aux cas où les modèles déterministes et d'observations sont des fonctions linéaires de l'état (*i.e.* quelque soit n, $\boldsymbol{\psi}_{n+1|n}$, $\boldsymbol{c}_n$ et $\boldsymbol{B}_n$ sont linéaires) [2]. Ainsi, dans la suite de cette section, nous ne ferons pas de distinction entre les applications linéaire $\boldsymbol{\psi}_{n+1|n}$, $\boldsymbol{c}_n$ et leur matrices associées dans la base canonique.

Après avoir initialisé $\boldsymbol{z}_0^f$ et P_0^f, nous effectuons les étapes de mise à jour et de prédiction aux instants des observations $\{t_0, ..., t_{N_{T_{obs}}}\}$.

Phase de mise à jour : pour tout $0 \leq n \leq N_{T_{obs}}$

A partir du vecteur d'innovation $\boldsymbol{y}_n - \boldsymbol{c}_n \boldsymbol{z}_n^a$, nous construisons l'état analysé par l'expression :

$$\boldsymbol{z}_n^a = \boldsymbol{z}_n^f + K_n(\boldsymbol{y}_n - \boldsymbol{c}_n \boldsymbol{z}_n^a) \tag{1.45}$$

où K_n est la matrice de gain, dite matrice de Kalman. Elle est défini par :

$$K_n = P_n^f \boldsymbol{c}_n^T \left(\boldsymbol{c}_n P_n^f \boldsymbol{c}_n + W_n^T \right)^{-1} \tag{1.46}$$

Nous obtenons ensuite la matrice de covariance de l'erreur d'analyse $z_n^a - z_n$ par l'expression suivante :

$$P_n^a = P_n^f - K_n \boldsymbol{c}_n P_n^f \tag{1.47}$$

Phase de prédiction : pour tout $0 \leq n \leq N_{T_{obs}} - 1$

Nous pouvons ensuite construire une ébauche de l'état du système au prochain instant de mesure t_{n+1} par l'expression : $z^f_{n+1} = \boldsymbol{\psi}_{n+1|n} z^a_n$, où z^a_n est l'état analysé au temps t_n et $\boldsymbol{\psi}_{n+1|n}$ est l'opérateur de transition entre les instants d'observation t_n et t_{n+1}.

Si nous notons P^f_n la matrice de covariance de l'erreur d'ébauche $z^f_n - z_n$, nous avons la formule :

$$P^f_{n+1} = \boldsymbol{\psi}_{n+1|n} P^a_n \boldsymbol{\psi}^T_{n+1|n} + \boldsymbol{B}_n \boldsymbol{Q}_n \boldsymbol{B}^T_n \tag{1.48}$$

Comme nous le verrons dans le chapitre 2, les modèles que nous considérerons ne seront pas linéaires. Dans ce cas, cette version des filtres de Kalman n'est pas recommandée. Il est préférable d'utiliser les *Extended Kalman Filter* ou les *Unscented Kalman Filter*. Pour plus détails à propos de l'implémentation des *Unscented Kalman Filter* et de la gestion de la non linéarité du modèle, le lecteur pourra se référer aux travaux d'Annabelle Collin et al. [7]. Lors de nos travaux sur l'assimilation séquentielle de données, nous utiliserons une des versions détaillée dans l'article précédemment cité.

L'assimilation variationnelle de données est une méthode largement utilisée dans la littérature pour l'étude de modèles pharmacocinétiques et pharmacodynamiques. Or, cette approche ne permet pas de prendre en compte une éventuelle évolution des paramètres du modèle au cours du temps, ce qui pourrait être intéressant dans certains modèles. Comme nous avons pu le voir dans la section 1.3, le fait de considérer l'état augmenté permet d'intégrer cette évolution. Mais si nous savons que l'assimilation séquentielle de données est efficace et largement utilisée lorsque les instants de mesures sont très proches (météorologie, pilotage automatique, location et cartographie simultanées), nous n'en connaissons pas les performances lorsque les données sont parcimonieuses. Ainsi, une des ambitions théorique ce travail est de comparer les deux approches dans le cas de données parcimonieuses.

Chapitre 2

Modèles et stratégies de modélisation

Dans ce chapitre nous cherchons à modéliser l'évolution de la charge virale et de la réponse immunitaire au cours du temps face au *SARS-CoV-2* chez des singes.

Dans un premier temps, nous présenterons brièvement la maladie, ses principaux mécanismes d'action et le fonctionnement du vaccin que nous étudions. Nous rappellerons les principaux résultats d'un article présentant un premier modèle mécaniste sur ces données [1]. Nous nous appuierons sur ce modèle tout au long du chapitre et nous l'appellerons *modèle de référence*.

Dans un second temps, nous expliciterons les trois modèles mécanistes que nous étudierons par la suite. Dans le premier modèle, nous transformerons le modèle d'observation du *modèle de référence*. Puis, nous mettrons en avant deux modèles permettant la modélisation conjointe de la charge virale et de la réponse immunitaire.

Pour terminer, nous expliquerons la méthode du profil de vraisemblance et l'appliquerons à un des modèles défini préalablement. Cette méthode est fréquemment utilisée lorsque l'information contenue dans les données n'est pas suffisante pour permettre une estimation de certains paramètres. Nous présenterons certains graphiques pouvant nous indiquer les modifications à faire pour améliorer le modèle, si possible. Nous verrons aussi les critères nous permettant d'affirmer que notre modélisation est satisfaisante – ou plutôt que les données que nous possédons ne nous permettent pas de rejeter notre modélisation.

2.1 Le cas d'étude

Le *SARS-CoV-2* (*severe acute respiratory syndrome coronavirus 2*) est un virus apparu en décembre 2019 dans la ville de Wuhan, en Chine. Face à la rapide propagation et virulence du virus, les dirigeants des pays ont mis en place des restrictions importantes pour protéger la population. Des équipes de recherche du monde entier ont étudié ce virus dans l'objectif de trouver un traitement curatif contre cet agent pathogène.

Nous allons étudier ici l'effet d'infections antérieures et du vaccin αCD40.RBD sur la charge virale, la quantité d'anticorps et leur fonction neutralisante face au virus. Permettons-nous de faire de succincts rappels d'immunologie et d'expliquer les principaux mécanismes d'actions du *SARS-CoV-2*.

2.1.1 SARS-CoV-2 et bases d'immunologie

Le SARS-CoV-2 possède les protéines structurales E, M, N et S. Les protéines S, E et M forment, ensemble, son enveloppe virale. La protéine S, ou plus connue sous le nom de protéine Spike, est constituée de deux sous-unités dont l'une d'elles contient le *receptor binding domain* (RBD). Le domaine RBD permet la liaison à l'enzyme de conversion de l'angiotensine 2 (ACE2)

des cellules susceptibles. Une fois lié à une cellule, le virus pénètre à l'intérieur de celle-ci par un enchainement de processus chimiques. On dit alors que la cellule est infectée. Le virus utilise les protéines présentes au sein de cette cellule afin de se répliquer. Au bout d'un certain temps, la cellule infectée produira des virions qui peuvent être infectieux ou non. Une fois relâchés dans le milieux extra-cellulaire, les virions infectieux infecteront d'autres cellules susceptibles et reproduiront le même schéma.

Face aux infections virales, le corps humain met en place principalement deux réactions immunitaires, la réponse innée et la réponse adaptative.

La réponse immunitaire innée a lieu dès qu'un agent pathogène est repéré dans le corps. Elle se met en place très rapidement et sans avoir déjà été exposé à ce pathogène. Cependant, lorsque cette réaction n'est pas suffisante pour éliminer l'infection, les cellules présentatrices d'antigène vont enclencher la réponse immunitaire adaptative en activant les lymphocytes T et B dans les organes lymphoïdes. C'est donc un réponse tardive qui est accentuée si l'individu a déjà rencontré l'agent pathogène. En effet, à chaque fois que cette réponse se met en place, les lymphocytes B et T correspondant au virus se multiplient et se différencient. Certains, appelés lymphocytes mémoires, sont conservés et permettront une réponse plus virulente lors d'une prochaine infection par ce virus.

Dans ce rapport, nous nous intéressons au vaccin αCD40.RBD. L'objectif de celui-ci est de présenter le domaine RBD de la protéine Spike aux récepteurs CD40 des cellules dendritiques, qui sont des cellules présentatrices d'antigènes. Elles rejoindront les organes lymphoïdes et présenteront le domaine aux lymphocytes B et T. Ces cellules vont se différencier afin de produire des anticorps et des cellules CD8. Les anticorps auront principalement deux modes d'actions. Les immunoglobulines G iront se fixer sur les cellules productrices de virions et les complexes seront détruits par les cellules NK ou les macrophages. D'autres anticorps iront se fixer sur les virions libres dans le milieu extra-cellulaire pour bloquer la liaison RBD-ACE2.

2.1.2 Les données

Nous étudions trois groupes de six macaques crabier que nous exposons à une forte dose (1×10^6 Unité de Formation de Plaque (pfu) soit 2.19×10^{10} virions) de *SARS-CoV-2* administrée simultanément par voie intra-nasale et intra-trachéale [16]. Deux groupes de macaques, soit douze macaques, sont infectés une première fois. 172 jours après, parmi ces macaques, six sont choisis aléatoirement pour recevoir un placebo, les autres reçoivent une injection de vaccin αCD40.RBD. Quatre semaines après la vaccination, six macaques n'ayant jamais été vacciné ou infecté rejoignent l'étude et les dix-huit macaques sont infectés.

Parmi les différentes mesures qui ont été effectuées tout au long de l'étude, nous nous servons principalement des données d'Acides RiboNucléiques (ARN) génomique et subgénomique, d'IgG anti-RBD et d'anticorps neutralisants (voir travaux de Marie Alexandre et al. [1]).

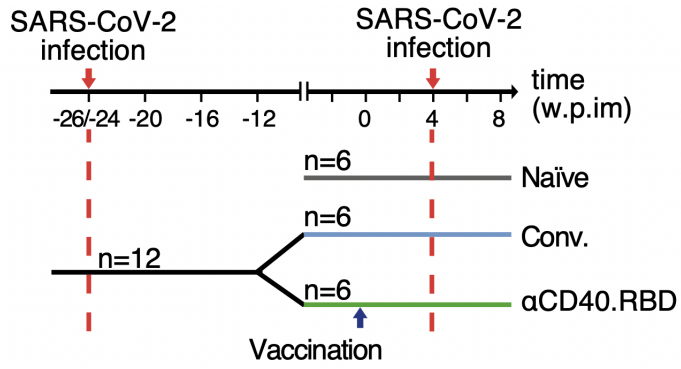
2.1.3 Le *modèle de référence*

L'objectif des travaux de Marie Alexandre et al. [1] était de construire un modèle afin de mieux caractériser l'impact de la réponse immunitaire sur les dynamiques de l'ARN génomique et l'ARN subgénomique viral. Pour faire cela, il a été étudié le modèle mécaniste constitué des trois parties habituelles :

— un modèle mathématique constitué d'un système d'EDO pour décrire les dynamiques dans les compartiments nasal et trachéal.

— un modèle statistique pour prendre en compte la variabilité inter-individuelle et les effets des covariables sur les paramètres.

FIGURE 2.1 – Design de l'étude.



— un modèle d'observation pour décrire le $\log_{10}$ de la charge virale observée dans les deux compartiments.

Le modèle mathématique est représenté par le système suivant où $X \in \{N = \text{Nasopharynx}, T = \text{Trachée}\}$ représente le compartiment en question :

$$\begin{cases} \dot{T}^X &= -\beta T^X V_i^X - \mu\beta T^X V_S^X \\ \dot{I_1}^X &= \beta T^X V_i^X + \mu\beta T^X V_S^X - kI_1^X \\ \dot{I_2}^X &= kI_1^X - \delta I_2^X \\ \dot{V_i}^X &= \mu P^X I_2^X - cV_i^X - \beta T^X V_i^X \\ \dot{V_{ni}}^X &= (1-\mu)P^X I_2^X - cV_{ni}^X \\ \dot{V_S}^X &= -\mu\beta T^X V_S^X - c_I V_S^X \end{cases} \tag{2.1}$$

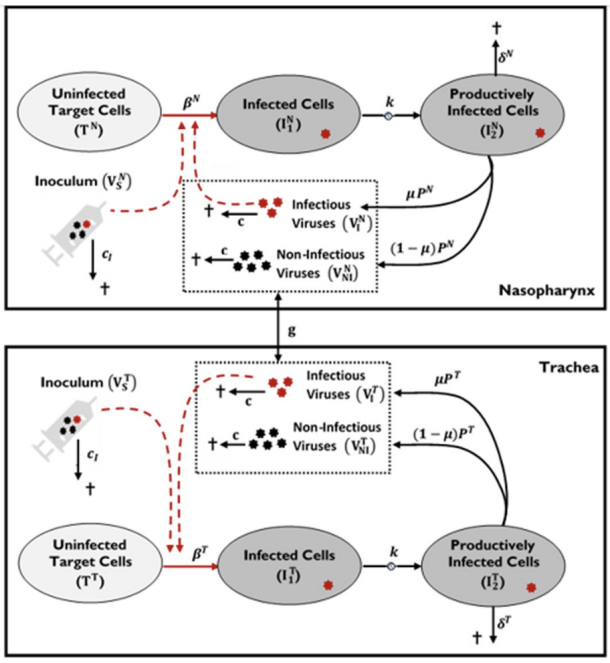
Les conditions initiales sont $I_1^X(0) = I_2^X(0) = V_i^X(0) = V_{ni}^X(0) = 0$. $V_S^X(0)$, la concentration inoculée dans chacun des deux compartiments est définie par : $V_S^T(0) = \dfrac{0.9 \times Inoc_0}{W^T}$ et $V_S^N(0) = \dfrac{0.1 \times Inoc_0}{W^N}$ où $Inoc_0 = 2.19 \times 10^{10}$ représente le nombre de virions inoculés et W^X le volume de cellules dans chacun des compartiments. Quant à la valeur initiale du nombre de cellules susceptibles, nous avons $T^T(0) = \dfrac{2.25 \times 10^4}{W^T}$ et $T^N(0) = \dfrac{1.25 \times 10^5}{W^N}$. Une précédente étude a permis d'approximer W^N et W^T en fonction du poids du macaque, qui est mesuré pour chacun des 18 macaques suivis [1] :

$$\begin{aligned} W_i^N &= \begin{cases} 4 & \text{si poids}_i \leq 4.5kg \\ 5.5 & \text{sinon} \end{cases} \\ W_i^T &= \begin{cases} 2 & \text{si poids}_i \leq 4.5kg \\ 3 & \text{sinon} \end{cases} \end{aligned} \tag{2.2}$$

Nous utiliserons ces conditions initiales pour les modèles suivants. La FIGURE 2.2 est une représentation schématique de ce modèle mathématique.

Pour chacun des deux compartiments, nasal et trachéal, le modèle contient des cellules susceptibles (T) qui peuvent être infectées (I_1) par le virus infectieux (V_i) ou le virus inoculé (V_S)

FIGURE 2.2 – Représentation schématique du modèle (2.1) avec $g = 0$. [1]



avec un taux d'infectivité β. Après une phase d'éclipse, les cellules infectées deviennent productrices de virions (I_2) à un taux P^X et peuvent être détruites à un taux δ. Une proportion μ du virus généré par les cellules est infectieux (V_i) tandis que la proportion $1-\mu$ est non infectieux (V_{ni}). Le virus produit par les cellules est éliminé à un taux c alors que le virus inoculé est éliminé à un taux c_I.

Le modèle statistique qui va nous intéresser dans la suite est donné par :

$$\begin{cases} \log_{10}(\beta_i) &= \log_{10}(\beta_{pop}) + u_i^\beta \\ \log(\delta_i) &= \log(\delta_{pop}) + u_i^\delta \end{cases} \tag{2.3}$$

où $u_i^\beta \sim \mathcal{N}(0, \omega_\beta^2)$ et $u_i^\delta \sim \mathcal{N}(0, \omega_\delta^2)$.
Dans toute la suite de ce rapport, au lieu d'estimer P^N et P^T, nous écrirons $P^T = P^N \exp(f_P^T)$ de manière à estimer les paramètres P^N et f_P^T. Similairement, nous estimons $\beta_i^{pow} := log_{10}(\beta_i)$ et $\beta_{pop}^{pow} := log_{10}(\beta_{pop}^{pow})$.

Pour terminer, nous avons le modèle d'observation défini pour chacun des deux compartiment – nasopharynx et trachée – de la manière suivante :

$$\begin{cases} gRNA_{ij}^X &= \log_{10}(V_i^X + V_{ni}^X + V_S^X)(t_{ij}, \psi_i^X) + \epsilon_{ij,g}^X \\ sgRNA_{ij}^X &= \alpha_{sgRNA} \log_{10}(I_1^X + I_2^X)(t_{ij}, \psi_i^X) + \epsilon_{ij,sg}^X \end{cases} \tag{2.4}$$

où $\epsilon_{ij,g}^X \sim \mathcal{N}(0, \sigma_{g,X}^2)$, $\epsilon_{ij,sg}^X \sim \mathcal{N}(0, \sigma_{sg,X}^2)$ et ψ_i^X représente les paramètres de l'individu i dans le compartiment X. Dans la suite de ce rapport, nous appellerons *modèle de référence* le modèle défini par les équations (2.1), (2.3) et (2.4).

2.2 Présentation des modèles

Dans cette section, nous définissons les quatre modèles qui nous intéresseront au cours de ce rapport. Le modèle de la section 2.2.1 sera analysé à la fin du présent chapitre et dans la section 3.1.2. Les trois autres modèles seront étudiés dans le chapitre 4.

2.2.1 Vérification d'hypothèses sur le modèle d'observation

Lorsque nous construisons un modèle, nous pouvons utiliser des connaissances déjà établies ou faire des hypothèses biologiques et/ou mathématiques. Dans le *modèle de référence*, le fait que l'ARN subgénomique dépende de $I_1 + I_2$ ou seulement de I_2 n'est pas encore établi. Ainsi, nous faisons face au problème inverse consistant à trouver le meilleur modèle mécaniste pour l'étude de la dynamique de la charge virale associée à nos données.

Pour cela, nous reprenons le *modèle de référence* défini dans la section 2.1.3 et modifions seulement le modèle d'observation par :

$$\begin{cases} gRNA_{ij}^X &= \log_{10}(V_i^X + V_{ni}^X + V_S^X)(t_{ij}, \psi_i^X) + \epsilon_{ij,g}^X \\ sgRNA_{ij}^X &= \alpha_{sgRNA} \log_{10}(I_2^X)(t_{ij}, \psi_i^X) + \epsilon_{ij,sg}^X \end{cases} \tag{2.5}$$

où $\epsilon_{ij,g}^X \sim \mathcal{N}(0, \sigma_{g,X}^2)$, $\epsilon_{ij,sg}^X \sim \mathcal{N}(0, \sigma_{sg,X}^2)$ et ψ_i^X représente le vecteur de paramètres de l'individu i dans le compartiment X. De cette manière, l'ARN subgénomique dépend uniquement de la quantité de cellules productrices de virions (I_2).

2.2.2 Modélisation conjointe : modèle 1

Après avoir montré que l'ajout de covariable de groupe dans le modèle statistique permettait d'améliorer les estimations des paramètres et la modélisation des dynamiques virales, Marie Alexandre et al. [1] ont mis en lumière deux marqueurs comme corrélat de protection. Les deux marqueurs sont le *RBD-ACE2 binding inhibition* qui quantifie la neutralisation des liaisons entre les virus et cellules susceptibles par les anticorps, et la concentration d'anticorps IgG anti-RBD qui accélèrent la destruction des cellules productrices de virus (I_2).

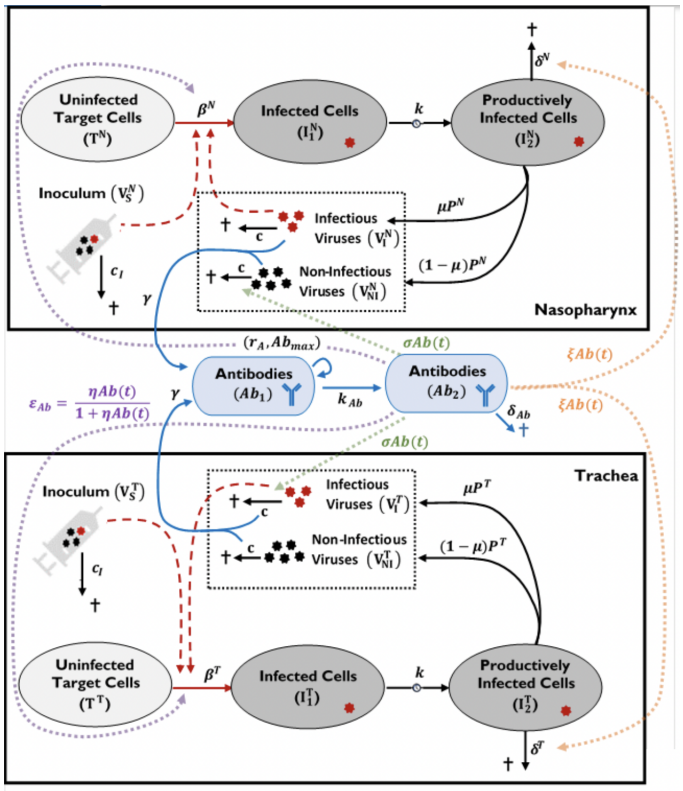
Les auteurs ont conclu que ces deux marqueurs ont un impact plus important sur les paramètres β et δ que les covariables de groupe. Naturellement, l'étape suivante est d'intégrer la réponse immunitaire au *modèle de référence* en ajoutant des compartiments et des équations au modèle mathématique. Ceci nous permettra de modéliser conjointement la charge virale et la réponse immunitaire humorale.

Dans cette continuité, un premier modèle mathématique a pu être défini par :

$$\begin{cases} \dot{T}^X &= -(1-\varepsilon_{Ab})\beta V_i^X T^X - \mu(1-\varepsilon_{Ab})\beta V_S^X T^X \\ \dot{I_1}^X &= (1-\varepsilon_{Ab})\beta V_i^X T^X + \mu(1-\varepsilon_{Ab})\beta V_S^X T^X - kI_1^X \\ \dot{I_2}^X &= kI_1^X - \delta I_2^X - \xi Ab_2 I_2^X \\ \dot{V_i}^X &= P^X \mu I_2^X - cV_i^X - (1-\varepsilon_{Ab})\beta V_i^X T^X \\ \dot{V_{ni}}^X &= P^X(1-\mu)I_2^X - cV_{ni}^X \\ \dot{V_S}^X &= -c_I V_S^X - \mu(1-\varepsilon_{Ab})\beta V_S^X T^X \\ \dot{Ab}_1 &= \gamma V + r_A\left(1 - \frac{Ab_1}{Ab_{max}}\right)Ab_1 - k_{Ab}Ab_1 \\ \dot{Ab}_2 &= k_{Ab}Ab_1 - \xi Ab_2 I_2 - \delta_{Ab}Ab_2 \end{cases} \tag{2.6}$$

où $I_2 = I_2^T + I_2^N$, $V = V_i^T + V_{ni}^T + V_i^N + V_{ni}^N$ et $\varepsilon_{Ab} = \dfrac{\eta Ab_2}{1+\eta Ab_2}$ avec les valeurs initiales supplémentaires $Ab_1(0) = 0$ et $Ab_2(0)$ la quantité d'anticorps mesurée au début de l'étude. Nous pouvons aussi représenter ce système par la FIGURE 2.3 où $\sigma = 0$.

FIGURE 2.3 – Représentation schématique du modèle (2.6) avec $\sigma = 0$.



Le modèle statistique est défini par :

$$\begin{cases} \log_{10}(\beta_i) &= \log_{10}(\beta_{pop}) + u_i^{\beta} \\ \log(\delta_i) &= \log(\delta_{pop}) + u_i^{\delta} \\ \log(r_{Ab_i}) &= \log(r_{Ab_{pop}}) + \phi_{conv}^{r_{Ab}} \mathbb{1}_{group=conv} + \phi_{CD40}^{r_{Ab}} \mathbb{1}_{group=CD40} + u_i^{r_{Ab}} \end{cases} \tag{2.7}$$

où $u_i^{\beta} \sim \mathcal{N}(0, \omega_{\beta}^2)$, $u_i^{\delta} \sim \mathcal{N}(0, \omega_{\delta}^2)$ et $u_i^{r_{Ab}} \sim \mathcal{N}(0, \omega_{r_{Ab}}^2)$. Puis, nous avons le modèle d'observation suivant :

$$\begin{cases} gRNA_{ij}^X &= \log_{10}(V_i^X + V_{ni}^X + V_S^X)(t_{ij}, \psi_i) + \epsilon_{ij,g}^X \\ sgRNA_{ij}^X &= \alpha_{sgRNA} \log_{10}(I_1^X + I_2^X)(t_{ij}, \psi_i) + \epsilon_{ij,sg}^X \\ IgG\ RBD_{ij} &= log_{10}(Ab_2)(t_{ij}, \psi_i) + \epsilon_{ij,ab} \\ ECLRBD_{ij} &= log_{10}(A \times \left(1 - \frac{\eta Ab_2^n}{1 + \eta Ab_2^n}\right) + B)(t_{ij}, \psi_i) + \epsilon_{ij,ecl} \end{cases} \tag{2.8}$$

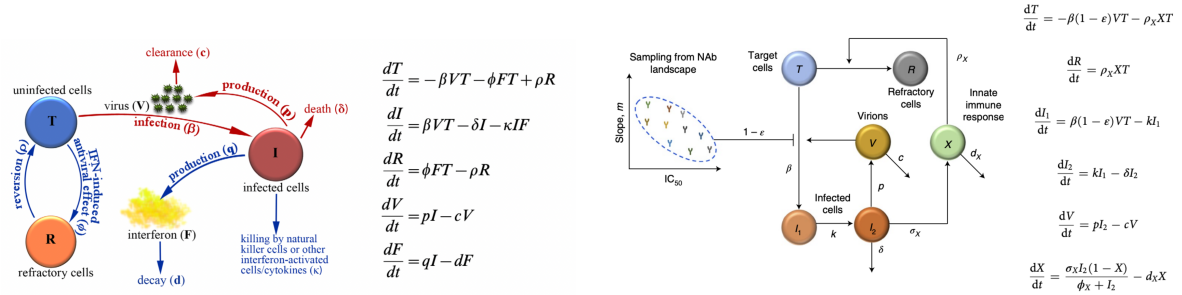
où $\epsilon_{ij,g}^X \sim \mathcal{N}(0, \sigma_{g,X}^2)$, $\epsilon_{ij,sg}^X \sim \mathcal{N}(0, \sigma_{sg,X}^2)$, $\epsilon_{ij,ab} \sim \mathcal{N}(0, \sigma_{ab}^2)$, $\epsilon_{ij,ecl} \sim \mathcal{N}(0, \sigma_{ecl}^2)$ et ψ_i représente les paramètres de l'individu i.

2.2.3 Modélisation conjointe : modèle 2

De nouveau, nous cherchons à résoudre le problème inverse pour déterminer le meilleur modèle mécaniste associé aux dynamiques que nous cherchons à modéliser. Ainsi, après avoir établi un état de l'art des modélisations de charge virale et/ou de réactions immunitaires pour différentes infections (grippe, *VIH*, *SARS-CoV-2*) [8, 19, 20], nous avons pu mettre en avant plusieurs modèles qui suscitent l'intérêt de la communauté scientifique.

Comme le montre la FIGURE 2.4, certains modèles utilisent un compartiment de cellules réfractaires représentant les cellules susceptibles ne pouvant être infectées. Cependant, pour des raisons d'identifiabilité biologique, nous n'avons pas considéré ces modèles. En effet, les anticorps neutralisants ne viennent pas protéger les cellules susceptibles d'une éventuelle infection mais viennent plutôt se fixer sur les virions libres pour bloquer leur entrée dans les cellules.

FIGURE 2.4 – Modèles intégrant un compartiment de cellules réfractaires.



(a) Voir Kasia A. PAWELEK et al. [19].

(b) Voir Pranesh PADMANABHAN et al. [18].

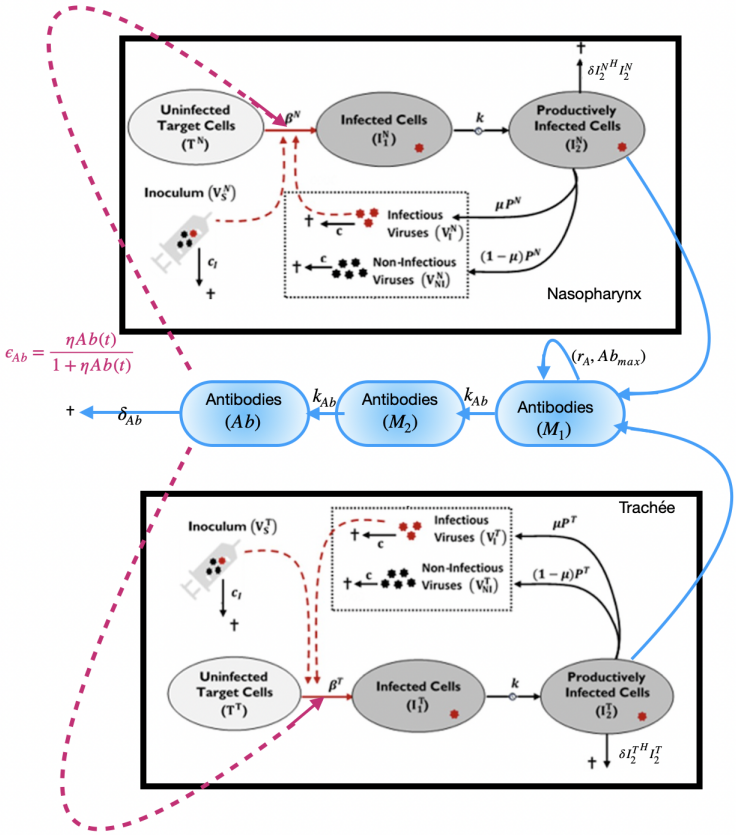
Nous avons défini un modèle, inspiré de travaux très novateurs dans la modélisation de la réponse immunitaire face au *SARS-CoV-2*, permettant de prendre en compte les deux phases de la réaction immunitaire [12, 20]. La réaction innée sera prise en compte par le terme dépendant de la quantité de cellules productrices de virus $\delta(I_2^X)^H$ et agira directement sur la destruction de ces cellule. Quant à la réaction adaptative, elle sera prise en compte par un terme à saturation dépendant de la quantité d'anticorps et permettra de réduire l'infection de nouvelles cellules par les virus libres.

Nous définissons le modèle mathématique de la manière suivante :

$$\begin{cases} \dot{T}^X &= -(1-\varepsilon_{Ab})\beta {V_i}^X T^X - \mu(1-\varepsilon_{Ab})\beta V_S^X T^X \\ \dot{I_1}^X &= (1-\varepsilon_{Ab})\beta {V_i}^X T^X + \mu(1-\varepsilon_{Ab})\beta V_S^X T^X - kI_1^X \\ \dot{I_2}^X &= kI_1^X - \delta(I_2^X)^H I_2^X \\ \dot{V_i}^X &= P^X \mu I_2^X - c{V_i}^X - (1-\varepsilon_{Ab})\beta {V_i}^X T^X \\ \dot{V_{ni}}^X &= P^X(1-\mu)I_2^X - cV_{ni}^X \\ \dot{V_S}^X &= -c_I V_S^X - \mu(1-\varepsilon_{Ab})\beta V_S^X T^X \\ \dot{M_1} &= \gamma I_2 + r_A\left(1 - \frac{M_1}{Ab_{max}}\right)M_1 - k_{Ab}M_1 \\ \dot{M_2} &= k_{Ab}M_1 - k_{Ab}M_2 \\ \dot{Ab} &= k_{Ab}M_2 - \delta_{Ab}Ab \end{cases} \tag{2.9}$$

où $I_2 = I_2^T + I_2^N$, $V = V_i^T + V_{ni}^T + V_i^N + V_{ni}^N$ et $\varepsilon_{Ab} = \dfrac{\eta Ab}{1+\eta Ab}$ avec les valeurs initiales supplémentaires $M_1(0) = M_2(0) = 0$ et $Ab(0)$ la quantité d'anticorps mesurée au début de l'étude. Nous pouvons aussi représenter ce système par la FIGURE 2.5.

FIGURE 2.5 – Représentation schématique du modèle 2.9



Puis nous définissons le modèle statistique et le modèle d'observation de la même manière que dans la section 2.2.2.

Les différences entre ces deux modèles sont :

— les anticorps du présent modèle agissent uniquement sur le taux d'infectivité β tandis que dans le modèle précédent ils agissent aussi sur la destruction des cellules I_2 avec le terme ξI_2 ;

— la réaction immunitaire du présent modèle est déclenchée par la présence de cellules infectées productrices de virus (I_2) tandis que dans le modèle précédent cette réaction est déclenchée par la présence de virus produit (V).

Nous étudierons l'impact de ces différences sur nos modélisations dans le chapitre 4.

2.3 Sélection de modèle

Une fois que nous avons estimé nos modèles, nous aimerions pouvoir les évaluer afin de savoir si nous avons une *bonne* estimation des paramètres et si nous arrivons à modéliser nos dynamiques observées. Nous aimerions aussi pouvoir déterminer quelles modifications pourraient améliorer notre modèle.

Dans un premier temps, il est possible que, quelque soit les modifications apportées à notre modèle d'observation ou notre modèle statistique, nous n'arrivons pas à améliorer significativement nos estimations. Cela peut être dû à un manque d'information dans nos données. Nous présentons une méthode permettant de résoudre ce problème lors de l'approche variationnelle.

Dans un second temps, nous utilisons des graphiques diagnostiques. Ils permettent de visualiser ce qu'il faudrait modifier pour améliorer notre modélisation (par exemple : ajout de covariable, ajout de corrélation entre les variables ou les observations, *et cetera*). Nous présentons les deux principaux graphiques que nous avons utilisés pour nous guider dans nos travaux.

Pour terminer, une fois que nous avons obtenu plusieurs modèles qui expliquent correctement le phénomène étudié, nous aimerions pouvoir les comparer entre eux grâce à des critères analytiques.

2.3.1 Profil de vraisemblance pour l'approche variationnelle

Lorsque nous souhaitons résoudre le problème inverse consistant à estimer les paramètres d'un modèle, il arrive fréquemment de rencontrer un problème d'identification ou un problème dû à un manque de données. Le premier se produit lorsque notre modèle n'est pas identifiable, c'est-à-dire que même avec une infinité d'observation nous n'arriverons pas à estimer les valeurs de certains paramètres. C'est un problème dû à la construction de notre modèle. Il existe des méthodes théoriques pour vérifier l'identifiabilité de notre modèle [4].

Durant ce stage, nous ne nous sommes pas concentrés sur ce problème théorique d'identification et nous supposons que nos modèles sont identifiables. Quand bien même, nous pouvons rencontrer le deuxième problème qui se produit lorsque nos données sont trop parcimonieuses ou que l'information contenue y est trop faible. Ici, nous détaillons une méthode pour résoudre ce problème en diminuant le nombre de paramètres à estimer.

Dans l'approche variationnelle, rappelons que notre objectif est de maximiser la vraisemblance $\mathcal{L}_y(\theta) = p(y;\theta)$. La méthode du profil de vraisemblance consiste à choisir un ensemble de valeurs potentielles pour les paramètres à fixer, à estimer le modèle pour toutes les combinaisons possibles et à calculer la vraisemblance. Puis, nous sélectionnons la combinaison menant à la valeur maximale de la vraisemblance. Il est important de remarquer qu'il est fort probable qu'il existe de meilleures combinaisons de paramètres que nous n'avons pas pu essayer.

Lorsque nous étudions des modèles non linéaires à effets mixtes, il n'existe pas d'expression analytique de la vraisemblance. Cependant, il existe deux méthodes algorithmiques implémentées

dans MONOLIX permettant une estimation de la vraisemblance. La première est plus rapide et moins précise, c'est la méthode de linéarisation. Nous avons déjà introduit cette méthode lors de l'estimation de la matrice d'information de Fisher dans la section 1.2.3. La seconde est une méthode de Monte Carlo par échantillonnage important. Elle doit être utilisée lors des dernières étapes de notre sélection de modèle ou lorsque deux modèles ont des scores, estimés par linéarisation, très proches. Pour plus d'informations sur ces deux algorithmes, le lecteur pourra se référer à l'ouvrage de Marc LAVIELLE [14].

2.3.2 Les graphiques diagnostiques

Dans cette section, nous présentons les principaux graphiques utilisés pour améliorer nos modèles. Ces graphiques permettent aussi de dire si les données nous permettent de rejeter ou non le modèle mais pas de l'accepter. En effet, il est probable qu'avec des données supplémentaires ou d'autres données, notre modèle ne soit pas aussi *bon* qu'il en ait l'air. Ainsi, les modèles que nous accepterons seront en réalité des modèles que nous ne rejetterons pas avec les données que nous avons à notre disposition.

Dans la suite du rapport, nous nous intéresserons principalement à deux types de graphiques utilisant les données observées : les *individual fits* et les *visual predictive checks*. Néanmoins, il existe d'autres graphiques représentant la distribution des paramètres individuels, la corrélation entre les paramètres individuels et les covariables possibles, la corrélation entre les effets aléatoires. Pour plus de détails sur ces graphiques, le lecteur pourra se référer à l'ouvrage de Marc LAVIELLE [14] et aux travaux de Thi Huyen Tram NGUYEN et al. [17].

Individual fits :
Dans le modèle défini par $y_{ij} = f(t_{ij};\psi_i)+a\epsilon_{ij}$, le fait d'estimer ψ_{pop} et ψ_i nous permet de tracer les courbes :

- $f(t;\widehat{\psi}_{pop})$ la dynamique estimée pour la population ;
- $f(t;\widehat{\psi}_i)$ la dynamique estimée pour l'individu i où $\widehat{\psi}_i$ est un estimateur de ψ_i les paramètres individuels de i.

On peut ajouter à ces deux courbes les observations y_{ij} avec $1 \leq j \leq n_i$. Ces graphiques nous permettent de voir si les trajectoires individuelles $f(t;\widehat{\psi}_i)$ modélisent correctement les données observées. Ils permettent aussi de visualiser comment la dynamique individuelle diffère de la population.

Cependant, lorsque le nombre d'individus étudié est grand, il peut être fastidieux d'analyser ce graphique pour chacun d'entre eux. Nous avons donc une alternative à ce problème qui nous permet d'affirmer si le modèle arrive à reproduire la dynamique centrale ainsi que la variabilité de nos données observées.

Visual Predictive Check (VPC) :
Ce type graphique permet de résumer en une seule figure la construction du modèle structurel et du modèle statistique. La construction d'un VPC se fait de la manière suivante (voir FIGURE 2.6) :

(a) Nous plaçons les observations $(y_{ij}, 1 \leq i \leq N, 1 \leq j \leq n_i)$ mesurées au temps (t_{ij}) dans un graphique.

(b) Les données sont regroupées par des intervalles de temps adjacents. Les quantiles empiriques sont calculés au sein de chaque intervalle. En général, nous utilisons le $10^{\text{ième}}$, le $50^{\text{ième}}$ et le $90^{\text{ième}}$ quantile.

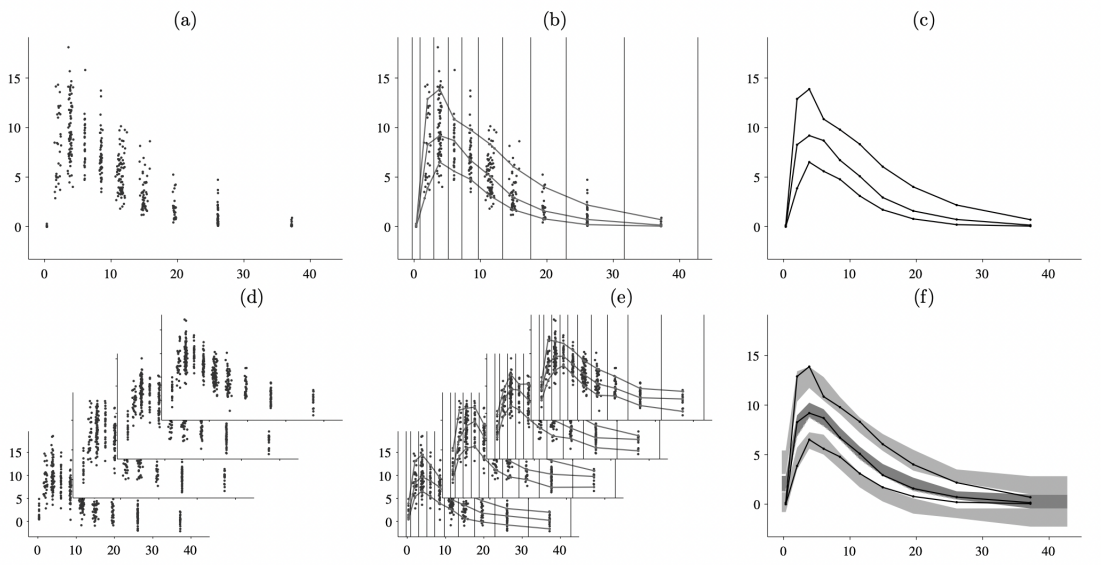
(c) Ces quantiles résument la distribution des observations. Nous enlevons les observations de l'étape (a).

(d) Nous simulons un grand nombre de données à partir du modèle à évaluer et dont les paramètres ont été estimés grâce aux observations à disposition.

(e) Les données simulées sont regroupées en utilisant les mêmes intervalles de temps de l'étape (b). Nous calculons aussi les mêmes quantiles dans chaque intervalle.

(f) Pour chaque quantile des données simulées, nous estimons l'intervalle de prédiction. Les quantiles observés de l'étape (b) sont comparés avec les intervalles de prédiction. En général, nous utilisons des intervalles de prédiction à 90%.

FIGURE 2.6 – Construction d'un *visual predictive check* [14].



Comme énoncé précédemment, il existe bien d'autres graphiques pouvant nous indiquer des modifications à faire dans notre modèle statistique ou notre modèle d'observation.

Lorsque nous obtenons des modélisations satisfaisantes grâce à plusieurs modèles, nous aimerions avoir des critères analytiques permettant de les comparer. Ces critères nous aideront à choisir le modèle le plus compatible avec le phénomène et les données étudiés.

2.3.3 Les critères d'information

Dans MONOLIX, il existe différents critères implémentés qui permettent d'évaluer les modèles entre eux. Rappelons que l'objectif de l'algorithme SAEM, qui permet d'estimer les paramètres de la population, est de maximiser la vraisemblance $\mathcal{L}_y(\theta) = p(y;\theta)$. Cela revient à minimiser $-2\mathcal{LL}_y(\theta) = -2\log(p(y;\theta))$. Ainsi, lors de la comparaison de modèles, il est important de vérifier quel modèle apporte une meilleure maximisation de la vraisemblance.

Comme nous pouvons être amené à comparer des modèles avec un nombre de paramètres ou d'individus différents, il parait judicieux que les critères que nous allons utiliser en tienne compte. En effet, avoir plus d'individus dans notre étude apporte plus d'information et peut conduire à une meilleure estimation des paramètres. De la même manière, avoir plus de paramètres dans notre modèle peut amener à plus de précision dans la modélisation, notamment lorsque nous avons l'information suffisante pour les estimer.
Nous définissons alors le critère d'information d'Akaike (AIC) (*Akaike information criterion*) et le critère d'information Bayésien (BIC) (*Bayesian information criterion*) par :

$$AIC = -2\mathcal{LL}_y(\theta) + 2P \tag{2.10}$$

$$BIC = -2\mathcal{LL}_y(\theta) + \log(N)P \tag{2.11}$$

où P est le nombre total de paramètres à estimer et N le nombre d'individus. Récemment, le critère d'information Bayésien corrigé (BICc) a été défini par :

$$BICc = -2\mathcal{LL}_y(\theta) + \dim(\theta_R)\log(N) + \dim(\theta_F)\log(n_{tot}) \tag{2.12}$$

où θ_R représente les paramètres de la partie aléatoire du modèle, θ_F représente les paramètres de la partie fixe du modèle et n_{tot} le nombre d'observation. Dans toute la suite de ce rapport, nous comparerons les modèles en fonction du BICc car son terme de pénalité permet de sélectionner des modèles plus parcimonieux en terme de nombre de paramètres au regard du nombre d'observation dans le problème étudié.

2.3.4 Applications des méthodes de profil de vraisemblance et de sélection de modèles

Pour fixer les paramètres k, c et c_I du modèle défini dans la section 2.1.3, la méthode du profil de vraisemblance a été utilisée (voir Marie ALEXANDRE et al. [1]). Comme nous souhaitons comparer les deux modèles d'observations (2.4) et (2.5) sous le même modèle statistique (2.3), nous allons appliquer un profil de vraisemblance pour fixer les paramètres k, c et c_I du modèle de la section 2.2.1.

De la même manière que dans l'article précédemment cité, nous allons appliquer deux profils de vraisemblance. Le premier nous permettra de fixer le paramètre k en supposant que $c = c_I$. Le deuxième nous permettra de fixer les deux paramètres restants c et c_I.

Premièrement, nous estimons le modèle pour toutes les combinaisons possibles avec $k \in \{1/3, 2/3, 1, 2, 3, 5, 6, 9\}$, $c \in \{1, 5, 10, 15, 20, 30\}$ et $c_I = c$. Nous minimisons le BICc lorsque $k = 3$ et $c = c_I = 20$. Dans la suite, nous fixons donc $k = 3$ et cherchons à optimiser les valeurs de c et c_I.

Deuxièmement, nous appliquons de nouveau la méthode pour toutes les combinaisons possibles avec $c, c_I \in \{1, 2, 3, 5, 7, 10, 12, 15, 17, 19, 21, 23, 25, 27, 29\}$ sans imposer $c = c_I$. Nous optimisons le BICc lorsque $c = 3$ et $c_I = 19$.

Ainsi, par la méthode du profil de vraisemblance, nous fixons $k = c = 3$ et $c_I = 19$. A titre comparatif, les paramètres obtenus dans les précédents travaux étaient $k = c = 3$ et $c_I = 20$ (voir Marie ALEXANDRE et al. [1]). Ce fait permet de confirmer la stabilité du profil de vraisemblance réalisé dans cet article.

Nous souhaitons maintenant comparer les estimations des modèles d'observation (2.4) et (2.5) sous le modèle statistique (2.3). Afin de garantir que la différence ne vient pas du paramètre c_I, nous fixerons dans la suite $k = c = 3$ et $c_I = 20$.

Dans un premier temps, en comparant les r.s.e. des deux modèles, nous pouvons remarquer que les estimations des paramètres sont fiables (voir FIGURES B.1 et B.2).

Dans un second temps, en comparant les VPC des quatre observations de ces deux modèles, nous pouvons affirmer qu'ils arrivent à modéliser la dynamique centrale et la variabilité individuelle des quantités étudiées (voir FIGURE2.7).

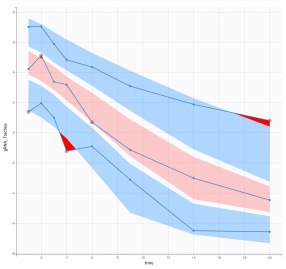
Pour terminer, le BICc obtenu avec le *modèle de référence* est égal à 794 tandis que le modèle défini section 2.2.1 obtient un BICc égal à 795.

Ainsi, les graphiques à disposition ou les estimations des paramètres et des critères d'information ne permettent pas d'affirmer que l'un des deux modèles est meilleur que l'autre. Il est donc nécessaire de mettre en place ce même procédé sur d'autres études pour répondre à ce problème inverse.

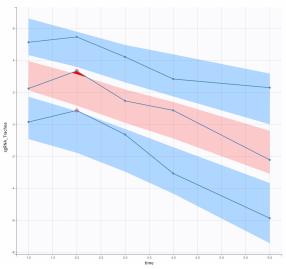
FIGURE 2.7 – VPC des modèles définis dans les sections 2.1.3 et 2.2.1

La première ligne représente les VPC du *modèle de référence*. La deuxième ligne représente les VPC du modèle de la section 2.2.1.

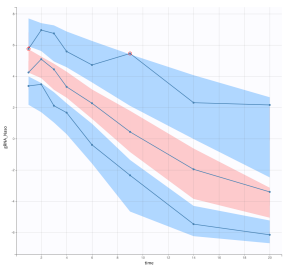
On utilise les abréviations suivantes ARN génomique (gRNA) et ARN subgénomique (sgRNA).



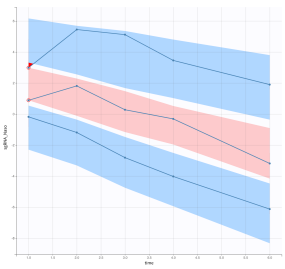
(a) gRNA trachée



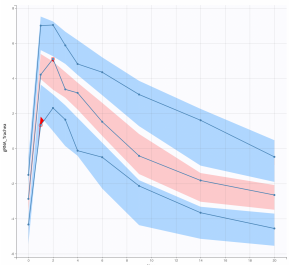
(b) sgRNA trachée



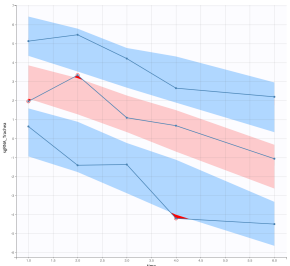
(c) gRNA nasal



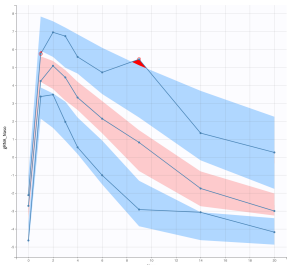
(d) sgRNA nasal



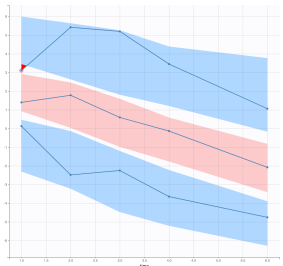
(e) gRNA trachée



(f) sgRNA trachée



(g) gRNA nasal



(h) sgRNA nasal

Chapitre 3

Problèmes théoriques pour la modélisation

Dans ce chapitre, nous présenterons les deux principaux problèmes théoriques auxquels nous nous sommes confrontés au cours de ce stage.

D'une part, les modèles que nous cherchons à estimer ont pour objectif de nous aider à comprendre et quantifier certains phénomènes biologiques (par exemple : le taux d'infection de cellules susceptibles, le taux de mortalité des cellules, le taux de production de cellules susceptibles, *et cetera*).

Comme nous avons pu l'entendre au début de l'épidémie de *SARS-CoV-2*, une quantité importante est le taux de reproduction du virus R_0. Les travaux de J. M. Heffernan, O. Diekmann et al. [10, 11, 13] ont permis d'établir une méthode algébrique, appelée *next generation matrix method* aboutissant à une formule du taux de reproduction R_0.

Dans la première section de ce chapitre, nous expliquerons et mettrons en application cette méthode pour le modèle défini section 2.2.1. Pour conclure, nous montrerons que le groupe d'appartenance du singe (naïf, convalescent et convalescent-vacciné) agit significativement sur la valeur du taux de reproduction.

D'autre part, dans les codes que nous avons obtenus pour l'assimilation séquentielle de données (voir Annabelle Collin et al. [7]), le schéma en temps implémenté pour la résolution du système d'équations différentielles est la méthode d'Euler explicite (voir (1.34)). Dans la deuxième section de ce chapitre, nous verrons que ce schéma en temps n'est pas adapté à notre système d'EDO. Par conséquent, nous expliciterons les différents schémas en temps (explicite, implicite, à pas multiple ou non) que nous avons implémentés. Nous terminerons en comparant ces différentes méthodes de résolution numérique.

3.1 Calcul du taux de reproduction

Le concept du taux de reproduction R_0 est fondamental lorsque nous souhaitons étudier la propagation d'un agent pathogène. En effet, il est défini comme le nombre moyen de cellules nouvellement infectées par une unique cellule infectée durant son temps de vie, en supposant que toutes les autres cellules sont susceptibles.

Ainsi, par définition, lorsque $R_0 < 1$, chaque cellule infectée infectera en moyenne moins d'une cellule, ce qui impliquera une élimination de l'infection. Cependant, lorsque $R_0 > 1$, l'agent pathogène est capable d'envahir le milieu.

3.1.1 *Next generation method*

D'un point de vue plus mathématique, le taux de reproduction R_0 est défini comme le rayon spectral du *next generation operator* [10]. En s'appuyant sur les travaux de Jane HEFFERNAN, O. DIEKMANN et al. [11, 13], nous expliquons comment obtenir cet opérateur .

Dans un premier temps, il faut définir un sous-système infectieux de notre système initial d'EDO. Ce système doit décrire la production de nouvelles infections et les changements entre les différentes classes infectées. De plus, il doit être linéarisé (en les variables de celui-ci.).

Supposons que nous avons n compartiments parmi lesquels m sont infectés. Nous définissons le vecteur $\bar{x}(t) = (x_i(t))_{i=1,\dots,n}$, où x_i représente la proportion d'individus dans le compartiment i à l'instant t.

Nous définissons pour chaque compartiment i deux fonctions.

— $F_i(\bar{x}(t))$ représente le taux d'apparition de nouvelles infections dans le compartiment i.

— $V_i(\bar{x}(t)) = V_i^-(\bar{x}(t)) - V_i^+(\bar{x}(t))$ où $V_i^-(\bar{x}(t))$ est le taux d'arrivée et $V_i^+(\bar{x}(t))$ le taux de départ du compartiment i.

Il est important de remarquer que F doit uniquement inclure les nouvelles infections. Les transferts entre les compartiments infectieux, par exemple entre les cellules infectées et les cellules infectées productrices de virus, ne sont pas pris en compte dans les fonctions F mais dans les fonctions V.

À partir de ces fonctions, nous obtenons deux matrices F et V définies de la manière suivante, où x_0 est l'équilibre sans maladies (*disease-free equilibrium* dans la littérature).

$$F = \left(\frac{dF_i(x_0)}{dx_j}\right)_{i,j=1,\dots,n} \qquad V = \left(\frac{dV_i(x_0)}{dx_j}\right)_{i,j=1,\dots,n} \tag{3.1}$$

Le taux de reproduction R_0 est ensuite défini par le rayon spectral (*i.e.* la plus grande valeur propre) de FV^{-1}.

3.1.2 Application au modèle de la section 2.2.1

Dans cette section, en utilisant la méthode de la *next generation matrix*, nous allons établir une formule du taux de reproduction R_0 pour le compartiment du nasopharynx et de la trachée dans le cas du modèle défini dans la section (2.2.1). Après avoir estimé les paramètres de ce modèle, nous calculerons le taux de reproduction pour chaque individu. Nous comparerons la distribution du R_0 entre nos trois groupes de singes.

Considérons le modèle mathématique défini en (2.1) pour les deux compartiments (nasopharynx et trachée). Lorsqu'il n'y a pas de maladie, notre système se réduit aux équations :

$$\begin{cases} \dot{T}^T &= 0 \\ \dot{T}^N &= 0 \end{cases} \tag{3.2}$$

Donc, $T^T(t) = T^T(0)$ et $T^N(t) = T^N(0)$. Les valeurs non nulles de l'état d'équilibre sans maladie sont alors $T^T(0)$ et $T^N(0)$.

Ensuite, nous définissons le sous-système infectieux par :

$$\begin{cases} \dot{I_1}^X &= \beta V_i^X T^X + \mu\beta V_S^X T^X - kI_1^X \\ \dot{I_2}^X &= kI_1^X - \delta I_2^X \\ \dot{V_i}^X &= P^X \mu I_2^X - cV_I^X - \beta V_i^X T^X \\ \dot{V_S}^X &= -c_I V_s^X - \mu\beta V_S^X T^X \end{cases} \tag{3.3}$$

Ce système est bien linéaire en ses variables. Il nous reste à déterminer les fonctions F_i, V_i et les matrices associées.

Dans la suite, l'indice 1 représente le compartiment I_1, l'indice 2 le compartiment I_2, l'indice 3 le compartiment V_i et l'indice 4 le compartiment V_S. Nous obtenons $F_1^X(\bar{x}) = \beta x_3^X T^X + \mu\beta x_4^X T^X - kx_1^X$ et $F_2^X(\bar{x}) = F_3^X(\bar{x}) = F_4^X(\bar{x}) = 0$. Nous avons aussi :

$$\begin{cases} V_1(\bar{x}) &= kx_1^X \\ V_2(\bar{x}) &= \delta x_2^X - kx_1^X \\ V_3(\bar{x}) &= cx_3^X + \beta x_3^X T^X - \mu P^X x_2^X \\ V_4(\bar{x}) &= c_I x_4^X + \mu\beta x_4^X T^X \end{cases} \tag{3.4}$$

En dérivant et en remplaçant T^X par les états d'équilibre sans maladie des deux compartiments, nous obtenons les matrices F et V :

$$F = \begin{pmatrix} 0 & 0 & \beta T^X(0) & \mu\beta T^X(0) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{et} \quad V = \begin{pmatrix} k & 0 & 0 & 0 \\ -k & \delta & 0 & 0 \\ 0 & -\mu P^X & c + \beta T^X(0) & 0 \\ 0 & 0 & 0 & c_I + \mu\beta T^X(0) \end{pmatrix}$$

A l'aide du logiciel MAPLE, nous obtenons le rayon spectral de FV^{-1} :

$$R_0^X = \frac{\beta T^X(0)\mu P^X}{\delta(c + \beta T^X(0))}. \tag{3.5}$$

Cette formule correspond avec celle établie dans d'autres travaux pour un modèle similaire (voir Peter CZUPPON et al. [8]). Pour des exemples d'application de cette méthode à d'autres modèles, le lecteur pourra se référer à l'ANNEXE A

Grâce à cette formule et à la statistique de Mann-Whitney, nous pouvons vérifier si le taux de reproduction entre les groupes de singes ont des distributions deux à deux différentes. Pour $((X_1, Y_1), ..., (X_n, Y_n)) \sim (\mathcal{P} \otimes \mathcal{Q})^{\otimes n}$ avec $\mathcal{P}$ et $\mathcal{Q}$ des lois de probabilités sans atome sur $\mathbb{R}$, on définit la statistique de Mann-Whitney par [6] :

$$U_n = \sum_{i,j=1}^{n} \mathbb{1}_{X_i < Y_i}. \tag{3.6}$$

Cette statistique dépend de l'ordre de l'échantillon global $(X_1, Y_1, ..., X_n, Y_n)$. Or, nous remarquons que les valeurs des taux de reproduction des deux compartiments – nasopharynx et trachée – sont ordonnées de la même manière. En effet, le singe ayant le k$^{\text{ième}}$ plus petit taux de reproduction dans la nasopharynx possède aussi le k$^{\text{ième}}$ plus petit taux de reproduction dans la trachée. De ce fait, il est suffisant d'effectuer les tests statistiques sur le compartiment nasal. Nous construisons les tests non paramétriques unilatéraux suivants dont les résultats respectifs se trouvent dans la TABLE 3.1.

1. $H_0 : \mathcal{P}_{conv} = \mathcal{P}_{naif}$ contre $H_1 : \mathcal{P}_{conv} \otimes \mathcal{P}_{naif}(T) > \frac{1}{2}$
2. $H_0 : \mathcal{P}_{conv} = \mathcal{P}_{CD-40}$ contre $H_1 : \mathcal{P}_{conv} \otimes \mathcal{P}_{CD-40}(T) < \frac{1}{2}$
3. $H_0 : \mathcal{P}_{naif} = \mathcal{P}_{CD-40}$ contre $H_1 : \mathcal{P}_{naif} \otimes \mathcal{P}_{CD-40}(T) < \frac{1}{2}$

où $T = \{(x, y) \in \mathbb{R}^2 : x \leq y\}$ et $\mathcal{P}_{naif}, \mathcal{P}_{conv}, \mathcal{P}_{CD-40}$ représentent respectivement les probabilités du taux de reproduction R_0 dans le nasopharynx des singes naifs, convalescents et convalescents-vaccinés.

Dans le premier cas, nous rejetons l'hypothèse H_0 lorsque U_n prend une valeur anormalement grande tandis que dans les deux cas suivants, nous rejetons H_0 lorsque U_n prend une valeur anormalement petite.

TABLE 3.1 – Tests de Mann-Whitney où les p-valeurs sont ajustées par la méthode de Bonferroni. Code : ***<0.001< ** <0.01<*<0.05

Test	P-valeur	P-valeur ajustée	Code
1. Convalescent vs. Naïf	0.004	0.013	*
2. Convalescent vs. CD-40	0.001	0.003	**
3. Naif vs. Vacciné	0.001	0.003	**

Ainsi, au niveau 0.05, nous rejetons l'hypothèse H_0 pour les trois tests effectués et nous déduisons que $\mathbb{P}(R_0^{conv} < R_0^{naif}) \geq \frac{1}{2}$, $\mathbb{P}(R_0^{CD-40} < R_0^{conv}) \geq \frac{1}{2}$ et que $\mathbb{P}(R_0^{CD-40} < R_0^{naif}) \geq \frac{1}{2}$ où $R_0^{naif} \sim \mathcal{P}_{naif}$, $R_0^{conv} \sim \mathcal{P}_{conv}$ et $R_0^{CD-40} \sim \mathcal{P}_{CD-40}$. Autrement dit, la vaccination et l'exposition antérieure au virus permettent de diminuer le taux de reproduction R_0 et donc, limiter la propagation de l'infection au sein de l'individu.

3.2 Les schémas en temps pour l'approche séquentielle

Rappelons que, dans le cas de données parcimonieuses, l'utilisation de l'approche séquentielle est en cours de développement. Ainsi, afin de simplifier le modèle à étudier, nous nous concentrons uniquement sur le compartiment nasal avec le modèle mathématique (3.7).

$$\begin{cases} \dot{T} &= -\beta T V_i - \mu\beta T V_S \\ \dot{I}_1 &= \beta T V_i + \mu\beta T V_S - k I_1 \\ \dot{I}_2 &= k I_1 - \delta I_2 \\ \dot{V}_i &= \mu P I_2 - c V_i - \beta T V_i \\ \dot{V}_{ni} &= (1-\mu) P I_2 - c V_{ni} \\ \dot{V}_S &= -\mu\beta T V_S - c_I V_S \end{cases} \tag{3.7}$$

Le modèle statistique est défini par l'expression (2.3) tandis que le modèle d'observation est lui aussi simplifié en se restreignant aux observations d'ARN génomique du nasopharynx :

$$gRNA_{ij} = log_{10}(V_i + V_{ni} + V_S)(t_{ij}, \psi_i) + \epsilon_{ij} \tag{3.8}$$

où $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

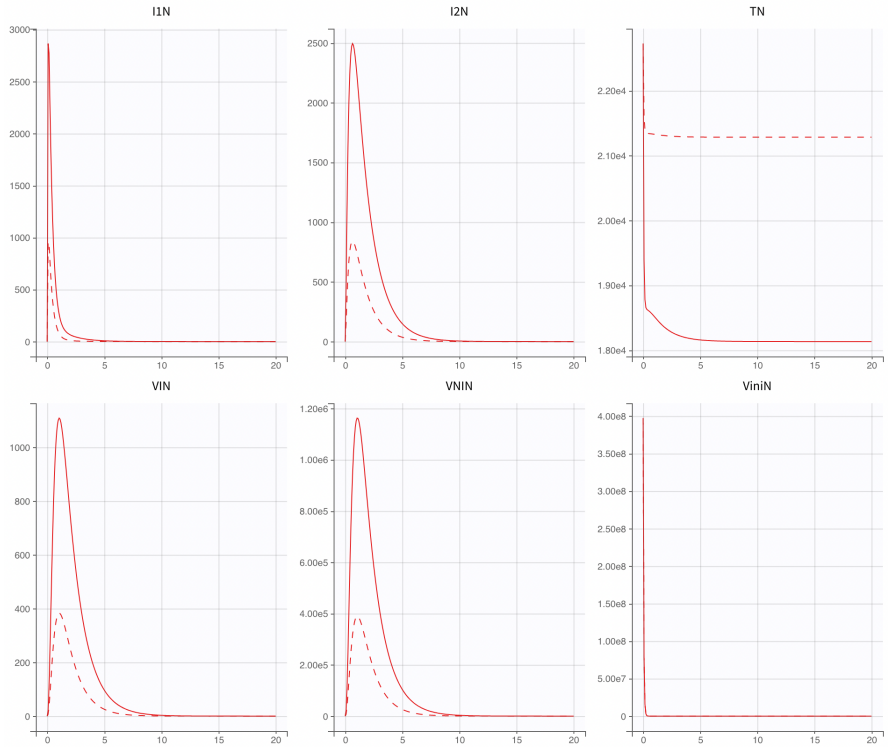
Pour réduire encore la taille du problème, nous fixons les paramètres de la manière suivante (voir section 2.3.1 et Maire Alexandre et al. [1]) :

$$k = c = 3 \quad ; \quad \mu = 10^{-3} \quad ; \quad c_I = 20 \quad \text{et} \quad P = 1600 \tag{3.9}$$

Comme il n'existe pas de solution analytique du modèle (3.7), nous devons les résoudre numériquement. Il existe de nombreuses méthodes de résolutions qui sont plus ou moins adaptées en fonction des équations que l'on cherche à résoudre.

Dans les modèles que nous avons précédemment définis, nous avons des EDO dites raides. Cela signifie que de petites variations des paramètres vont impacter grandement la dynamique de la solution. Nous pouvons constater ce fait pour le paramètre β du modèle (3.7) grâce à la Figure 3.1.

Figure 3.1 – Résolution du système (3.7) avec différentes valeurs de β et $\delta = 0.85$.
Ligne discontinue : $\beta = 10^{-5.5}$. Ligne continue : $\beta = 10^{-5}$



Ainsi, la résolution par des méthodes numériques explicites va être difficile et il faudra privilégier des méthodes implicites telles que les méthodes BDF ou de Runge-Kutta. Par exemple, dans le cas d'EDO raides, Monolix utilise une méthode BDF. Cependant, le projet initial entre les équipes SISTM et MONC (voir page 3 et Annabelle Collin et al. [7]) utilise la méthode explicite d'Euler (voir équation (1.34)) pour la résolution du système. Lorsque nous avons utilisé l'approche séquentielle avec ce schéma en temps pour estimer les paramètres du modèle nos estimations étaient anormalement mauvaises.

Ce fait pouvant être causé par la méthode de résolution utilisée, nous présentons, dans cette section, les différents schémas numériques que nous avons implémentés pour remplacer la méthode d'Euler explicite. Pour terminer, nous comparons les simulations obtenues en fonction des schémas et des pas de temps utilisés.

Nous reprenons ici les notations introduites dans la section 1.3.1 que nous appliquons au modèle (3.7). Rappelons que nous avons deux notations possibles pour la modélisation des

dynamiques. La première représentation (1.33) considère uniquement les états du système dans les dynamiques. Tandis que la deuxième (1.35) y ajoute en plus les paramètres individuels. Par goût de simplicité, nous considérerons la première notation pour l'implémentation des schémas numériques.

Ainsi, en notant $x(t) = (T(t), I_1(t), I_2(t), V_i(t), V_{ni}(t), V_S(t))$ et θ le vecteur de paramètres du modèle, nous définissons $f : \mathbb{R}^6 \times \mathbb{R}^{N_\theta} \times \mathbb{R} \longrightarrow \mathbb{R}^6$ par :

$$f(x(t), \theta, t) = \begin{pmatrix} -\beta T(t) V_i(t) - \mu\beta T(t) V_S(t) \\ \beta T(t) V_i(t) + \mu\beta T(t) V_S(t) - k I_1(t) \\ k I_1(t) - \delta I_2(t) \\ \mu P I_2(t) - c V_i(t) - \beta T(t) V_i(t) \\ (1-\mu) P I_2(t) - c V_{ni}(t) \\ -\mu\beta T(t) V_S(t) - c_I V_S(t) \end{pmatrix} \tag{3.10}$$

Comme énoncé précédemment, nous serons amenés à implémenter des schémas en temps implicites. Pour cela, nous aurons besoin d'appliquer l'algorithme de Newton pour trouver le – ou un – zéro d'une fonction $F : \mathbb{R}^6 \longrightarrow \mathbb{R}^6$. Etant donné $x_0 \in \mathbb{R}^6$, pour tout $n \geq 1$ l'algorithme de Newton se définit de la manière suivante :

$$x_n = x_{n-1} - Jac(F)^{-1}(x_{n-1}) F(x_{n-1}) \tag{3.11}$$

où $Jac(F)$ est la jacobienne de la fonction F. Nous arrêtons les itérations lorsque $\|F(x_n)\| \leq 10^{-5}$ ou que $\|Jac(F)^{-1}(x_n)F(x_n)\| \leq 10^{-5}$ où $\|.\|$ représente la norme euclidienne dans $\mathbb{R}^6$.

Pour faciliter cette étape, la jacobienne par rapport à $x = (T, I_1, I_2, V_i, V_{ni}, V_S)$ de la fonction f est donnée par :

$$Jac_x(f)(x(t), \theta, t) = \begin{pmatrix} -\beta V_i(t) - \mu\beta V_S(t) & 0 & 0 & -\beta T(t) & 0 & -\mu\beta T(t) \\ \beta V_i(t) + \mu\beta V_S(t) & -k & 0 & \beta T(t) & 0 & \mu\beta T(t) \\ 0 & k & -\delta & 0 & 0 & 0 \\ -\beta V_i(t) & 0 & \mu P & -c - \beta T(t) & 0 & 0 \\ 0 & 0 & (1-\mu)P & 0 & -c & 0 \\ -\mu\beta V_S(t) & 0 & 0 & 0 & 0 & -\mu\beta T(t) - c_I \end{pmatrix} \tag{3.12}$$

3.2.1 Crank-Nicolson et BDF d'ordre 2

Afin d'implémenter un schéma en temps proche de celui utilisé par Monolix, nous avons choisi d'implémenter un BDF d'ordre 2. Comme cette méthode est implicite à pas d'ordre 2, pour commencer à l'implémenter nous avons besoin de z_0^i et z_1^i. Ainsi, pour obtenir z_1^i nous devons utiliser une méthode implicite à pas d'ordre 1. Pour cela, nous utiliserons la méthode de Crank-Nicolson définie de la manière suivante pour chaque individu $1 \leq i \leq N$:

$$x_{n+1}^i = x_n^i + \frac{\delta T_n}{2} \left(f(x_n^i, \theta^i, T_n) + f(x_{n+1}^i, \theta^i, T_{n+1}) \right) \tag{3.13}$$

Comme la méthode de Crank-Nicolson est implicite, nous devons à chaque itération appliquer l'algorithme de Newton pour approximer le (ou un) zéro de la fonction :

$$F(x) = x - x_n^i - \frac{\delta T_n}{2} \left(f(x_n^i, \theta^i, T_n) + f(x, \theta^i, T_{n+1}) \right) \tag{3.14}$$

dont la jacobienne est :

$$Jac(F)(x) = Id_{\mathbb{R}^6} - \frac{\delta T_n}{2} Jac_x(f)(x, \theta^i, T_{n+1}) \tag{3.15}$$

Après la première itération de la méthode de Crank-Nicolson, nous pouvons alors utiliser la méthode BDF d'ordre 2 définie pour chaque individu de la manière suivante :

$$x^i_{n+2} = -\frac{1}{3}x^i_n + \frac{4}{3}x^i_{n+1} + \delta T_n \frac{2}{3} f(x^i_{n+2}, \theta^i, T_{n+2}) \tag{3.16}$$

Cette méthode étant implicite, nous devons aussi appliquer l'algorithme de Newton à la fonction :

$$F_2(x) = x + \frac{1}{3}x^i_n - \frac{4}{3}x^i_{n+1} - \delta T_n \frac{2}{3} f(x, \theta^i, T_{n+2}) \tag{3.17}$$

dont la jacobienne est donnée par :

$$Jac(F_2)(x) = Id_{\mathbb{R}^6} - \delta T_n \frac{2}{3} Jac_x(f)(x, \theta^i, T_{n+2}) \tag{3.18}$$

3.2.2 Runge-Kutta d'ordre 3 et 4 implicites

Les méthodes implicites étant propices aux équations différentielles raides, nous avons implémentés deux méthodes de Runge-Kutta d'ordre 3 et d'ordre 4. Ce sont des méthodes à un seul pas qui sont plus adaptées à l'assimilation de données par les filtres de Kalman. Pour d'autres implémentations de ces deux méthodes, nous pouvons nous référer aux travaux de Matthieu Brachet [5].

Runge-Kutta d'ordre 3 implicite

Il existe différentes implémentations de cette méthode, nous utilisons ici la méthode de Crouzeix qui présente l'avantage d'être diagonalement implicite. En posant $\alpha = \frac{\sqrt{3}}{6}$, la méthode est définie par :

$$\begin{cases} K^{(1)}_n &= f(x^i_n + \delta T_n(\frac{1}{2} + \alpha)K^{(1)}_n, \theta^i, T_n + \delta T_n(\frac{1}{2} + \alpha)) \\ K^{(2)}_n &= f\left(x^i_n + \delta T_n(-2\alpha K^{(1)}_n + (\frac{1}{2} + \alpha)K^{(2)}_n), \theta^i, T_n + \delta T_n(\frac{1}{2} - \alpha)\right) \\ x^i_{n+1} &= x^i_n + \frac{\delta T_n}{2}(K^{(1)}_n + K^{(2)}_n) \end{cases} \tag{3.19}$$

Pour déterminer $K^{(1)}_n$ et $K^{(2)}_n$, il faut effectuer, à chaque itération, deux fois l'algorithme de Newton aux fonctions respectives :

$$\begin{aligned} F(x) &= x - f(x^i_n + \delta T_n(\frac{1}{2} + \alpha)x, \theta^i, T_n + \delta T_n(\frac{1}{2} + \alpha)) \\ G(x) &= x - f\left(x^i_n + \delta T_n(-2\alpha F_0 + (\frac{1}{2} + \alpha)x), \theta^i, T_n + \delta T_n(\frac{1}{2} - \alpha)\right) \end{aligned} \tag{3.20}$$

où F_0 représente le zéro de la fonction F déterminé par l'algorithme de Newton. Les jacobiennes de ces deux fonctions sont :

$$\begin{aligned} Jac(F)(x) &= Id_{\mathbb{R}^6} - \delta T_n(\frac{1}{2} + \alpha) Jac_x(f)\left(x^i_n + \delta T_n(\frac{1}{2} + \alpha)x, \theta^i, T_n + \delta T_n(\frac{1}{2} + \alpha)\right) \\ Jac(G)(x) &= Id_{\mathbb{R}^6} - \delta T_n(\frac{1}{2} + \alpha) Jac_x(f)\left(x^i_n + \delta T_n(-2\alpha F_0 + (\frac{1}{2} + \alpha)x), \theta^i, T_n + \delta T_n(\frac{1}{2} - \alpha)\right) \end{aligned} \tag{3.21}$$

Runge-Kutta d'ordre 4 implicite

De même que pour la méthode précédente, il existe différentes implémentations possibles. Nous choisissons d'utiliser ici la méthode de Norsett qui est diagonalement implicite d'ordre 4. Elle est définie par :

$$\begin{cases} K_n^{(1)} &= f(x_n^i + \delta T_n \alpha K_n^{(1)}, \theta^i, T_n + \delta T_n \alpha) \\ K_n^{(2)} &= f\left(x_n^i + \delta T_n((\frac{1}{2} - \alpha)K_n^{(1)} + \alpha K_n^{(2)}), \theta^i, T_n + \frac{1}{2}\delta T_n\right) \\ K_n^{(3)} &= f\left(x_n^i + \delta T_n(2\alpha K_n^{(1)} + (1-4\alpha)K_n^{(2)} + \alpha K_n^{(3)}, \theta^i, T_n + (1-\alpha)\delta T_n\right) \\ x_{n+1}^i &= x_n^i + \delta T_n\left(\frac{1}{6(1-2\alpha)^2}K_n^{(1)} + \frac{3(1-2\alpha)^2 - 1}{3(1-2\alpha)^2}K_n^{(2)} + \frac{1}{6(1-2\alpha)^2}K_n^{(3)}\right) \end{cases} \tag{3.22}$$

où α est solution de $x^3 - \frac{3}{2}x^2 + \frac{1}{2}x - \frac{1}{24} = 0$. Les solutions de cette équation peuvent être approchées par les valeurs de l'ensemble suivant : $\{1,06858\ ;\ 0,30254\ ;\ 0.12889\}$. Nous obtenons les meilleurs propriétés de stabilité en prenant $\alpha = 1,06858$.

Pour déterminer $K_n^{(1)}$, $K_n^{(2)}$ et $K_n^{(3)}$, il faut effectuer, à chaque itération, trois fois l'algorithme de Newton aux fonctions respectives : respectives :

$$\begin{aligned} F(x) &= x - f(x_n^i + \delta T_n \alpha x, \theta^i, T_n + \delta T_n \alpha) \\ G(x) &= x - f\left(x_n^i + \delta T_n((\frac{1}{2} - \alpha)F_0 + \alpha x), \theta^i, T_n + \frac{1}{2}\delta T_n\right) \\ H(x) &= x - f\left(x_n^i + \delta T_n(2\alpha F_0 + (1-4\alpha)G_0 + \alpha x, \theta^i, T_n + (1-\alpha)\delta T_n\right) \end{aligned} \tag{3.23}$$

où F_0 et G_0 représente les zéros des fonctions F et G, respectivement, déterminés par l'algorithme de Newton. Les jacobiennes de ces trois fonctions sont données par les expressions suivantes :

$$\begin{aligned} Jac(F)(x) &= x - \delta T_n \alpha Jac_x(f)(x_n^i + \delta T_n \alpha x, \theta^i, T_n + \delta T_n \alpha) \\ Jac(G)(x) &= x - \delta T_n \alpha Jac_x(f)\left(x_n^i + \delta T_n((\frac{1}{2} - \alpha)F_0 + \alpha x), \theta^i, T_n + \frac{1}{2}\delta T_n\right) \\ Jac(H)(x) &= x - \delta T_n \alpha Jac_x(f)\left(x_n^i + \delta T_n(2\alpha F_0 + (1-4\alpha)G_0 + \alpha x, \theta^i, T_n + (1-\alpha)\delta T_n\right) \end{aligned} \tag{3.24}$$

3.2.3 Runge-Kutta d'ordre 4 explicite

Pour terminer , nous avons aussi choisi d'implémenter la méthode de Runge-Kutta explicite d'ordre 4. Cette méthode est théoriquement plus précise que la méthode explicite d'Euler car nous augmentons significativement l'ordre. Elle comporte aussi l'avantage de réduire le temps de calcul en comparaison avec les méthodes implicites présentées précédemment. En effet, pour déterminer les valeurs de $K_n^{(1)}$, $K_n^{(2)}$, $K_n^{(3)}$ et $K_n^{(4)}$ nous n'utilisons plus l'algorithme de Newton. Ce dernier ne converge pas nécessairement ou peut être très couteux en temps lorsque notre valeur d'initialisation est trop éloignée de la solution.

Cette méthode est donc définie par les équations suivantes :

$$\begin{cases} K_n^{(1)} &= f(x_n^i, \theta^i, T_n) \\ K_n^{(2)} &= f\left(x_n^i + \dfrac{\delta T_n}{2} K_n^{(1)}, \theta^i, T_n + \dfrac{\delta T_n}{2}\right) \\ K_n^{(3)} &= f\left(x_n^i + \dfrac{\delta T_n}{2} K_n^{(2)}, \theta^i, T_n + \dfrac{\delta T_n}{2}\right) \\ K_n^{(4)} &= f\left(x_n^i + \delta T_n K_n^{(3)}, \theta^i, T_n + \delta T_n\right) \\ x_{n+1}^i &= x_n^i + \dfrac{\delta T_n}{6}\left(K_n^{(1)} + 2K_n^{(2)} + 2K_n^{(3)} + K_n^{(4)}\right) \end{cases} \tag{3.25}$$

3.2.4 Comparaisons des simulations obtenues

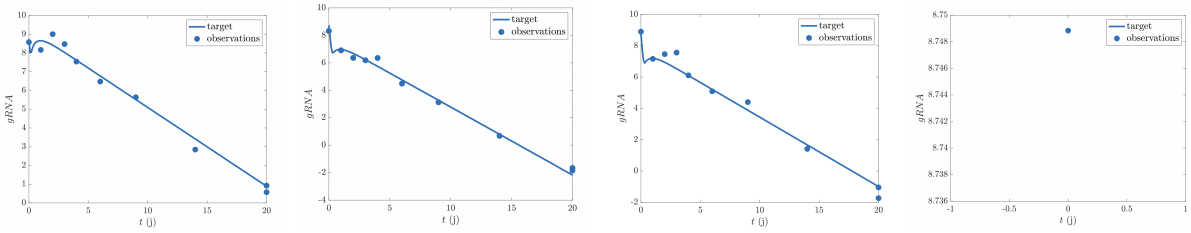
Dans cette section, nous comparons les différents schémas en temps implémentés en fonction des pas de temps utilisés et des valeurs des paramètres β et δ. Nous souhaitons que les schémas en temps convergent pour un large choix de valeurs des paramètres avec un pas de temps acceptable. Nous terminons en expliquant les difficultés que nous rencontrons lors de l'assimilation séquentielle avec ces schémas en temps.

Dans un premier temps, comparons les méthodes d'Euler explicite et de Runge-Kutta d'ordre 4 explicite. En nous appuyant sur la FIGURE 3.2, nous remarquons que la méthode de Runge-Kutta d'ordre 4 explicite ne converge pas systématiquement vers la solution, même en choisissant un pas de temps très petit (10^{-4}). Ce phénomène se produit lorsque nous définissons β_{pop} et δ_{pop} trop éloignés des valeurs biologique probables ($\beta_{pop} = 10^{-5,44}$, $\delta_{pop} = 0,85$).

Cependant, la méthode d'Euler explicite ne possède pas ce problème à condition de choisir un pas de temps suffisamment petit. Le pas 5×10^{-4} permet une convergence vers la solution numérique dans les cas que nous avons testés.

En complément, nous pouvons nous référer aux FIGURES C.1, C.2, C.3, C.4 et C.5 en ANNEXES. Dans ces graphiques, chaque ligne représente les observations simulées de deux singes fictifs et nous faisons décroitre le pas de haut en bas.

FIGURE 3.2 – Comparaison des méthodes explicites.
Les trois premiers graphiques représentent les simulations en utilisant la méthode d'Euler explicite. Le dernier graphique utilise la méthode de Runge-Kutta d'ordre 4 explicite.



(a) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 10^{-3}

(b) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 5×10^{-4}

(c) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 1×10^{-4}

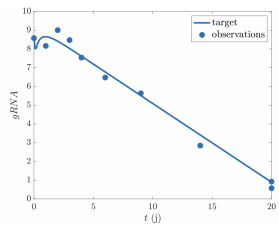
(d) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 1×10^{-4}

Dans un second temps, nous montrons que les méthodes implicites implémentées sont plus performantes que les méthodes explicites pour gérer les EDO raides. Dans la FIGURE 3.3, nous comparons les méthodes de Runge-Kutta d'ordre 3 implicite et d'Euler explicite. Le lecteur pourra vérifier avec les figures des sections C.2, C.3 et C.4 que l'on obtient les mêmes conclusions

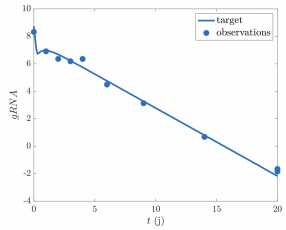
si l'on utilise la méthode de Runge-Kutta d'ordre 4 implicite. Par ailleurs, nous obtenons des différences légèrement moins franches en utilisant la méthode de Crank-Nicolson et BDF d'ordre 2 (voir FIGURES C.8).

Il est bienvenu de rappeler qu'en contrepartie les méthodes implicites doivent effectuer à chaque itération une ou plusieurs fois l'algorithme de Newton, qui globalement augmente les temps de calcul.

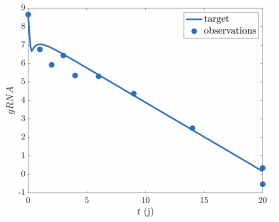
FIGURE 3.3 – Comparaison des méthodes explicites et implicites.
Les deux premières colonnes représentent les simulations en utilisant la méthode d'Euler explicite. Les deux dernières utilisent la méthode de Runge-Kutta d'ordre 3 implicite.



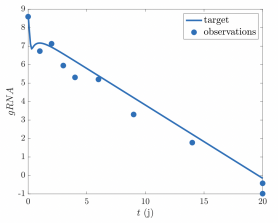
(a) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 10^{-3}



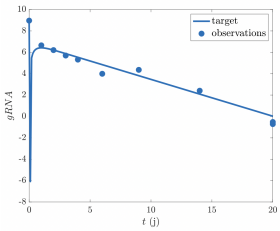
(b) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 5×10^{-4}



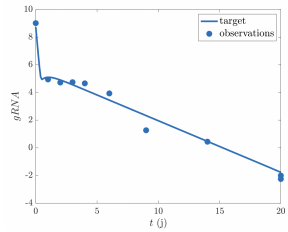
(c) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 5×10^{-2}



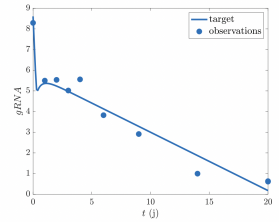
(d) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 1×10^{-2}



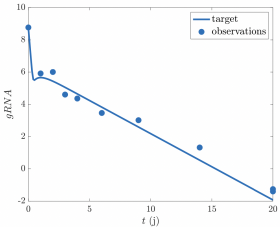
(e) $\beta = 10^{-5.44}$
et $\delta = 0.85$
Pas de temps 10^{-1}



(f) $\beta = 10^{-5.44}$
et $\delta = 0.85$
Pas de temps 5×10^{-2}



(g) $\beta = 10^{-5.44}$
et $\delta = 0.85$
Pas de temps 10^{-1}



(h) $\beta = 10^{-5.44}$
et $\delta = 0.85$
Pas de temps 5×10^{-2}

Bien que les schémas en temps implémentés convergent plus ou moins bien vers la solution numérique de notre système d'EDO, l'utilisation de ces schémas pour l'assimilation séquentielle de nos données ne fonctionne pas encore. A ce jour nous rencontrons plusieurs difficultés : les schémas explicites n'estiment pas correctement les paramètres en utilisant des valeurs initiales très proches des valeurs attendues, l'algorithme de Newton utilisé pour les schémas implicites ne converge pas.

Chapitre 4

Les résultats

Ce chapitre est consacré à la présentation des principaux résultats numériques obtenus au cours de ce travail.

Dans un premier temps, nous présenterons un schéma de simulations associé au modèle (3.7). Quelque soit la méthode d'estimation utilisée, cet outil permet de montrer les bonnes propriétés de notre méthode en calculant les biais de nos estimations par rapport aux paramètres réels. Nous calculerons aussi les taux de couverture qui représentent les proportions d'intervalles de confiance contenant la valeur réelle d'un paramètre.

Dans un second temps, nous appliquerons ce schéma de simulation à la méthode variationnelle. Lorsque nous aurons terminé l'implémentation de la méthode séquentielle, nous lui appliquerons ce même schéma de simulation. Ceci nous permettra de comparer les deux approches dans le cas de données parcimonieuses.

Pour terminer, nous utiliserons l'assimilation variationnelle de données pour l'étude des modèles définis dans les sections 2.2.2 et 2.2.3. Nous comparerons ces deux modèles et étudierons comment leurs différences agissent. De manière similaire, l'approche séquentielle aurait pu être envisagée si son implémentation avait été terminée.

4.1 Schéma de simulations

Nous allons simuler des données d'observations à partir du modèle (3.7). Ainsi, nous connaitrons les paramètres *réels* du modèle ayant permis de telles observations. Nous ferons cela pour 100 groupes constitués de 10 singes fictifs. Par la suite, nous estimerons les paramètres qui ont permis ces simulations et les comparerons avec les paramètres réels.

Pour la comparaison des paramètres estimés et réels, nous faisons la différence entre les paramètres individuels et les paramètres de population, où une population est un groupe de 10 singes. Dans la suite, nous supposons qu'il existe une fonction monotone h telle que $h(\theta_i) = h(\theta_{pop}) + u_i^\theta$ où $u_i^\theta \sim \mathcal{N}(0, \omega_\theta^2)$. Ainsi, les paramètres de population sont θ_{pop} et ω_θ tandis que les paramètres individuels sont θ_i et u_i^θ.

4.1.1 Comparaison des paramètres de population

Nous définissons le biais relatif d'un paramètre de population de la manière suivante :

$$\frac{h(\theta_{pop}^{\text{target}}) - h(\theta_{pop}^{\text{estim}}))}{h(\theta_{pop}^{\text{target}})} \quad \text{si } h(\theta_{pop}^{\text{target}}) \neq 0 \qquad \text{et} \qquad \frac{\omega_\theta^{\text{target}} - \omega_\theta^{\text{estim}}}{\omega_\theta^{\text{target}}} \quad \text{si } \omega_\theta^{\text{target}} \neq 0 \tag{4.1}$$

Similairement, nous définissons le biais relatif de l'écart-type de l'erreur résiduelle σ du modèle d'observation par :

$$\frac{\sigma^{\text{target}} - \sigma^{\text{estim}}}{\sigma^{\text{target}}} \quad \text{si } \sigma^{\text{target}} \neq 0 \tag{4.2}$$

En complément, nous définissons le taux de couverture des paramètres de population par la proportion d'intervalles de confiance estimés contenant le paramètre réel $h(\theta_{pop}^{\text{estim}})$. L'intervalle de confiance à 95% est défini par :

$$CI_{\theta_{pop}} = \left[h(\theta_{pop}^{\text{estim}}) - 1.96\widehat{s.e.}(h(\theta_{pop}^{\text{estim}}))\ ;\ h(\theta_{pop}^{\text{estim}}) + 1.96\widehat{s.e.}(h(\theta_{pop}^{\text{estim}}))\right] \tag{4.3}$$

où $\widehat{s.e.}(h(\theta_{pop}^{\text{estim}}))$ est l'erreur standard du paramètre estimée dans le modèle de population par la matrice d'information de Fisher.

4.1.2 Comparaison des paramètres individuels

Pour l'étude des paramètres individuels, nous rappelons la notion de *shrinkage* introduite dans la section 1.2.2. Pour chaque paramètre θ possédant une variabilité individuelle (*i.e.* $\omega_\theta^{\text{target}} \neq 0$) , nous définissons le *shrinkage* par :

$$\text{shrinkage} = 1 - \frac{Var(u_i^\theta)}{{\omega_\theta^{\text{estim}}}^2} \tag{4.4}$$

où $Var(u_i^\theta)$ est la variance empirique des estimations des effets aléatoires u_i^θ au sein de chaque groupe et $\omega_\theta^{\text{estim}}$ est l'écart-type du paramètre individuel estimé dans le modèle populationnel.

Lorsque le *shrinkage* est proche de 1 cela signifie que les paramètres des différents individus sont proches du mode de la distribution du paramètre dans la population. Dans ce cas, les outils d'évaluation individuels ne représentent pas la variabilité individuelle et ne permettent pas une évaluation correcte du modèle (voir Thi Huyen Tram Nguyen et al. [17]).

Pour éviter ce phénomène de *shrinkage*, Monolix peut simuler pour chaque individu un échantillon de 10 paramètres individuels en utilisant la distribution conditionnelle du paramètre en question. Ceci permet de prendre en compte la variabilité de cette distribution au lieu de considérer uniquement le mode – la valeur la plus probable. Ainsi, on multiplie par 10 la taille de l'échantillon utilisé pour calculer $Var(u_i^\theta)$ dans la formule du *shrinkage* et dans les outils d'évaluation individuels.

Nous définissons ensuite le biais individuel par :

$$h(\theta_i^{\text{target}}) - h(\theta_i^{\text{estim}}(EBE)) \tag{4.5}$$

où $\theta_i^{\text{estim}}(EBE)$ est l'estimation du mode de la distribution conditionnelle du paramètre θ_i.

De la même manière que précédemment, nous définissons le taux de couverture comme la proportion d'intervalle de confiance estimé contenant le paramètre réel $h(\theta_i^{\text{target}})$, où l'intervalle de confiance à 95% est défini par :

$$CI_{\theta_i} = \left[h(\theta_i^{\text{estim}}(EBE)) - 1.96\sqrt{Var(h(\theta_i))}\ ;\ h(\theta_i^{\text{estim}}(EBE)) + 1.96\sqrt{Var(h(\theta_i))}\right] \tag{4.6}$$

où $Var(h(\theta_i)) = Var(h(\theta_{pop})) + Var(u_i^\theta)$. Pour rappel, $Var(h(\theta_{pop})) = \widehat{s.e.}(h(\theta_{pop}^{estim}))^2$ est le carré de l'erreur standard du paramètre, estimée dans le modèle de population par la matrice d'information de Fisher (voir section 1.2.3) et $Var(u_i^\theta)$ représente la variance empirique des effets aléatoires de 10 simulations suivant la loi conditionnelle du paramètre θ_i.

4.2 Application à l'approche variationnelle

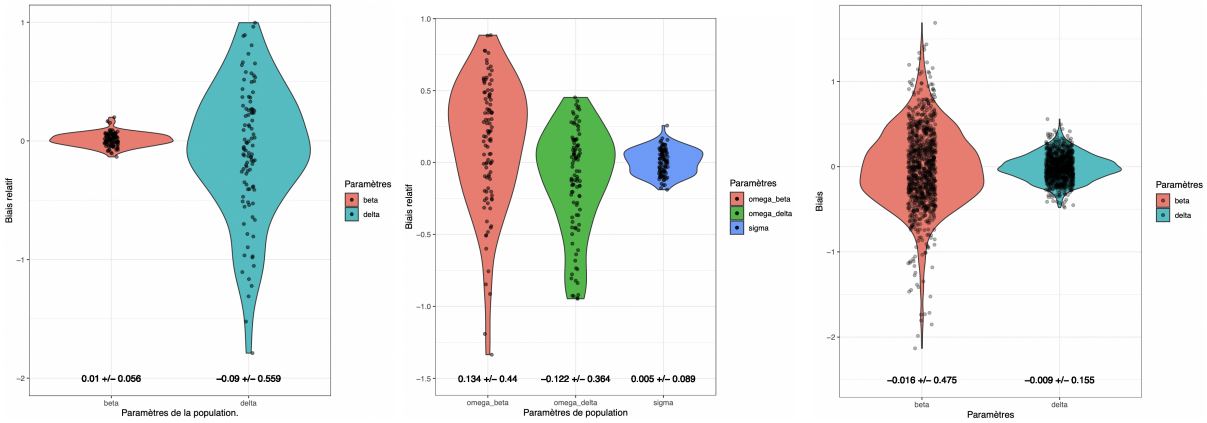
Dans cette section, nous appliquons le schéma de simulation précédemment défini dans le cas de l'approche variationnelle. Lorsque l'implémentation de l'approche séquentielle sera terminée,

nous comparerons les deux approches en appliquant ce même schéma de simulation à l'approche séquentielle.

La FIGURE 4.1 représente la distribution biais des paramètres de la population et des paramètres individuels. Nous pouvons résumer les différents résultats numériques obtenus dans le tableau suivant.

Approche	Paramètre	Population		Individuel	
		Biais relatif	Taux de couverture	Biais individuel	Taux de couverture
Variationnelle	β	0.01 ± 0.056	0.95	-0.016 ± 0.475	0.92
	ω_β	0.134 ± 0.44	0.85	Non défini	
	δ	-0.09 ± 0.559	0.90	-0.009 ± 0.155	0.88
	ω_δ	-0.122 ± 0.364	0.92	Non défini	
	σ	0.005 ± 0.089	0.94	Non défini	
Séquentielle	...	...	...	...	...

FIGURE 4.1 – Biais des paramètres du modèle (3.7) estimé par l'approche variationnelle.
Les moyennes et écart-types des biais respectivement à chaque paramètre étudié sont inscrits sur le graphique.



(a) Biais relatifs des paramètres β_{pop} et δ_{pop}.

(b) Biais relatifs des paramètres ω_β, ω_δ et σ

(c) Biais individuels des paramètres β et δ.

Les biais obtenus par l'approche variationnelle sur le modèle étudié sont très satisfaisants. En revanche, nous pouvons remarquer que certains taux de couverture sont en-dessous des 95% attendus. En effet, la construction des intervalles de confiance et des taux de couvertures reposent sur le Théorème Central Limite [3]. Comme nous appliquons ce théorème sur des échantillons de petite taille, les taux de couverture pourraient être améliorés si nous augmentions la taille de nos échantillons. Quand bien même, les taux obtenus restent acceptables et nous permettent de conclure que l'approche variationnelle est fiable.

Lorsque l'implémentation de l'approche séquentielle sera terminée, nous pourrons appliquer le même schéma de simulation et compléter le précédent tableau avec les résultats obtenus. Cela nous permettra de comparer les deux approches sur un même modèle avec des données parcimo-

nieuses et fictives de singes. Dans la poursuite des travaux, nous pourrions ensuite enlever, au fur et à mesure, les simplifications faites sur le modèle (3.7) et l'estimer par l'approche séquentielle.

4.3 Études sur données réelles avec l'approche variationnelle

Dans cette section, nous utilisons l'approche variationnelle pour estimer les paramètres des modèles définis dans les section 2.2.2 et 2.2.3. Pour permettre cette estimation nous utilisons les données réelles présentées dans la section 2.1.2 et utilisées dans différents travaux [1, 16].

Dans un premier temps, nous présentons les premiers résultats des deux modèles et nous analysons leurs différences au travers des modélisations obtenues.

Dans un second temps, nous cherchons un meilleur modèle intermédiaire et les raisons pour lesquelles le modèle 2.2.2 est significativement meilleur que le modèle 2.2.3.

4.3.1 Les résultats des modèles 2.2.2 et 2.2.3

Avant toute chose, nous analysons les estimations des paramètres des deux modèles (voir FIGURES B.3 et B.5). Nous remarquons que le deuxième modèle n'est pas précis pour l'estimation du paramètre H. Ce paramètre représente la réaction immunitaire innée face au virus *SARS-CoV-2*. Cependant, le modèle estime sa valeur très proche 0 ce qui signifierait qu'il n'y a pas réaction immunitaire innée ce qui parait peu cohérent [12].

Ensuite, nous observons et comparons les graphiques diagnostiques. En observant les deux graphiques individuels de la FIGURE 4.2, nous pouvons conclure que le premier modèle représente de manière plus précise et avec une tendance à sur-apprendre les dynamiques individuelles que le second. Les VPC nous montrent que les dynamiques centrales et la variabilité inter-individuelle sont mieux prises en compte dans le premier modèle. En complément, le lecteur pourra se référer aux FIGURES B.4 et B.6.

Pour terminer, nous comparons les critères d'informations de ces deux modèles. Comme énoncé précédemment, nous utilisons dans ce rapport le BICc défini par l'expression (2.12) et estimé par l'algorithme Monte Carlo par échantillonnage important. Le modèle défini dans la section 2.2.2 possède un BICc de 768 tandis que celui du deuxième modèle vaut 823. Ces résultats incitent donc à sélectionner le premier modèle.

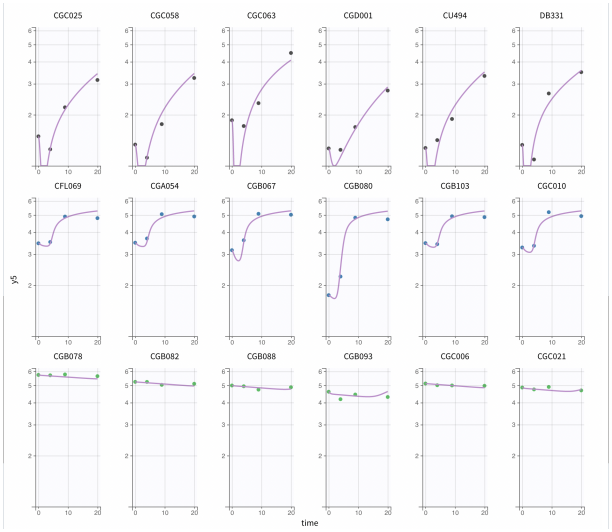
En plus de conclure que le premier modèle est meilleur que le second, nous aimerions déterminer quelle est la cause de cela et s'il existe un meilleur modèle intermédiaire. En effet, comme nous l'avons déjà énoncé, les différences entre ces deux modèles sont relativement limitées. Dans la section suivante, nous allons mettre en concurrence plusieurs modèles afin d'analyser les deux différences les plus notables.

4.3.2 Analyse des différences entre les modèles

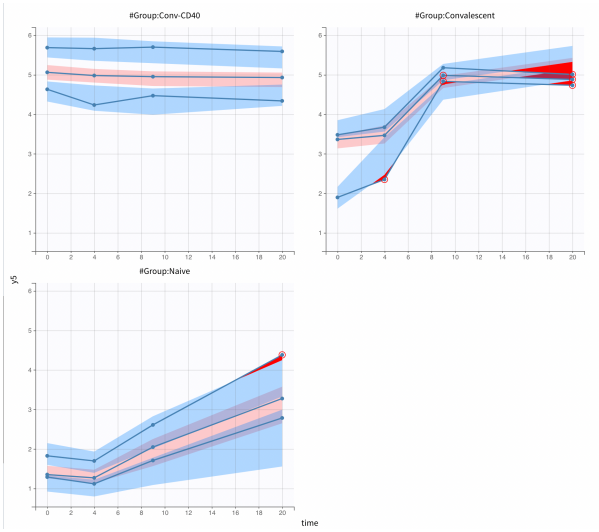
La première différence que nous souhaitons analyser correspond au déclenchement de la réaction immunitaire. Le premier modèle considère que la réaction immunitaire est déclenchée par la présence de virus (V) produit par les cellules I_2, tandis que le second modèle considère que cela provient de la présence de cellules infectées productrices de virions (I_2).

La deuxième concerne le terme $\delta I_2^{X\,H}$ dans le modèle de la section 2.2.3. D'une part, nous avons vu que l'algorithme a tendance à estimer H très proche de 0, ce qui reviendrait à dire qu'il n'y a pas de réponse immunitaire innée [12]. D'autre part, nous pouvons difficilement faire confiance à cette estimation du paramètre H au vu de la valeur de la r.s.e.

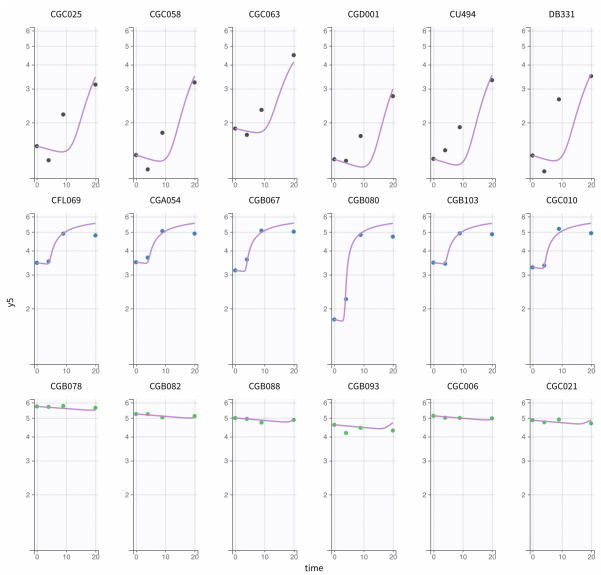
FIGURE 4.2 – Graphiques diagnostiques des modèles 2.2.2 et 2.2.3
La première ligne est issue du modèle 2.2.2 et la deuxième du modèle 2.2.3.



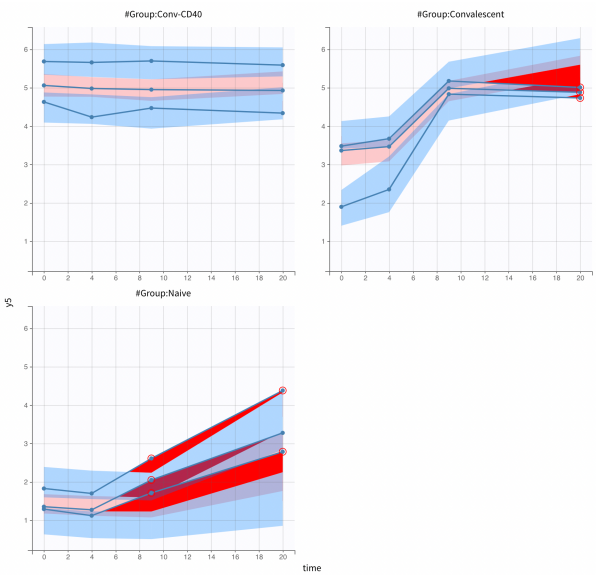
(a) Graphiques individuels des IgG RBD.



(b) VPC des IgG RBD par groupes.



(c) Graphiques individuels des IgG RBD.



(d) VPC des IgG RBD par groupes.

Déclenchement de la réaction immunitaire

Nous rappelons les définitions suivantes :

$$V = V_i^T + V_{ni}^T + V_i^N + V_{ni}^N \qquad \text{et} \qquad I_2 = I_2^T + I_2^N \tag{4.7}$$

Dans un premier temps, nous remplaçons l'équation différentielle ordinaire de Ab_1 dans le modèle (2.6) par :

$$\dot{Ab_1} = \gamma I_2 + r_A \left(1 - \frac{Ab_1}{Ab_{max}}\right) Ab_1 - k_{Ab} Ab_1 \tag{4.8}$$

et comparons les estimations et graphiques obtenus avec ceux présentés dans la section précédente.

Nous remarquons que les dynamiques des IgG des singes naïfs et convalescents sont moins précises qu'avec le modèle originel (voir Annexes B.4). Par ailleurs, avec ce modèle, nous obtenons un BICc égal à 800. Nous concluons donc que cette modification détériore le modèle défini dans la section 2.2.2.

Profil de vraisemblance sur H

Les précédentes conclusions nous motivent à modifier le modèle (2.9) en remplaçant l'EDO de M_1 par :

$$\dot{M_1} = \gamma V + r_A \left(1 - \frac{M_1}{Ab_{max}}\right) M_1 - k_{Ab} M_1 \tag{4.9}$$

En complément de cette modification, nous effectuons un profil de vraisemblance sur la variable H (voir section 2.3.1 pour la méthode). Nous estimons alors notre modèle (2.9) modifié avec les valeurs de $H \in \{0,005\ ;\ 0,007\ ;\ 0,008\ ;\ 0,0085\ ;\ 0,009\ ;\ 0,0095\ ;\ 0,01\ ;\ 0,015\ ;\ 0,02\ ;\ 0,025\ ;\ 0,03\ ;\ 0,05\}$. Nous minimisons notre BICc à 789 avec $H = 0,025$. Nous réitérons un profil de vraisemblance avec un pas plus petit autour de $0,025$ sans que cela améliore notre critère.

Pour terminer, nous analysons les estimations des paramètres et les graphiques diagnostiques obtenus dans l'Annexe B.5. Bien que nous ayons une augmentation du BICc par rapport au modèle de la section 2.2.2, nous réduisons l'estimation de l'erreur résiduelle de l'ARN génomique du nasopharynx.

Pour conclure cette section, lorsque nous souhaitons modéliser conjointement la charge virale et la réponse immunitaire, il est plus vraisemblable que la réaction immunitaire soit déclenchée par la présence de virus dans le milieu plutôt que de cellules productrices de virus. Dans le tableau suivant, nous résumons les différents BICc obtenus lors de l'estimation des différents modèles.

Modèle de base	Modification	BICc
Modèle 2.2.2	Aucune	768
	Remplacement de V par I_2 dans la mise en place de la réaction immunitaire	800
	Ajout du terme $\delta(I_2^X)^H I_2^X$	≥ 1000
Modèle 2.2.3	Aucune	823
	Remplacement de I_2 par V dans la mise en place de la réaction immunitaire	789
	$H = 0$	964

Conclusion

Dans ce rapport, nous avons présenté deux méthodes d'assimilation de données. La première, l'approche variationnelle, est déjà utilisée dans de nombreuses équipes de recherche pour l'estimation de modèles mécanistes en présence de données parcimonieuses. La deuxième, l'approche séquentielle par les filtres de Kalman, est efficace dans les domaines où il est possible d'obtenir des observations en temps quasi-continu.

Dans un premier temps, nous cherchons à savoir s'il est possible d'appliquer l'approche séquentielle en présence de données parcimonieuses, ce qui est le cas lors d'études médicales. Pour ce faire, nous utilisons un modèle simplifié modélisant la charge virale de *SARS-CoV-2* dans le nasopharynx chez des singes. Le modèle que nous considérons ici reste complexe en deux points : le système d'équations différentielles ordinaires considéré est non linéaire et raide. Nous avons montré que, dans ce cas, le choix du schémas en temps utilisé pour la résolution du système et la convergence vers la solution numérique est important.

Ensuite, pour comparer les deux approches – variationnelle et séquentielle – nous avons défini un schéma de simulation permettant, quelque soit la méthode utilisée, de quantifier la qualité de nos estimations. En appliquant ce schéma de simulation, nous avons vérifié la fiabilité de l'approche variationnelle dans le cas de données parcimonieuses. Dans la poursuite de ces travaux, nous finirons l'implémentation de l'approche variationnelle et appliquerons le schéma de simulation. Ainsi, nous pourrons comparer les deux approches dans le cas d'assimilation des données parcimonieuses.

Dans un second temps, nous souhaitions modéliser conjointement la charge virale et la réaction immunitaire. Nous nous sommes donc confrontés au double problème inverse consistant à la modélisation du phénomène puis à l'estimation des paramètres de ce modèle. En faisant une revue de la littérature scientifique, nous avons mis en avant deux modèles mécanistes que nous avons estimés en utilisant l'approche variationnelle. De fait, nous aurions aussi pu utiliser l'approche séquentielle si son implémentation était terminée.

Dans l'objectif de trouver un modèle plus vraisemblable, nous avons estimé les modèles intermédiaires. Ceci nous a permis de déterminer de quelle manière les différences agissent entre ces deux modèles. Avec les données à notre disposition, nous en avons conclu qu'il est plus vraisemblable que la réaction immunitaire soit déclenchée par la présence de virus produits par les cellules infectées que par la présence de cellules infectées productrices de virus. D'un point de vue biologique, cette questions est toujours ouverte. Il serait alors bienvenu d'effectuer ces mêmes comparaisons en utilisant des données issues d'autres études.

Bibliographie

[1] Marie Alexandre et al. “Modelling the response to vaccine in non-human primates to define SARS-CoV-2 mechanistic correlates of protection”. In : *eLife* (2022) (cf. p. 23-27, 34, 41, 50).

[2] Didier Auroux. “Étude de quelques méthodes d’assimilation de données pour l’environnement”. Université de Nice Sophia Antipolis, 2003 (cf. p. 9, 11, 21).

[3] P. Barbe et M. Ledoux. *Probabilité*. 2007 (cf. p. 49).

[4] Giuseppina Bellu et al. “DAISY : A New Software Tool to Test Global Identifiability of Biological and Physiological Systems”. In : *Computer methods and programs in biomedicine* (2007) (cf. p. 31).

[5] Matthieu Brachet. *Autour des méthodes de Runge-Kutta*. 2020. url : https://matthieubrachet.files.wordpress.com/2020/04/runge_kutta.pdf (cf. p. 43).

[6] Benoît Cadre et Céline Vial. *Statistique mathématique, cours et exercices corrigés*. 2012 (cf. p. 39).

[7] Annabelle Collin, Mélanie Prague et Philippe Moireau. “Estimation for dynamical systems using a population-based Kalman filter – Applications in computational biology”. In : *MathematicS In Action* (2022) (cf. p. 18, 22, 37, 41).

[8] Peter Czuppon et al. “Success of prophylactic antiviral therapy for SARS-CoV-2 : Predicted critical efficacies and impact of different drug-specific mechanisms of action”. In : *PLOS Computational Biology* (2021) (cf. p. 29, 39).

[9] B. Delyon, Marc Lavielle et Eric Moulines. “Convergence of a stochastic approximation version of EM algorithm”. In : *The Annals of Statistics* 27 (1999) (cf. p. 14, 15).

[10] Odo Diekmann, J.A.P. Heesterbeek et Johan Metz. “On the Definition and the Computation of the Basic Reproduction Ratio R0 in Models For Infectious-Diseases in Heterogeneous Populations”. In : *Journal of mathematical biology* (1990) (cf. p. 37, 38).

[11] Odo Diekmann, J.A.P. Heesterbeek et M.G. Roberts. “The construction of next-generation matrices for compartmental epidemic models”. In : *Journal of the Royal Society, Interface / the Royal Society* (2009) (cf. p. 37, 38).

[12] Ashish Goyal, E. Fabian Cardozo-Ojeda et Joshua T. Schiffer. “Potency and timing of antiviral therapy as determinants of duration of SARS-CoV-2 shedding and intensity of inflammatory response”. In : *Science Advances* (2020) (cf. p. 29, 50).

[13] Jane Heffernan, R.J. Smith et L.M. Wahl. “Perspectives on the Basic Reproductive Ratio”. In : *Journal of the Royal Society, Interface / the Royal Society* (2005) (cf. p. 37, 38).

[14] Marc Lavielle. *Mixed Effects Models for the Population Approach : Models, Tasks, Methods and Tools*. 2014 (cf. p. 13, 17, 18, 32, 33).

[15] Lixoft. *Monolix Documentation*. 2021. url : https://monolix.lixoft.com (cf. p. 13).

[16] Romain Marlin et al. “Targeting SARS-CoV-2 receptor-binding domain to cells expressing CD40 improves protection to infection in convalescent macaques”. In : (2021) (cf. p. 24, 50).

[17] Thi Huyen Tram Nguyen et al. “Model Evaluation of Continuous Data Pharmacometric Models : Metrics and Graphics”. In : *CPT : Pharmacometrics and Systems Pharmacology* (2016) (cf. p. 32, 48).

[18] Pranesh Padmanabhan, Rajat Desikan et Narendra Dixit. “Modeling how antibody responses may determine the efficacy of COVID-19 vaccines”. In : *Nature Computational Science* (2022) (cf. p. 29).

[19] Kasia A. Pawelek et al. “Modeling Within-Host Dynamics of Influenza Virus Infection Including Immune Responses”. In : *PLOS Computational Biology* (2012) (cf. p. 29).

[20] Alan Perelson et Ruian Ke. “Mechanistic Modeling of SARS-CoV-2 and Other Infectious Diseases and the Effects of Therapeutics”. In : *Clinical Pharmacology and Therapeutics* (2021) (cf. p. 29).

Acronymes

ACE2 Enzyme de Conversion de l'Angiotensine 2. 23, 24, 27

AIC Critère d'Information d'Akaike. 33

ARN Acides RiboNucléiques. 24, 25, 27, 40, 52

BDF *Backward Differenciation Formula*. 41, 46, 85

BIC Critère d'Information Bayésien. 33

BICc Critère d'Information Bayésien corrigé. 33, 34, 50, 52, 53

EBE *Empirical Bayes Estimate*. 17

EDO Équation Différentielle Ordinaire. 9, 11–13, 25, 37, 38, 41, 45, 46, 52

FIM Matrice d'Information de Fisher. 13, 18

gRNA ARN génomique. 34

IgG Immunoglobuline G. 24, 27, 51, 52

pfu Unité de Formation de Plaque. 24

r.s.e. Erreur Standard Relative. 18, 34, 50

RBD *Receptor Binding Domain*. 23, 24, 27, 51

s.e. Erreur Standard. 18

SAEM *Stochastic Approximation Expectation-Maximization*. 13–16, 32

sgRNA ARN subgénomique. 34

VPC *Visual Predictive Check*. 32, 34, 50, 51

Annexes

Annexe A

Taux de reproduction

Nous ajoutons deux exemples supplémentaires pour le calcul du taux de reproduction. Nous détaillons comment définir le sous-système infectieux, comment déterminer l'équilibre sans maladies et les matrices F et V. La dernière étape, consistant à trouver le rayon spectral, se réalise aisément en utilisant le logiciel MAPLE.

A.1 Modèle SEIR

Commençons avec le modèle *Susceptible-Exposed-Infected-Recovered* défini par le système suivant :

$$\begin{cases} \dot{S} &= \lambda - \mu S - \beta SI \\ \dot{E} &= \beta SI - kE - \mu E \\ \dot{I} &= kE - \gamma I - \mu I \\ \dot{R} &= \gamma I - \mu R \end{cases} \tag{A.1}$$

où λ représente le taux de production des cellules susceptibles, μ le taux de mortalité, β le taux d'infection des cellules susceptibles, k la vitesse à laquelle un individu latent devient infectieux (ou taux d'incubation) et γ est le taux de guérison.

Lorsque la population est sans maladie, notre système se réduit à une seule équation non triviale $\dot{S} = \lambda - \mu S$. A l'équilibre, lorsque $\dot{S} = 0$, nous obtenons que seulement la première coordonnée de x_0 est non nulle et est donnée par : $\hat{S} = \dfrac{\lambda}{\mu}$.

Nous définissons, maintenant, notre sous-système infectieux. Il doit nécessairement contenir les équations de E et I qui sont les compartiments infectés et nous pouvons ici nous limiter à ces deux équations. Ainsi, notre sous-système infectieux est défini par :

$$\begin{cases} \dot{E} &= \beta SI - kE - \mu E \\ \dot{I} &= kE - \gamma I - \mu I \end{cases} \tag{A.2}$$

Nous remarquons que le système est linéaire en E et I. Il nous reste à déterminer les fonction F_i, V_i et les matrices associées.

Dans la suite de cet exemple, l'indice 1 représente le compartiment E et l'indice 2 représente le compartiment I. Nous obtenons $F_1(\bar{x}) = \beta S x_2$ et $F_2(\bar{x}) = 0$. En effet, toutes les cellules arrivant dans le compartiment I ont déjà été infectées et ont été comptabilisées dans l'expression de F_1.

Nous avons de plus $V_1(\bar{x}) = (k+\mu)x_1$ et $V_2(\bar{x}) = (\gamma+\mu)x_2 - kx_1$. En dérivant et en remplaçant S par $\hat{S}$ la valeur à l'équilibre sans infection, nous obtenons les matrices F et V :

$$F = \begin{pmatrix} 0 & \beta\dfrac{\lambda}{\mu} \\ 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} k+\mu & 0 \\ -k & \gamma+\mu \end{pmatrix} \tag{A.3}$$

Puis, en utilisant par exemple le logiciel MAPLE, nous pouvons facilement obtenir le rayon spectral de FV^{-1}, ce qui nous donne $R_0 = \dfrac{\beta\lambda k}{\mu(\mu+k)(\mu+\gamma)}$.

A.2 Modèle de Malaria

Étudions désormais un exemple plus concret qui modélise les contaminations de la malaria entre les moustiques et les humains. Pour ces deux espèces H et M, nous avons deux compartiments possibles : susceptibles (H_S et M_S) et infectés (H_I et M_I). Le modèle est décrit par le système suivant :

$$\begin{cases} \dot{H_S} &= -\beta_{MH} M_I H_S + \alpha H_I \\ \dot{H_I} &= \beta_{MH} M_I H_S - (\mu_H + \alpha) H_I \\ \dot{M_S} &= -\beta_{HM} M_S H_I \\ \dot{M_I} &= \beta_{HM} M_S H_I - \mu_M M_I \end{cases} \tag{A.4}$$

Les humains susceptibles H_S sont infectés avec une efficacité β_{MH} par les moustiques porteurs du virus M_I. Nous supposons que les humains infectés meurent avec un taux μ_H et guérissent avec un taux α.
Les moustiques s'infectent avec une efficacité β_{HM} lorsqu'ils vont piquer des humains porteurs du virus H_I. Nous supposons que les moustiques ne peuvent pas quitter le compartiment infecté et meurent naturellement avec un taux μ_M.

Comme nous souhaitons nous concentrer sur les premières étapes de l'infection, nous supposons que les humains et moustiques n'ont pas le temps de se reproduire et que seule la population infectée peut décéder.
Ainsi, lorsque la population est sans maladie, notre système se réduit aux deux équations des compartiments susceptibles :

$$\begin{cases} \dot{H_S} &= 0 \\ \dot{M_S} &= 0 \end{cases} \tag{A.5}$$

Donc, $H_S(t) = H_S(0)$ et $M_S(t) = M_S(0)$ le nombre initial d'humains et de moustiques susceptibles. L'état d'équilibre sans maladie est donc $(H_S(0), 0, M_S(0), 0)$.

Nous définissons ensuite notre sous-système infectieux par :

$$\begin{cases} \dot{H_I} &= \beta_{MH} M_I H_S - (\mu_H + \alpha) H_I \\ \dot{M_I} &= \beta_{HM} M_S H_I - \mu_M M_I \end{cases} \tag{A.6}$$

Nous remarquons de nouveau que ce système est linéaire en H_I et M_I. Il nous reste à déterminer les fonctions F_i, V_i et les matrices associées.

Dans la suite de cet exemple, l'indice 1 représente le compartiment H_I et l'indice 2 le compartiment M_I. Nous obtenons $F_1(\bar{x}) = \beta_{MH} H_S x_2$ et $F_2(\bar{x}) = \beta_{HM} M_S x_1$. Nous remarquons la

différence avec l'exemple précédent où une seule fonction F_i était non nulle. Cela s'explique par le fait que nous avons ici deux populations distinctes (humains et moustiques) qui peuvent être infectées.

Nous avons aussi $V_1(\bar{x}) = (\mu_H + \alpha)x_1$ et $V_2(\bar{x}) = \mu_M x_2$. En dérivant et en remplaçant H_S et M_S par les états d'équilibre sans maladie, nous obtenons les matrices F et V :

$$F = \begin{pmatrix} 0 & \beta_{MH} H_S(0) \\ \beta_{HM} M_S(0) & 0 \end{pmatrix} \qquad V = \begin{pmatrix} \mu_H + \alpha & 0 \\ 0 & \mu_M \end{pmatrix} \tag{A.7}$$

De la même manière, nous obtenons le rayon spectral de FV^{-1}, ce qui nous donne : $R_0 = \sqrt{\dfrac{\beta_{MH}\beta_{HM}H_S(0)M_S(0)}{\mu_M(\mu_H + \alpha)}}$

Lorsque nous avons plusieurs classes de populations infectées, cette méthode nous donne le nombre moyen de nouvelles infections par un individu infecté de n'importe quelle classe. Il existe une autre méthode qui donne le nombre de nouvelles infections au sein de la même classe que la personne infectée.

Annexe B

Résultats de l'approche variationnelle

B.1 Estimation du *modèle de référence* et de la section 2.2.1

FIGURE B.1 – Estimations des paramètres de la population et des erreurs standards du *modèle de référence*.

	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-7.15	0.33	4.60
delta_N_pop	1.46	0.22	15.4
c_pop	3		
cl_pop	20		
k_pop	3		
mu_pop	0.001		
P_N_pop	13765.62	6328.28	46.0
f_P_T_pop	-2.56	0.42	16.4
alpha_VLSG_pop	1.48	0.14	9.23
thresh_Weight_pop	4.5		
Standard Deviation of the Random Effects			
omega_beta_N_pow	1.02	0.19	19.0
omega_delta_N	0.43	0.13	30.4
Error Model Parameters			
sigma_g_T	1.29	0.13	9.81
sigma_sg_T	1.36	0.2	14.6
sigma_g_N	1.03	0.095	9.27
sigma_sg_N	1.61	0.26	16.1

FIGURE B.2 – Estimations des paramètres de la population et des erreurs standards du modèle de la section 2.2.1.

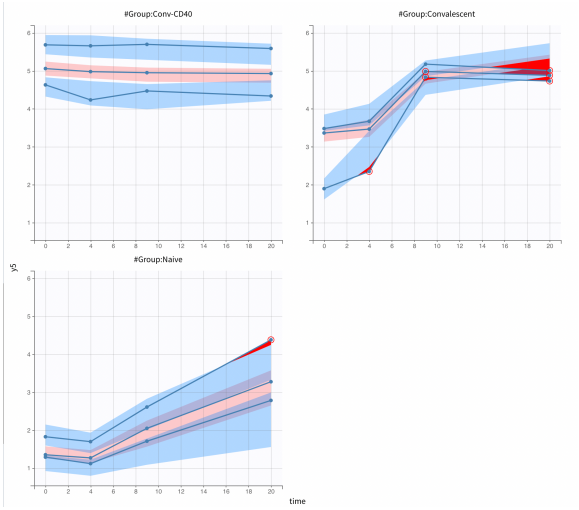
	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-7.16	0.33	4.64
delta_N_pop	1.44	0.22	15.5
c_pop	3		
cl_pop	15		
k_pop	3		
mu_pop	0.001		
P_N_pop	10092.14	4510.12	44.7
fact_P_T_pop	-2.55	0.76	29.6
alpha_VLSG_pop	1.43	0.11	7.98
thresh_Weight_pop	4.5		
Standard Deviation of the Random Effects			
omega_beta_N_pow	1.12	0.23	20.2
omega_delta_N	0.39	0.12	30.9
Error Model Parameters			
sigma_g_T	1.31	0.14	10.5
sigma_sg_T	1.28	0.19	14.5
sigma_g_N	1.03	0.1	9.78
sigma_sg_N	1.52	0.22	14.2

B.2 Estimation du modèle défini section 2.2.2

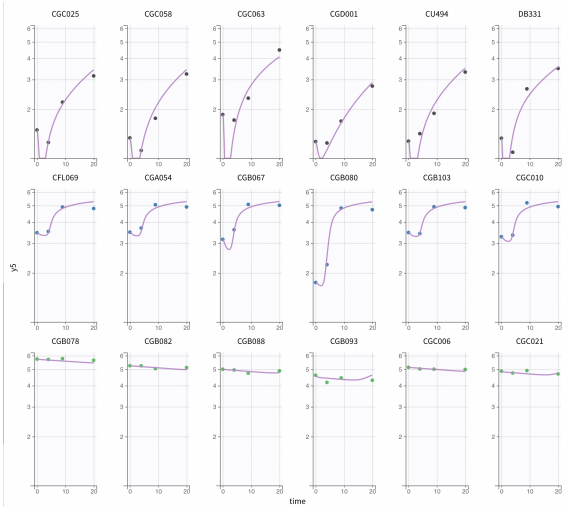
FIGURE B.3 – Estimation des paramètres de la population et des erreurs standards.

	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-5.85	0.86	14.7
fact_beta_T_pop	0		
fact_delta_T_pop	0		
delta_N_pop	1.04	0.99	96.2
c_pop	3		
cl_pop	20		
k_pop	3		
mu_pop	0.001		
P_N_pop	1260.15	736.66	58.5
fact_P_T_pop	0.00003	0.000023	77.9
alpha_VLSG_pop	1.1	0.17	15.9
gamma_pop	0.00001		
kAb_pop	0.16	0.013	7.85
rAb_pop	0.43	0.021	4.99
beta_rAb_groupVSnaive_G_Conv_CD40	0.56	0.17	29.8
beta_rAb_groupVSnaive_G_Convalescent	1.87	0.21	11.5
Abmax_pop	100000		
sigma_pop	0		
delta_Ab_pop	0.033		
eta_pop	0.00000000011	0.00000000013	120
ksi_pop	0.0003	0.00034	114
n_pop	2.89	0.13	4.47
A_pop	251169.59	15622.46	6.22
B_pop	57.28	4.05	7.07
thresh_Weight_pop	4.5		
Standard Deviation of the Random Effects			
omega_beta_N_pow	0.78	0.21	26.9
omega_delta_N	0.4	1.56	387
omega_rAb	0.051	0.048	93.9
Error Model Parameters			
sigma_g_T	1.34	0.17	12.4
sigma_sg_T	1.4	0.25	17.9
sigma_g_N	1.27	0.21	16.3
sigma_sg_N	1.46	0.22	14.8
sigma_ab	0.19	0.019	9.67
sigma_ecl	0.12	0.012	9.69

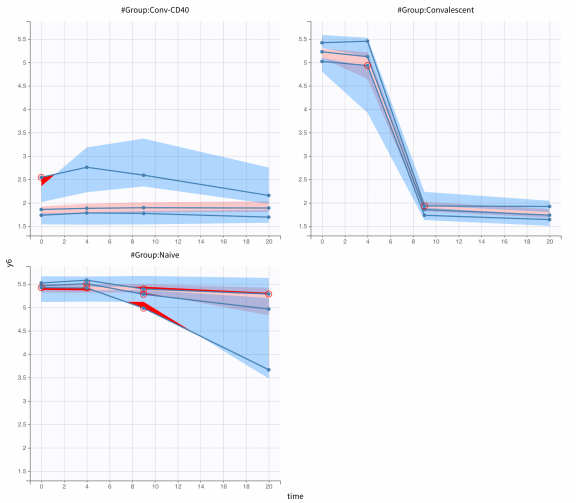
FIGURE B.4 – Graphiques du modèle défini section 2.2.2.



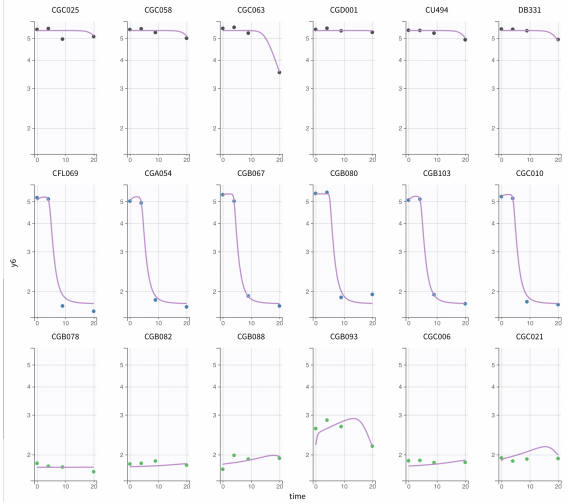
(a) VPC des IgG RBD par groupes.



(b) Graphiques individuels des IgG RBD.



(c) VPC des ECL RBD par groupes.



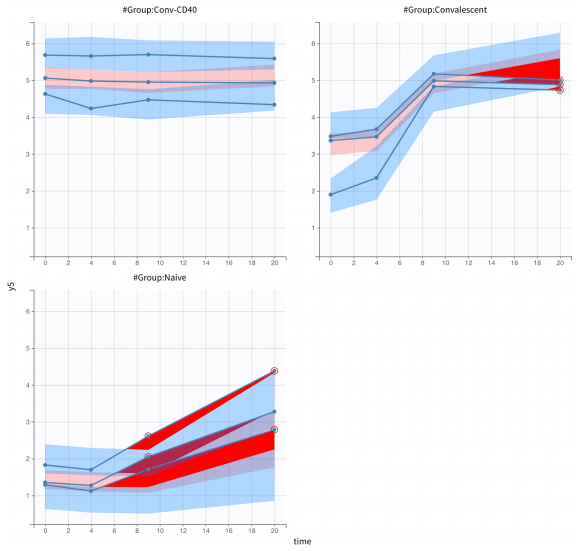
(d) Graphiques individuels des ECL RBD.

B.3 Estimation du modèle défini section 2.2.3

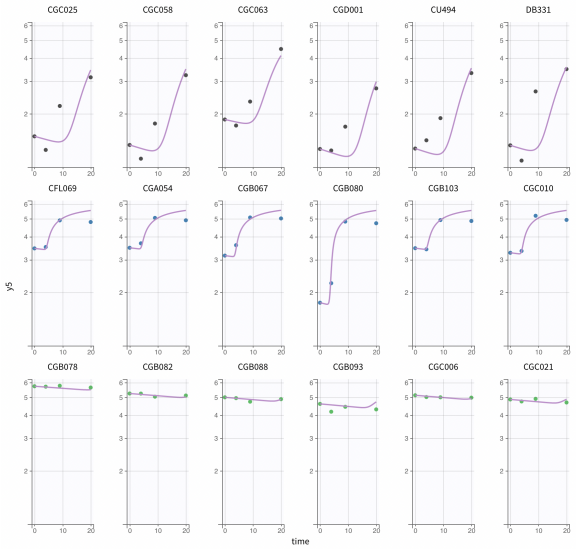
FIGURE B.5 – Estimation des paramètres de la population et des erreurs standards.

	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-6.91	0.44	6.35
delta_N_pop	1.36	0.18	13.2
fact_beta_T_pop	0		
fact_delta_T_pop	0		
c_pop	3		
cl_pop	20		
k_pop	3		
mu_pop	0.001		
P_N_pop	7330.51	5109.05	69.7
fact_P_T_pop	-2.56	0.44	17.3
alpha_VLSG_pop	1.31	0.16	12.2
gamma_pop	0.00001		
kAb_pop	0.35	0.032	9.18
rAb_pop	1.04	0.042	4.01
beta_rAb_Group2_A2	1.63	0.057	3.53
beta_rAb_Group2_A3	0.37	0.18	47.6
Abmax_pop	100000		
eta_pop	0.00000000059	0.00000000091	155
delta_Ab_pop	0.033		
A_pop	252419.06	15074.68	5.97
B_pop	54.61	3.7	6.77
thresh_Weight_pop	4.5		
n_pop	2.69	0.16	6.12
H_pop	0.000023	0.00055	2.33e+3
Standard Deviation of the Random Effects			
omega_beta_N_pow	1.12	0.23	20.2
omega_delta_N	0.34	0.13	38.4
omega_rAb	0.021	0.019	91.7
Error Model Parameters			
sigma_g_T	1.31	0.13	10.1
sigma_sg_T	1.32	0.19	14.5
sigma_g_N	1.02	0.1	10.3
sigma_sg_N	1.52	0.22	14.5
sigma_ab	0.33	0.029	8.69
sigma_ecl	0.12	0.011	9.64

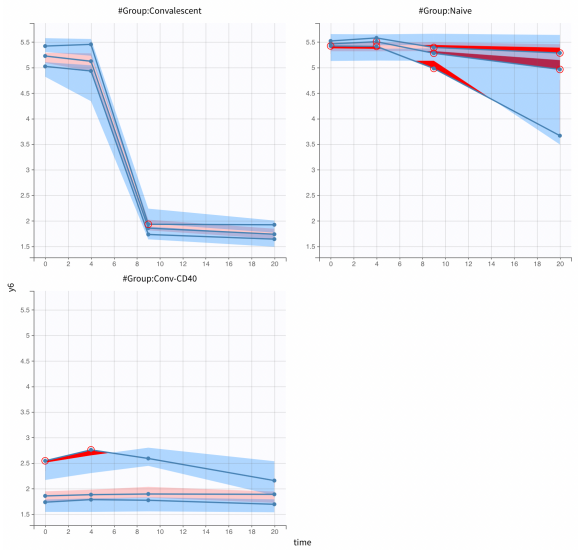
FIGURE B.6 – Graphiques du modèle défini section 2.2.3.



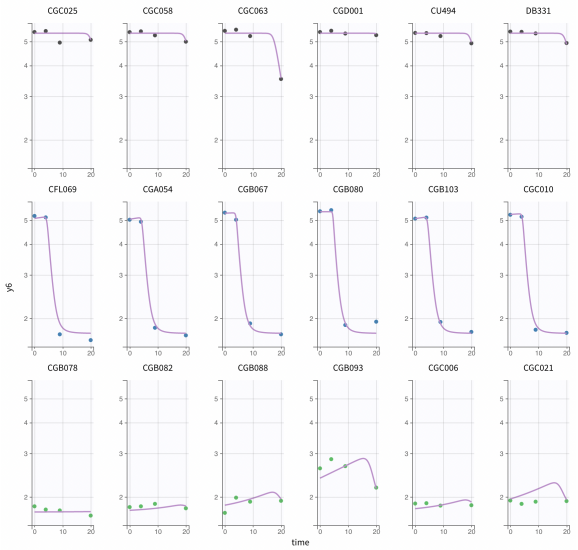
(a) VPC des IgG RBD par groupes.



(b) Graphiques individuels des IgG RBD.



(c) VPC des ECL RBD par groupes.



(d) Graphiques individuels des ECL RBD.

B.4 Estimation du modèle 2.2.2 modifié

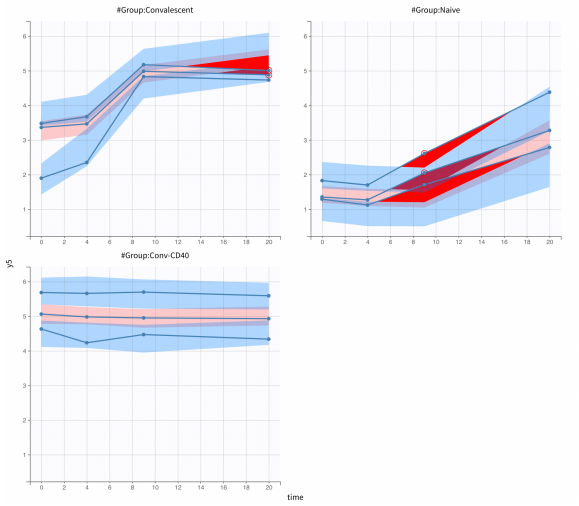
Ici, nous avons remplacé l'équation différentielle ordinaire de Ab_1 dans le modèle (2.6) par :

$$\dot{Ab_1} = \gamma I_2 + r_A \left(1 - \frac{Ab_1}{Ab_{max}}\right) Ab_1 - k_{Ab} Ab_1 \tag{B.1}$$

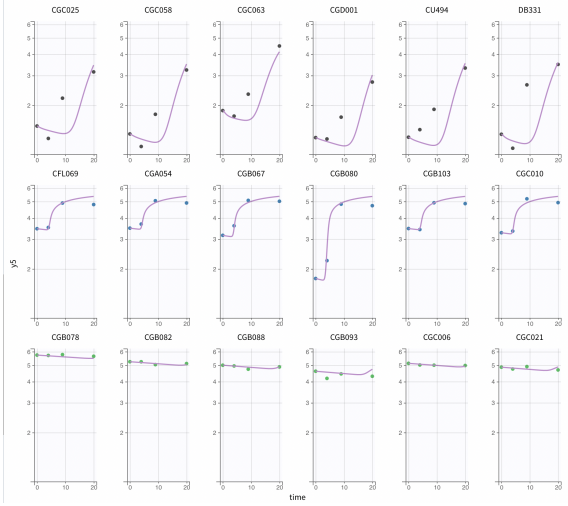
FIGURE B.7 – Estimation des paramètres de la population et des erreurs standards.

	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-7.24	0.21	2.87
fact_beta_T_pop	0		
fact_delta_T_pop	0		
delta_N_pop	1.05	0.18	16.8
c_pop	3		
cl_pop	20		
k_pop	3		
mu_pop	0.001		
P_N_pop	49062.54	19939.04	40.6
fact_P_T_pop	-2.89	0.45	15.6
alpha_VLSG_pop	1.75	0.17	9.75
gamma_pop	0.00001		
kAb_pop	0.19	0.027	13.6
rAb_pop	0.96	0.028	2.95
beta_rAb_groupVSnaive_G_Conv_CD40	0.31	0.044	14.4
beta_rAb_groupVSnaive_G_Convalescent	1.64	0.042	2.58
Abmax_pop	100000		
sigma_pop	0		
delta_Ab_pop	0.033		
eta_pop	0.00000000048	0.00000000051	106
ksi_pop	0.0001	0.000094	90.0
n_pop	2.7	0.1	3.88
A_pop	252374.84	14804.77	5.87
B_pop	54.1	3.72	6.88
thresh_Weight_pop	4.5		
Standard Deviation of the Random Effects			
omega_beta_N_pow	0.55	0.14	25.4
omega_delta_N	0.31	0.17	53.7
omega_rAb	0.025	0.025	98.1
Error Model Parameters			
sigma_g_T	1.3	0.13	10.0
sigma_sg_T	1.53	0.25	16.0
sigma_g_N	1.09	0.1	9.50
sigma_sg_N	1.91	0.31	16.4
sigma_ab	0.32	0.029	8.93
sigma_ecl	0.12	0.011	9.45

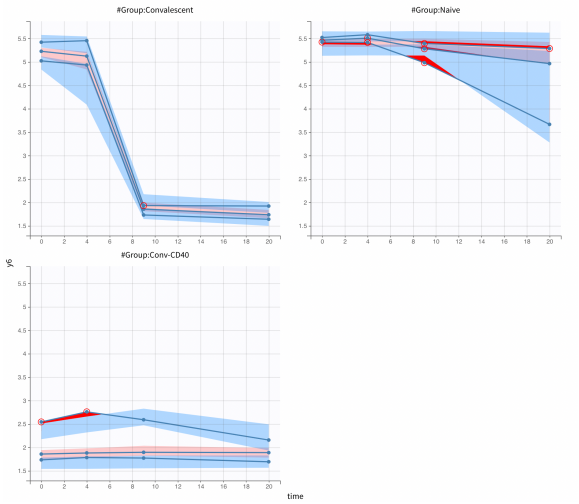
FIGURE B.8 – Graphiques du modèle défini section 2.2.2 modifié.



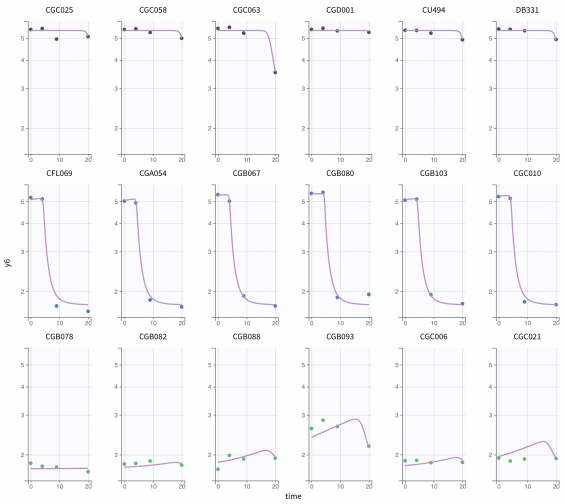
(a) VPC des IgG RBD par groupes.



(b) Graphiques individuels des IgG RBD.



(c) VPC des ECL RBD par groupes.



(d) Graphiques individuels des ECL RBD.

B.5 Estimation du modèle 2.2.3 modifié

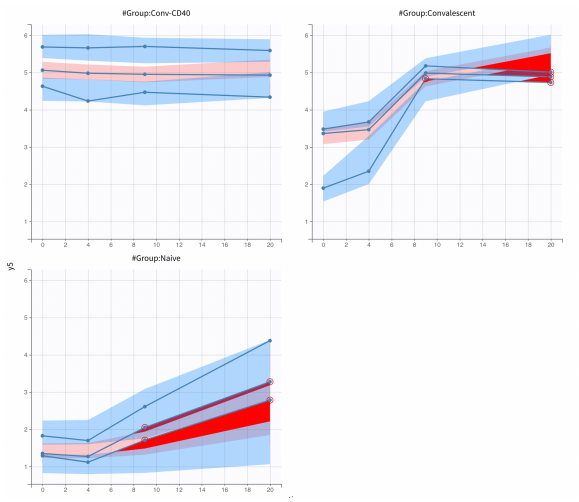
Ici, nous avons remplacé l'équation différentielle ordinaire de M_1 dans le modèle (2.9) par :

$$\dot{M}_1 = \gamma V + r_A \left(1 - \frac{M_1}{Ab_{max}}\right) M_1 - k_{Ab} M_1 \tag{B.2}$$

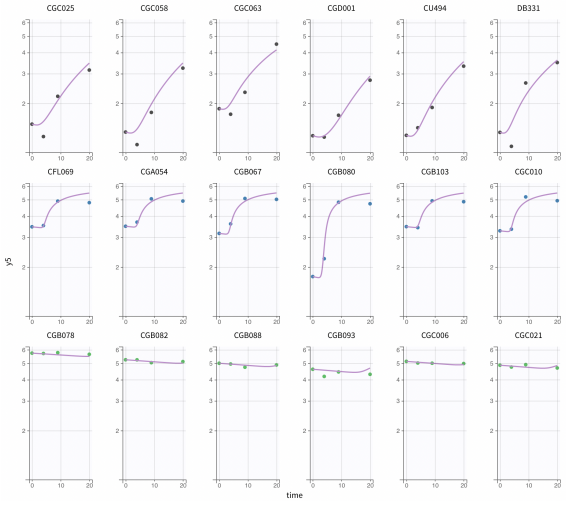
FIGURE B.9 – Estimation des paramètres de la population et des erreurs standards.

	VALUE	STOCH. APPROX.	
		S.E.	R.S.E.(%)
Fixed Effects			
beta_N_pow_pop	-6.95	0.28	3.96
delta_N_pop	1.48	0.25	17.2
fact_beta_T_pop	0		
fact_delta_T_pop	0		
c_pop	3		
cl_pop	20		
k_pop	3		
mu_pop	0.001		
P_N_pop	10367.9	2967.29	28.6
fact_P_T_pop	-2.58	0.44	16.9
alpha_VLSG_pop	1.43	0.11	7.97
gamma_pop	0.00001		
kAb_pop	0.31	0.029	9.59
rAb_pop	0.61	0.034	5.56
beta_rAb_Group2_A2	1.83	0.11	5.96
beta_rAb_Group2_A3	0.57	0.071	12.6
Abmax_pop	100000		
eta_pop	0.00000000018	0.00000000022	124
delta_Ab_pop	0.033		
A_pop	249715.08	15944.77	6.39
B_pop	55.06	3.89	7.07
thresh_Weight_pop	4.5		
n_pop	2.81	0.12	4.34
H_pop	0.025		
Standard Deviation of the Random Effects			
omega_beta_N_pow	0.93	0.2	21.0
omega_delta_N	0.49	0.14	29.0
omega_rAb	0.035	0.05	144
Error Model Parameters			
sigma_g_T	1.33	0.13	9.93
sigma_sg_T	1.4	0.2	14.3
sigma_g_N	0.99	0.088	8.83
sigma_sg_N	1.57	0.22	14.1
sigma_ab	0.24	0.023	9.46
sigma_ecl	0.12	0.013	10.2

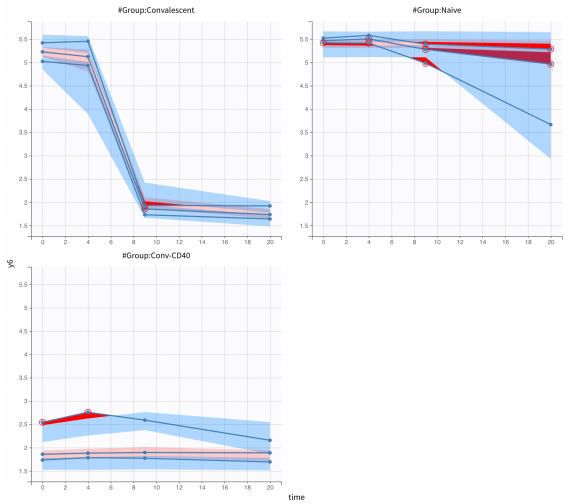
FIGURE B.10 – Graphiques du modèle défini section 2.2.3 modifié.



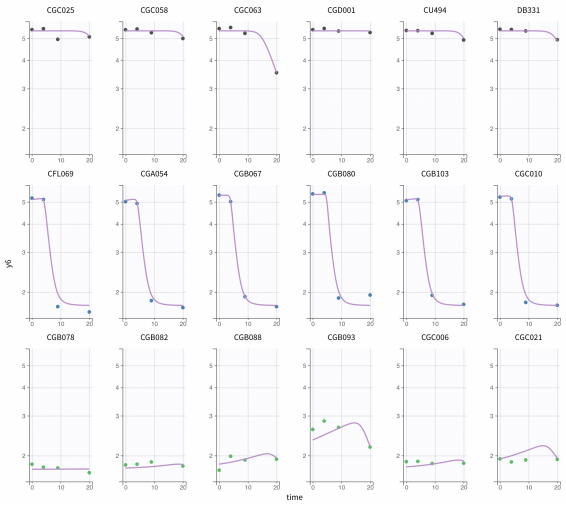
(a) VPC des IgG RBD par groupes.



(b) Graphiques individuels des IgG RBD.



(c) VPC des ECL RBD par groupes.



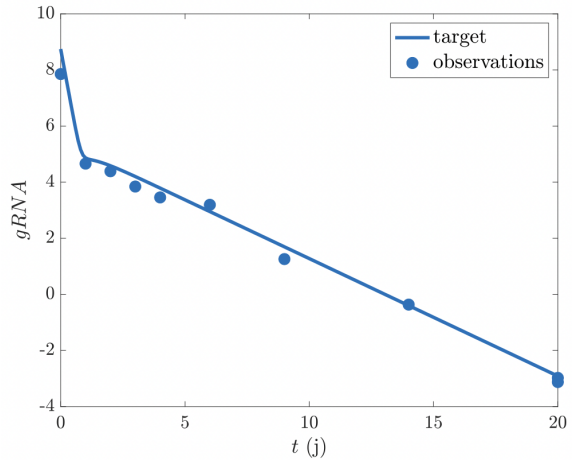
(d) Graphiques individuels des ECL RBD.

Annexe C

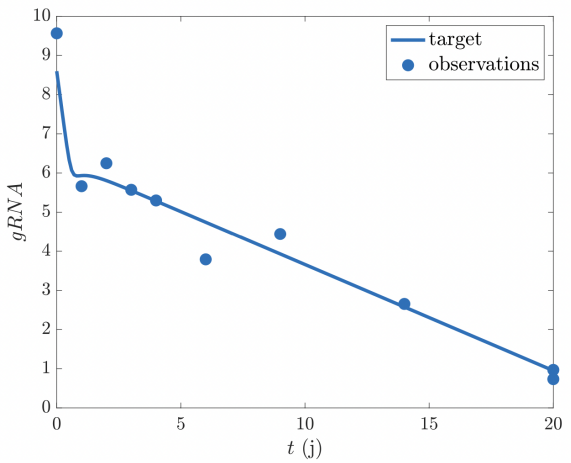
Résultats de l'approche séquentielle

C.1 Runge-Kutta d'ordre 4 explicite

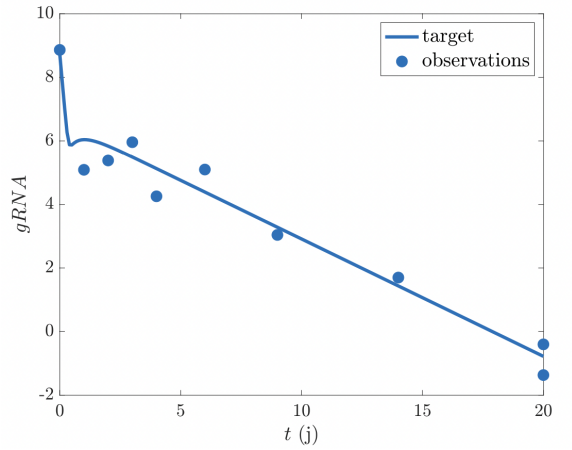
FIGURE C.1 – Runge-Kutta 4 explicite avec $\beta_{pop} = 10^{-5,44}$ et $\delta_{pop} = 0,85$



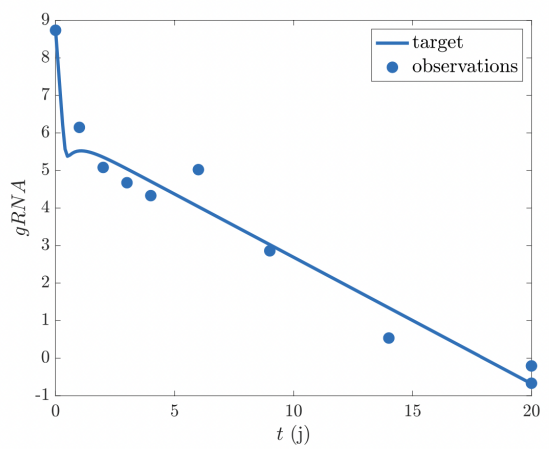
(a) Pas de temps $0,1$.



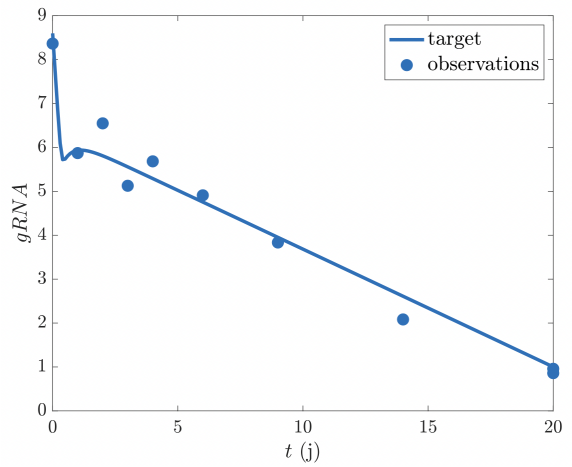
(b) Pas de temps $0,1$.



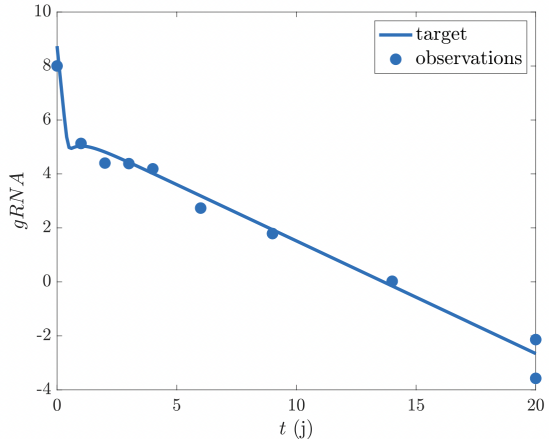
(c) Pas de temps $0,05$.



(d) Pas de temps $0,05$.

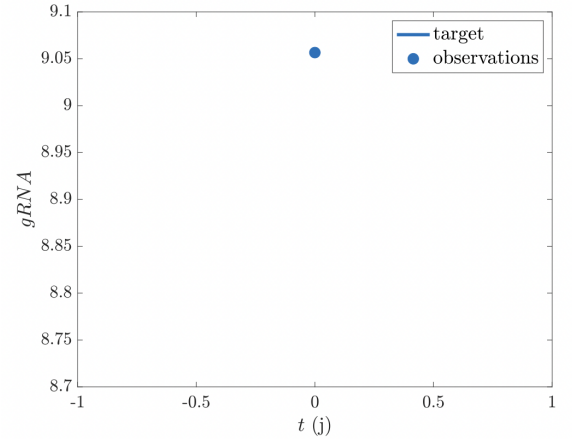


(e) Pas de temps $0,01$.

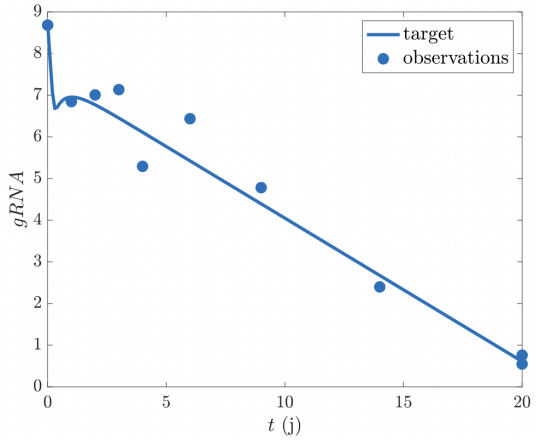


(f) Pas de temps $0,01$.

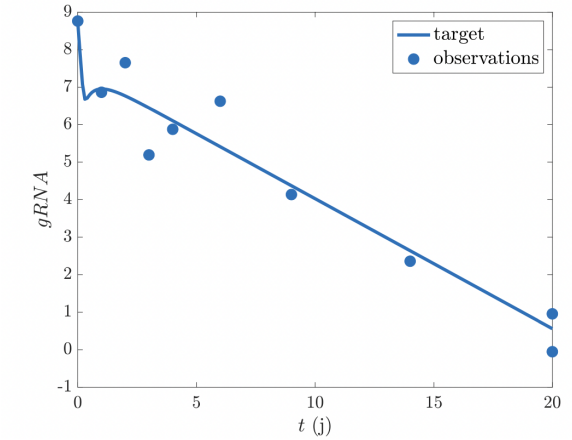
FIGURE C.2 – Runge-Kutta 4 explicite.



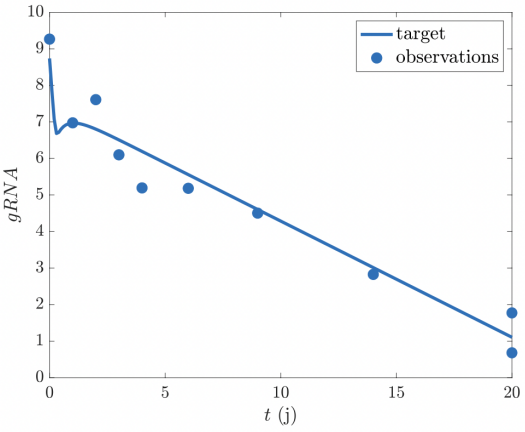
(a) $\beta_{pop} = 10^{-2}$ et $\delta_{pop} = 1$
Pas de temps $0,001$.



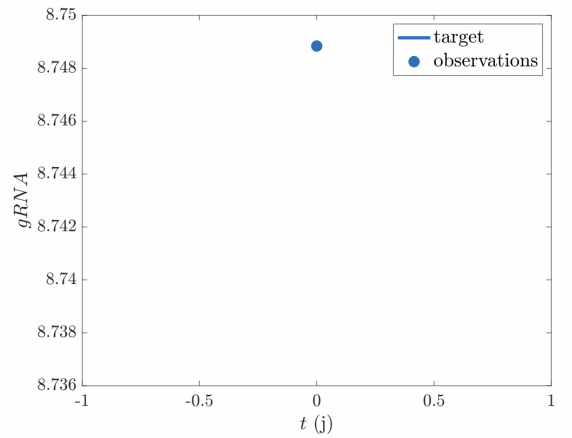
(b) $\beta_{pop} = 10^{-2}$ et $\delta_{pop} = 1$
Pas de temps $0,001$.



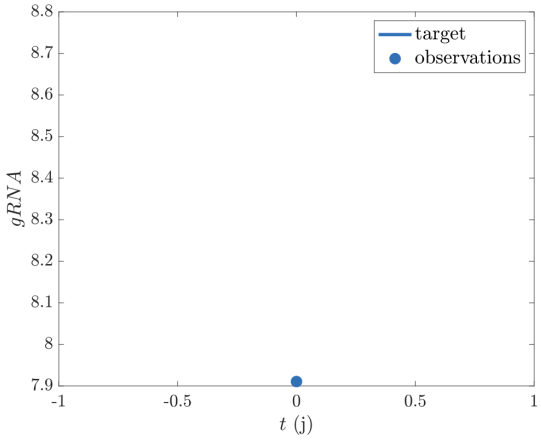
(c) $\beta_{pop} = 10^{-2}$ et $\delta_{pop} = 1$
Pas de temps 5×10^{-4}.



(d) $\beta_{pop} = 10^{-2}$ et $\delta_{pop} = 1$
Pas de temps 5×10^{-4}.



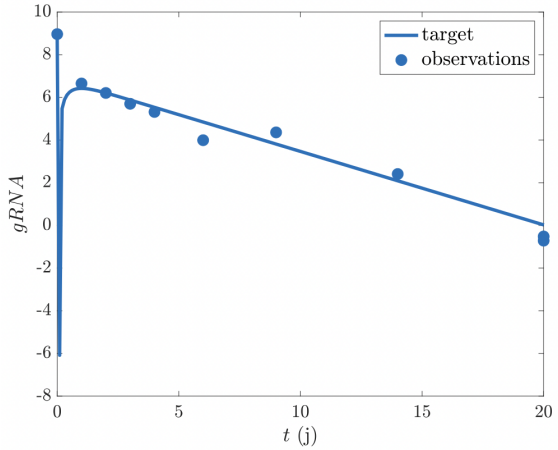
(e) $\beta = 10^{-2}$ et $\delta = 1$
Pas de temps 1×10^{-4}



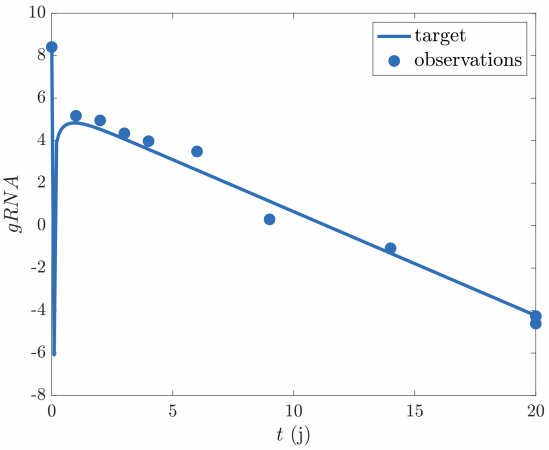
(f) $\beta_{pop} = 1$ et $\delta_{pop} = 1$
Pas de temps 5×10^{-4}.

C.2 Euler explicite

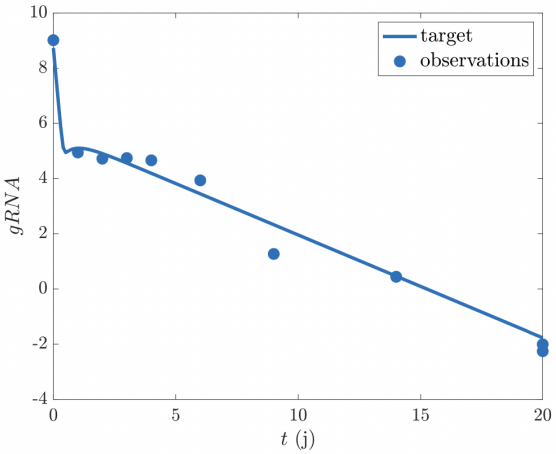
FIGURE C.3 – Euler explicite avec $\beta_{pop} = 10^{-5,44}$ et $\delta_{pop} = 0,85$



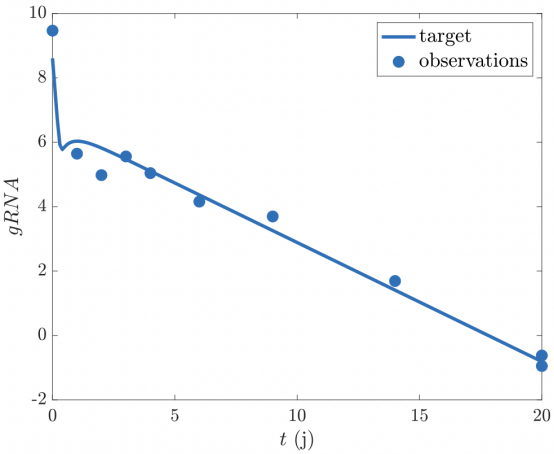
(a) Pas de temps $0,1$.



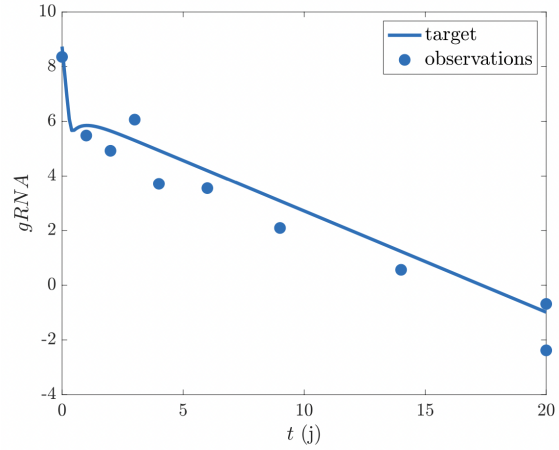
(b) Pas de temps $0,1$.



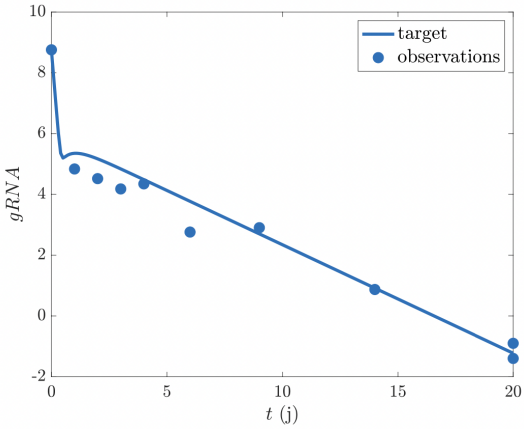
(c) Pas de temps $0,05$.



(d) Pas de temps $0,05$.

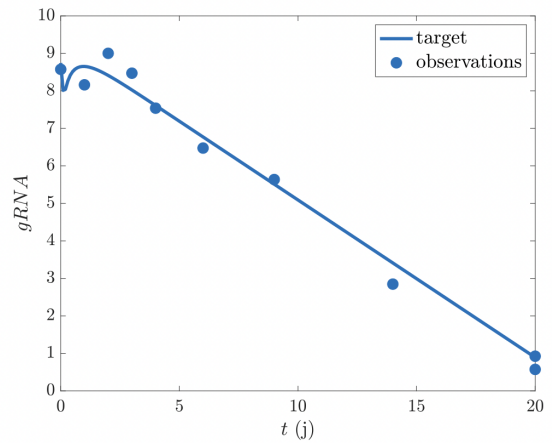


(e) Pas de temps $0,01$.

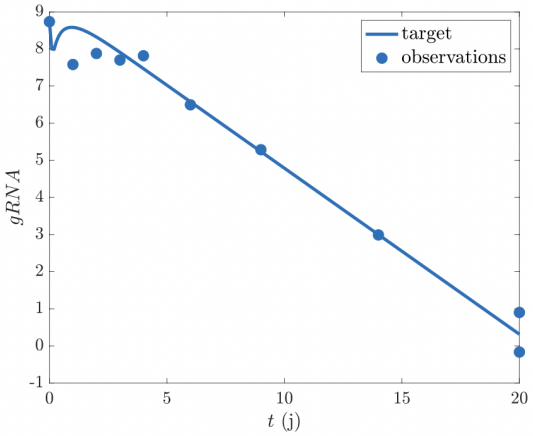


(f) Pas de temps $0,01$.

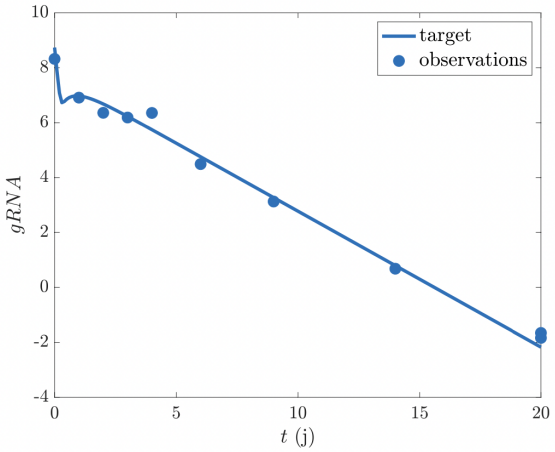
FIGURE C.4 – Euler explicite avec $\beta_{pop} = 10^{-2}$ et $\delta_{pop} = 1$



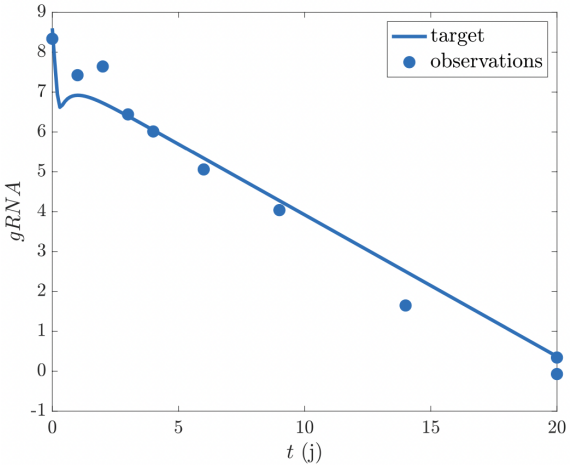
(a) Pas de temps 10^{-3}.



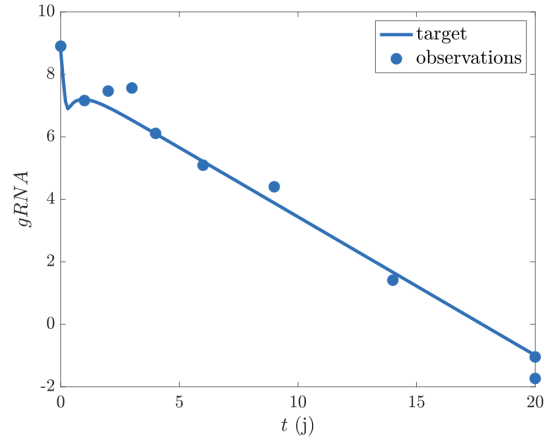
(b) Pas de temps 10^{-3}.



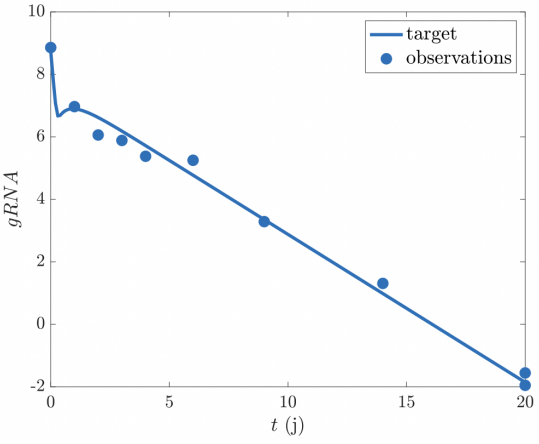
(c) Pas de temps 5×10^{-4}.



(d) Pas de temps 5×10^{-4}.

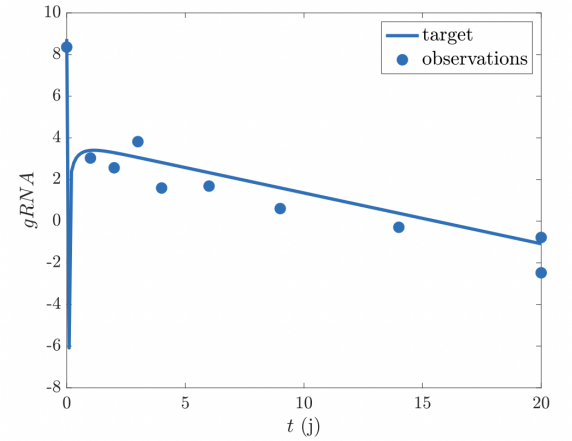


(e) Pas de temps 10^{-4}.

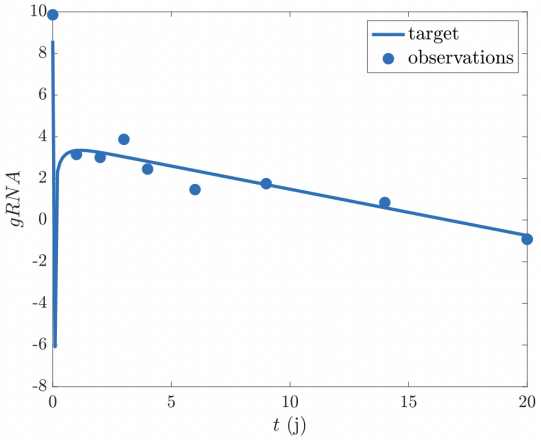


(f) Pas de temps 10^{-4}.

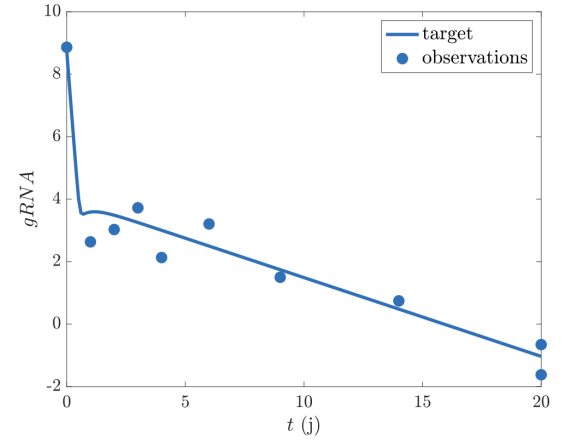
FIGURE C.5 – Euler explicite avec $\beta_{pop} = 10^{-8}$ et $\delta_{pop} = 0,85$



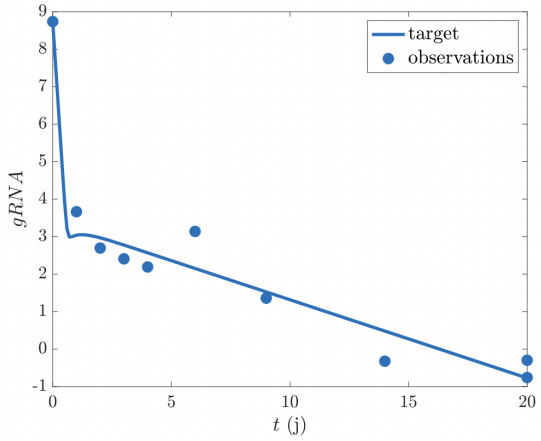
(a) Pas de temps $0,05$.



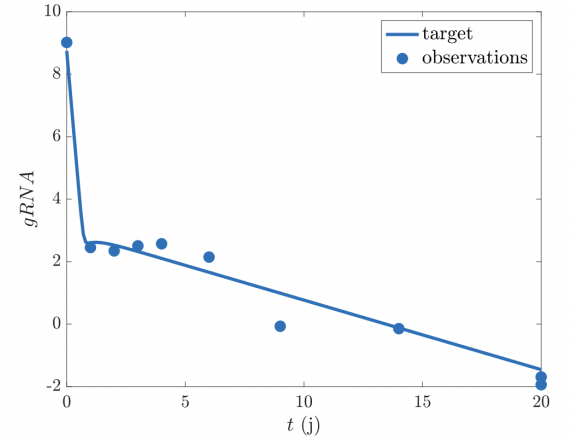
(b) Pas de temps $0,05$.



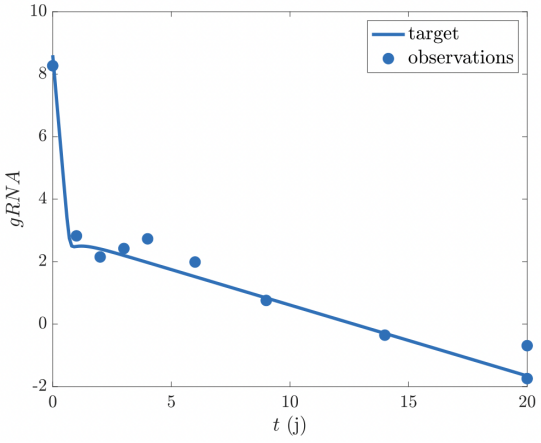
(c) Pas de temps $0,01$.



(d) Pas de temps $0,01$.



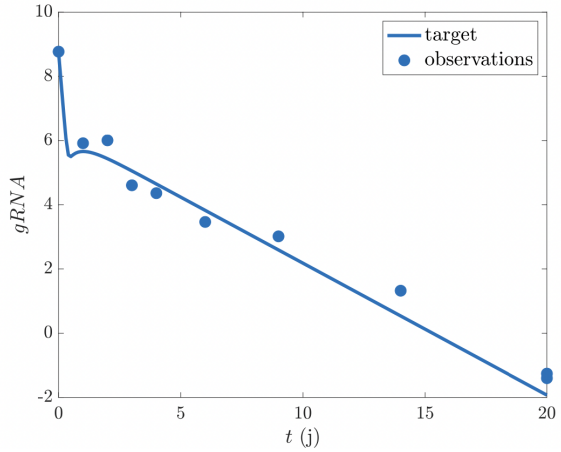
(e) Pas de temps 5×10^{-4}.



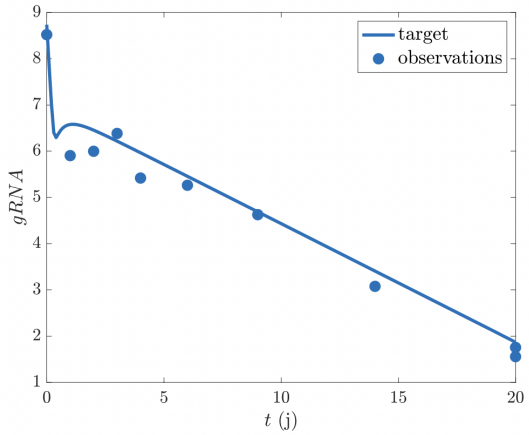
(f) Pas de temps 5×10^{-4}.

C.3 Runge-Kutta d'ordre 3 implicite

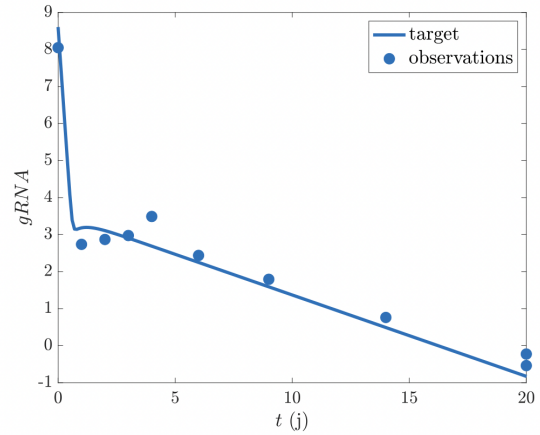
FIGURE C.6 – Runge-Kutta d'ordre 3 implicite.



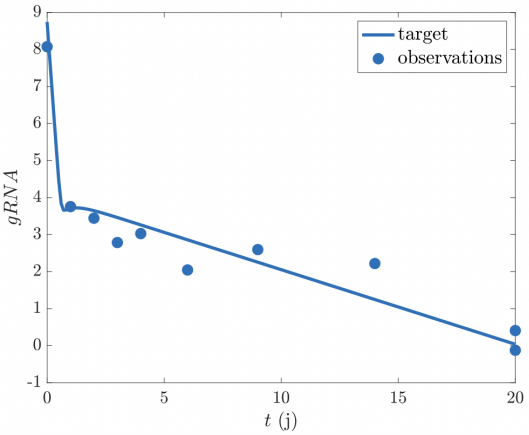
(a) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,05$.



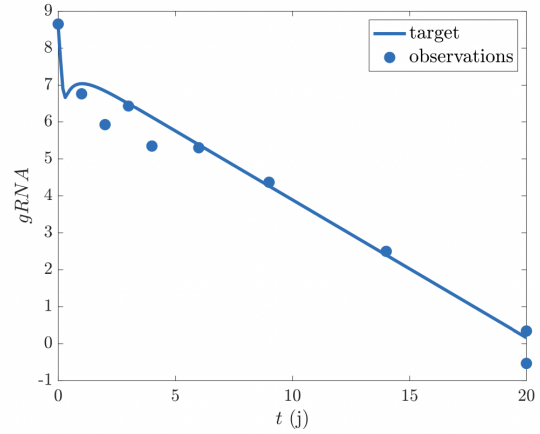
(b) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,01$.



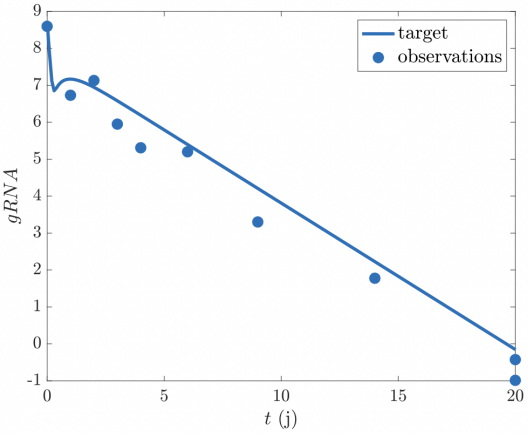
(c) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,05$.



(d) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,01$.



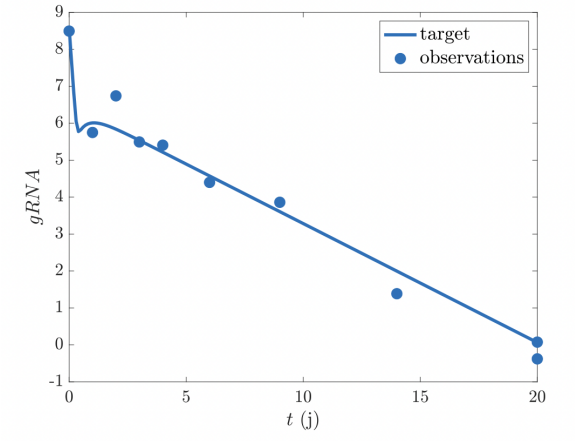
(e) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,05$.



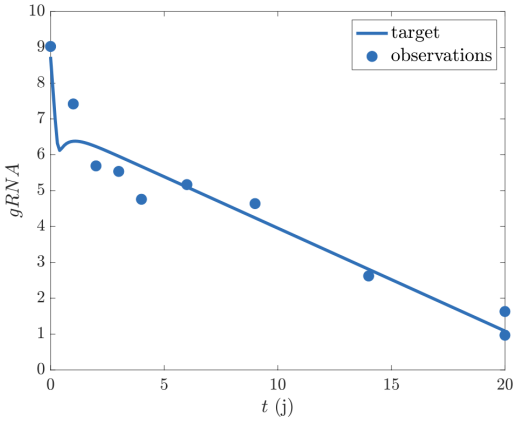
(f) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,01$.

C.4 Runge-Kutta d'ordre 4 implicite

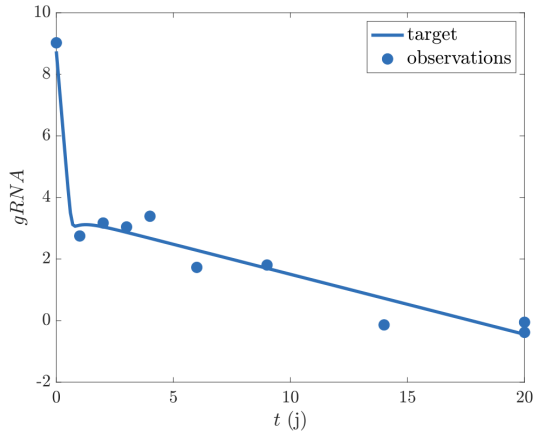
FIGURE C.7 – Runge-Kutta d'ordre 4 implicite.



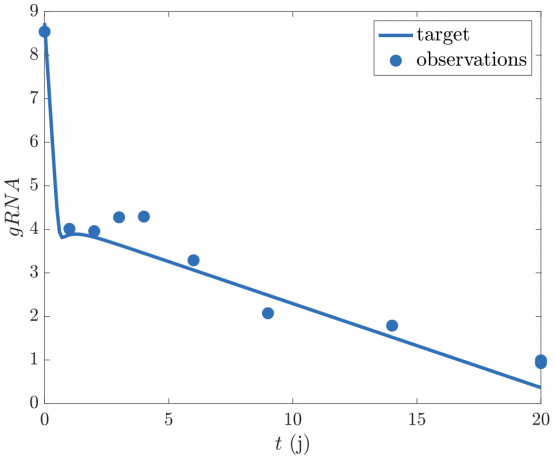
(a) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,05$.



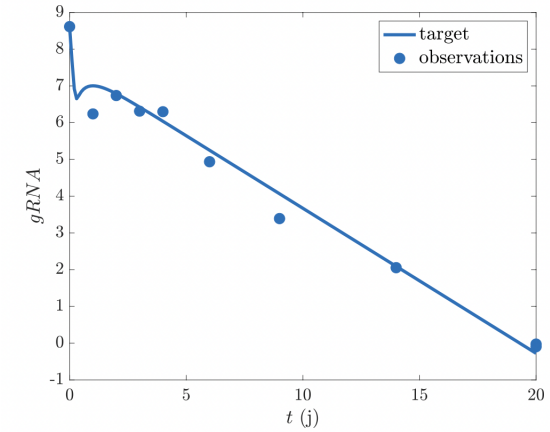
(b) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,01$.



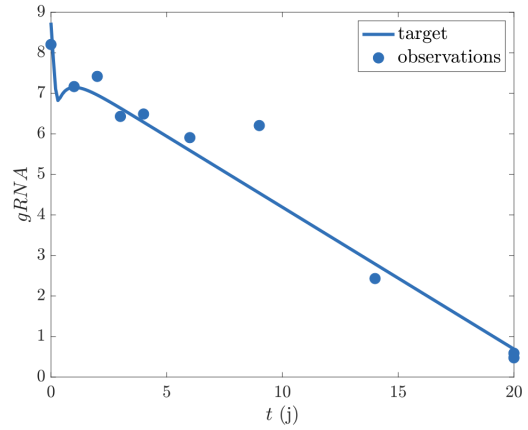
(c) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,05$.



(d) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,01$.



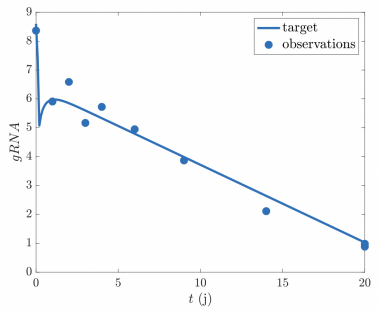
(e) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,05$.



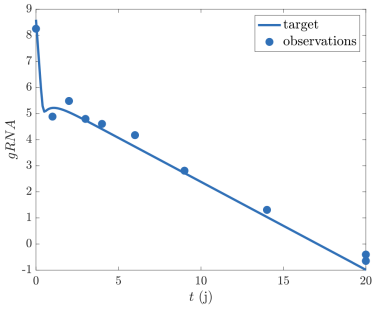
(f) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,01$.

C.5 Crank-Nicolson et BDF d'ordre 2

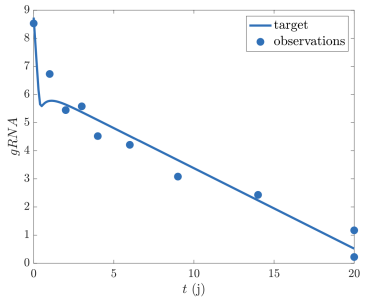
FIGURE C.8 – Crank-Nicolson et BDF d'ordre 2



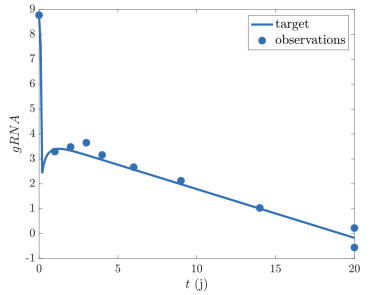
(a) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,05$.



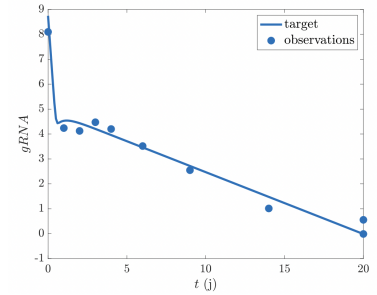
(b) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,01$.



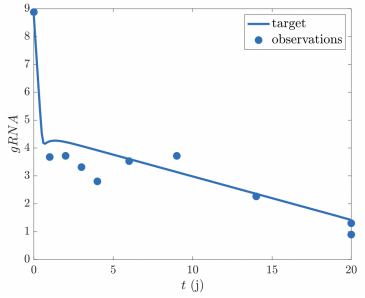
(c) $\delta_{pop} = 10^{-5,44}$ et $\delta = 0,85$
Pas de temps $0,005$.



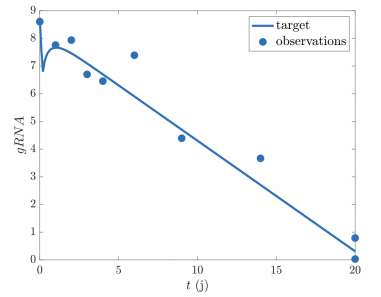
(d) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,05$.



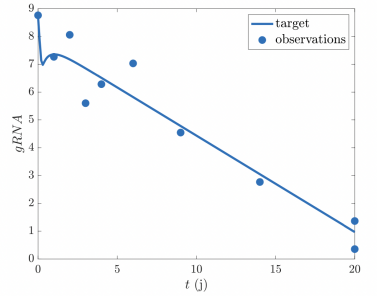
(e) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,01$.



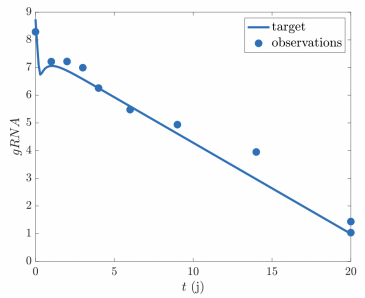
(f) $\delta_{pop} = 10^{-8}$ et $\delta = 0,5$
Pas de temps $0,005$.



(g) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,05$.



(h) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,01$.



(i) $\delta_{pop} = 10^{-2}$ et $\delta = 1$
Pas de temps $0,005$.