UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

Spécialité :

BIOSTATISTIQUES

Présentée par

Mme Adeline Samson

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PARIS 6

Sujet de la thèse :

Estimation dans les modèles non-linéaires à effets mixtes : extensions de l'algorithme SAEM pour l'analyse de la dynamique virale sous traitement anti-VIH

soutenue le 19 mai 2006

devant le jury composé de :

Monsieur Jean-Marc AZAIS	Rapporteur
Monsieur Daniel COMMENGES	Rapporteur
Monsieur Philippe FLANDRE	Examinateur
Monsieur Marc LAVIELLE	Directeur de thèse
Madame France MENTRÉ	Directeur de thèse
Monsieur Alain-Jacques VALLERON	Président du jury

École Doctorale 393

SANTÉ PUBLIQUE : ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE

Remerciements

- Au Professeur Alain-Jacques Valleron, qui me fait l'honneur de participer à l'évaluation de cette thèse. Je vous remercie de m'avoir fait confiance et de m'avoir accordé un financement lorsque nous vous avons soumis ce sujet de thèse à l'interface entre modélisation, statistique et épidémiologie clinique. Je vous remercie également pour les discussions fructueuses et enrichissantes que nous avons eues, en particulier à l'occasion des séminaires de l'école doctorale.
- A Daniel Commenges et Jean-Marc Azais, pour avoir accepté d'examiner mes travaux en qualité de rapporteurs. J'espère que l'honneur que vous me faites sera compensé par l'intérêt que vous pourrez porter à ce manuscrit.
- A Philippe Flandre, de me faire l'honneur de participer à mon jury de thèse. Merci d'avoir accepté d'examiner mes travaux.
- A France Mentré, directrice de mes travaux de recherche, qui m'a accueillie lors de mon stage de DEA puis dans le cadre de ma thèse. Je te remercie d'avoir su me proposer un sujet sur mesure si passionnant, alliant développements statistiques théoriques et application pratique de ces méthodes. Merci de m'avoir encouragée et fait confiance en acceptant que ma première année de thèse soit bousculée par mon second DEA. Merci également de m'avoir fait confiance en me laissant naviguer librement entre Bichat et Orsay et en établissant cette collaboration bénéfique avec Marc.
- A Marc Lavielle, co-directeur de mes travaux de recherche. Je te remercie de m'avoir accueillie et de m'avoir fait confiance en me proposant ces différents axes de recherche si passionnants et motivants. Merci pour ton ouverture vers les Sciences de la Vie et la Biologie Humaine, qui représente pour moi l'avenir de la recherche en statistiques, à la fois moderne et rationnelle, et largement tournée vers des collaborations élargies. Merci pour ton dynamisme si impressionnant et communicatif.
- A Sacha Tsybakov, pour m'avoir accepté dans votre DEA. Je vous remercie de votre soutien et de votre confiance qui sont très importants pour moi. Merci de votre ouverture et de l'intérêt que vous avez toujours porté à mon sujet de recherche, à l'interface entre statistiques et biostatistiques.
- A Philippe Ravaud, pour son accueil au sein du Département d'épidémiologie, biostatistiques et recherche clinique de l'hôpital Bichat. Je vous remercie pour votre soutien permanent et vos encouragements tout au long de ma thèse. Merci d'exiger autant et toujours plus de nous, de savoir aiguiser nos sens critiques et nos appétits scientifiques.
- A Sophie, pour ces années de collaboration quotidienne (ou quasiment). Je te

remercie pour tes compétences, ta confiance, ta patience et ton amitié, qui ont fait de ces années de thèse des années si agréables et constructives. Merci de ton soutien constant dans les moments difficiles et de doute. Merci aussi pour tous les moments partagés en dehors du cadre de la thèse ! J'espère que nous aurons encore l'occasion de travailler ensemble.

- A Xavière, pour nos diverses collaborations. Je te remercie également pour ta motivation, tes compétences, ta patience et ta pédagogie à répondre à toutes (sans exception) mes questions dans des domaines si divers ! Merci de ton soutien quotidien et de ton amitié. Nous serons amenées à continuer à travailler ensemble, et j'en suis ravie.
- A Ségolène, Sophie, Xavière, Odile, Jérôme, Karl, Gabriel, Vincent et Alexis, mes relecteurs attentionnés d'articles et de thèse, anglais ou francais. Merci d'avoir accepté de m'aider en prenant sur votre temps libre, merci de votre amitié si franche, honnête et si précieuse.
- A tous les membres du Département d'épidémiologie, biostatistiques et recherche clinique de Bichat, Isabelle, Florence, Xavière, Lucie, Carine, Sylvie, Agnès, Carole, Morgane, Emmanuelle, Karl, Gabriel, Gabriel, Xavier, et tous les étudiants, pour la bonne humeur et la complicité qui ont régné dans nos bureaux pendant ces années.
- A mes amis, pour votre soutien sans faille, spécialement Soisick, Francois, Odile, Vincent, Sophie, Alexis, Jean-Maxime, Najette, Jérôme, Véronique, Marc, Aline, Séverine et Delphine.
- A Nathanaël, Anne et Ségolène, à Laurent et Benjamin, à mes parents et beaux-parents et à toute ma famille, pour votre soutien tout au long de cette thèse.
- A Jérôme, pour ton aide, ta confiance et tes encouragements quotidiens.

Table des matières

Résumé

Cette thèse est consacrée au développement de nouvelles méthodes d'estimation adaptées à l'analyse de la dynamique virale sous traitement dans l'infection par le virus de l'immunodéficience humaine (VIH).

L'analyse statistique des données longitudinales de dynamique virale recueillies chez plusieurs patients repose sur l'utilisation de modèles non-linéaires à effets mixtes, qui permettent de distinguer une variabilité inter-patient d'une variabilité résiduelle. L'estimation des paramètres de ces modèles est délicate, l'expression de la vraisemblance n'étant pas explicite. Les travaux de cette thèse sont fondés sur l'algorithme d'estimation par maximum de vraisemblance SAEM, une version stochastique de l'algorithme *Expectation Maximisation* (EM) adaptée à ces modèles.

Nous avons développé des tests de Wald et du rapport de vraisemblance fondés sur l'algorithme SAEM permettant de comparer l'effet traitement de deux traitements sur les paramètres modélisant la décroissance de la charge virale. La matrice de Fisher, utilisée pour le test de Wald, est estimée par approximation stochastique et la vraisemblance par échantillonnage préférentiel. Dans ce cadre, nous avons également proposé une méthode de calcul du nombre de sujets nécessaires.

L'analyse de données de charge virale est compliquée par l'existence d'une censure de la mesure de ce marqueur. En effet, dès que le nombre de virus est trop faible, la concentration ne peut pas être mesurée précisément. Cette censure, si elle n'est pas prise en compte dans l'analyse statistique des données de charge virale, induit un biais dans l'estimation des paramètres du modèle. Nous avons proposé une extension de l'algorithme SAEM intégrant un algorithme de Gibbs hybride de simulation des données censurées, s'affranchissant de ce biais. Nous avons adapté les tests de Wald et du rapport de vraisemblance à ce cadre. L'utilisation de cette méthode de modélisation pour l'analyse de la décroissance de charge virale de l'essai clinique TRIANON-ANRS 81 de l'Agence Nationale de Recherche sur le Sida (ANRS) a mis en évidence une meilleure réponse des patients à l'un des deux traitements comparés, ce qui n'a pas pu être montré par une approche classique de traitement des données censurées dans le cadre d'un modèle mixte.

Les traitements anti-rétroviraux étant de plus en plus efficaces, la part de données de charge virale censurées augmente. Une évaluation à long terme des traitements ne peut donc pas reposer uniquement sur l'analyse de ce marqueur, et doit également s'appuyer sur l'évolution de la concentration de cellules lymphocytes $CD4^+$, qui est corrélée à celle de la charge virale. La dynamique conjointe de ces deux marqueurs est décrite par des systèmes différentiels complexes. Nous avons développé des versions de l'algorithme SAEM adaptées à l'estimation des paramètres de modèles mixtes définis par systèmes différentiels ordinaires ou stochastiques, en intégrant respectivement une méthode de résolution numérique du système différentiel ou du

processus de diffusion. Dans ce cadre, l'estimation des paramètres de ces modèles est réalisée sur un modèle statistique approché, dont la fonction de régression est une approximation numérique du système différentiel considéré. Nous avons montré la convergence des deux algorithmes d'estimation des paramètres de ces modèles statistiques approchés, tout en contrôlant l'erreur induite par l'approximation de la solution du système différentiel sur l'estimation des paramètres. Nous avons ensuite appliqué l'algorithme développé à l'analyse simultanée de la décroissance de la charge virale VIH et la croissance des lymphocytes $CD4^+$ observées chez des patients sous lopinavir dans l'essai COPHAR II-ANRS 111. Cette analyse nous a permis d'estimer les paramètres biologiques impliqués dans la dynamique virale, ainsi que leur variabilité inter-patient.

Enfin plusieurs médicaments étant administrés conjointement dans les traitements anti-rétroviraux, nous nous sommes intéressés aux méthodologies permettant d'étudier les interactions pharmacocinétiques entre ces molécules : dans les essais cliniques d'interaction pharmacocinétique, chaque patient est suivi lors de plusieurs périodes consécutives. Nous avons étendu l'algorithme SAEM à la modélisation de ce niveau de variabilité supplémentaire liée à la période d'observation. Nous avons utilisé cette méthode pour étudier l'interaction d'un inhibiteur nucléotidique de la transcriptase inverse, sur la pharmacocinétique d'un inhibiteur de protéase, deux médicaments prescrits dans les traitements anti-VIH, en analysant des données de concentration recueillies lors de l'essai Puzzle 2- ANRS 107.

Cette thèse propose donc des méthodes d'estimation adaptées à l'analyse de la dynamique virale dans l'infection par le VIH. Ces méthodes sont également applicables aux études pharmacodynamiques mises en place dans d'autres affections chroniques (hépatites, cancer, etc). De plus, l'analyse d'études pharmacocinétiques conduisant à ces mêmes problématiques, l'ensemble de nos résultats peut être utilisé dans ce domaine.

Productions scientifiques liées à la thèse

Articles acceptés

Adeline Samson, Marc Lavielle and France Mentré. Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model : application to HIV dynamics model. *Computational Statistics and Data Analysis*, 2006. (sous réserve de modifications mineures)

Sophie Donnet and **Adeline Samson**. Estimation of parameters in incomplete data models defined by dynamical systems. *Journal of Statistical Planning and Inference*, 2006. (sous réserve de modifications mineures)

Article en révision

Adeline Samson, Marc Lavielle and France Mentré. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. en révision pour *Statistics in Medicine*.

Article soumis

Sophie Donnet and **Adeline Samson**. Parametric inference for diffusion processes from discrete-time and noisy observations. Soumis à *Scandinavian Journal of Statistic*.

Article en préparation

Adeline Samson, Xavière Panhard, Xavier Duval, Marc Lavielle and France Mentré. Estimation of parameters of a simultaneous long-term model of HIV and CD4+ dynamics in patients initiating a lopinavir containing HAART. En préparation pour *Antiviral Therapy*.

Communications Orales

Adeline Samson, Marc Lavielle, France Mentré. Stochastic Approximation EM algorithm in nonlinear mixed effects models : an evaluation by simulation. *13th Meeting of the Population Approach Group in Europe*, Upssala, Sweden, 17-18 juin 2004.

Adeline Samson, Sophie Donnet. Estimation paramétrique dans des modèles définis par un système d'équations différentielles ordinaires. *2ème congrès de la Société Mathématiques Appliquées et Industrielles*, Evian, France, 23-27 mai 2005.

Adeline Samson, Sophie Donnet. Estimation paramétrique dans des modèles définis par un système d'équations différentielles ordinaires. *37èmes journées de la Société Francaise de Statistiques*, Pau, France, 6-10 juin 2005.

Adeline Samson, France Mentré, Marc Lavielle. Using SAEM, a new maximum likelihood estimation method in nonlinear mixed-effects models, for comparison of longitudinal responses. *5th International Meeting on Statistical Methods in Biopharmacy "Statistical innovations in clinical trials"*, Paris, France, 26-27 septembre 2005.

Adeline Samson, Marc Lavielle, France Mentré. The SAEM algorithm for nonlinear mixed models with left-censored data and differential systems : application to the joint modeling of HIV viral load and CD4 dynamics under treatment. *Sheiner Student Session, 15th Meeting of the Population Approach Group in Europe*, Brugge, Belgium, 14-16 juin 2006.

Posters

Adeline Samson, Xavière Panhard, Marc Lavielle, France Mentré. Generalisation of the SAEM algorithm to nonlinear mixed effects model defined by differential equations : application to HIV viral dynamic models. *14th Meeting of the Population Approach Group in Europe*, Pamplona, Spain, 16-17 juin 2005.

Sylvie Retout, Emmanuelle Comets, **Adeline Samson**, France Mentré. Designs in nonlinear mixed effects models : application to HIV viral load decrease with evaluation, optimisation and determination of the power of the test of a treatment effect. *14th Meeting of the Population Approach Group in Europe*, Pamplona, Spain, 16-17 juin 2005.

Adeline Samson, Marc Lavielle, France Mentré. Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model : application to HIV dynamics model. *5th International Symposium on measurement and kinetics of in vivo drug*

effects, Leyden, The Netherlands, 26-28 avril 2006.

Xavière Panhard, **Adeline Samson**. Extension of the SAEM algorithm for the estimation of Inter-Occasion Variability : application to the population pharmacokinetics of nelfinavir and its metabolite M8. *15th Meeting of the Population Approach Group in Europe*, Brugge, Belgium, 14-16 juin 2006. (soumis)

Première partie

Introduction

Chapitre 1

Contexte

Du fait du vieillissement de la population, l'incidence des maladies chroniques (infection par le VIH[1], hépatites, cancer, affections rhumatismales) est croissante dans nos sociétés. La prise en charge thérapeutique de ces maladies chroniques est donc l'un des champs majeurs de la recherche biomédicale. L'évaluation des traitements de ces pathologies repose classiquement sur l'analyse statistique de *données longitudinales*, recueillies lors d'essais cliniques, et mesurant au cours du temps et pour chaque individu considéré un critère de jugement principal, sensible à l'efficacité du traitement. Ce critère peut par exemple être la charge virale (quantité de molécules virales présentes dans un échantillon de plasma sanguin) lors d'une infection virale, le volume tumoral en oncologie, la hauteur de l'interligne articulaire dans l'arthrose, la pression artérielle dans l'hypertension ou encore l'état de la fonction respiratoire dans l'asthme. Ces études, qui étudient l'effet d'un médicament sur l'organisme, sont aussi appelées études de *pharmacodynamie* (PD). A partir de ces données longitudinales, l'efficacité de nouveaux traitements est alors comparée à celle du ou des traitements de référence par des tests statistiques de comparaison de groupes de patients.

En parallèle, il est nécessaire d'évaluer précisément le devenir dans l'organisme des médicaments impliqués dans la prise en charge thérapeutique de ces maladies chroniques. Cette évaluation appelée *pharmacocinétique* (PK), étudie la relation entre la dose administrée et la concentration du médicament dans l'organisme. Pour cela, la concentration du médicament dans le sang est mesurée au cours du temps chez plusieurs sujets. L'analyse de ces données longitudinales permet d'estimer les paramètres pharmacologiques caractérisant ces médicaments. Un nouveau médicament peut être alors comparé au médicament de référence à travers des tests statistiques de comparaison de ces paramètres pharmacocinétiques.

Nous avons choisi dans cette thèse de nous intéresser à l'infection par le VIH. Après avoir rappelé dans la section 1.1 les spécificités de cette maladie, nous présen-

[1] Virus de l'immunodéficience humaine

tons deux façons d'évaluer les traitements anti-rétroviraux, qui reposent d'une part sur les modèles, présentés dans la section 1.2, décrivant la dynamique du VIH et d'autre part sur les modèles décrivant la pharmacocinétique des médicaments anti-rétroviraux utilisés dans les traitements anti-VIH, et qui font l'objet de la section 1.3.

1.1 Évolution de l'infection par le VIH

Le VIH affecte le système immunitaire en infectant les cellules centrales de ce système : les lymphocytes T $CD4^+$. Il s'en suit un déficit à la fois quantitatif et qualitatif des lymphocytes T $CD4^+$ L'évolution de l'infection par le VIH peut donc être mesurée principalement par deux marqueurs biologiques impliqués dans cette infection, la charge virale et la concentration des principales cellules hôtes de ce virus, les lymphocytes T $CD4^+$. Les premières semaines suivant l'infection se traduisent par une augmentation extrêmement rapide de la charge virale parallèlement à une chute de la concentration de $CD4^+$ (phase aiguë). Suit alors une période de latence, pendant laquelle la charge virale se stabilise, alors que la concentration de $CD4^+$ augmente légèrement. Après une période dont la durée peut varier entre les patients (de quelques mois à plusieurs années), le virus se multiplie, provoquant une baisse inexorable de la concentration de $CD4^+$, l'apparition de maladies opportunistes, puis le décès du patient. Cette évolution est illustrée sur la figure 1.1 (a).

L'apparition des traitements anti-rétroviraux a permis d'allonger l'espérance de vie des patients, en essayant de maintenir la charge virale à un niveau faible, et ainsi d'augmenter la concentration de $CD4^+$. Cette période pendant laquelle l'infection est contrôlée peut durer plusieurs dizaines d'années, et varie selon les patients. Cependant à partir des mutations extrêmement fréquentes du virus, finit par émerger une ou plusieurs nouvelles formes du virus, résistantes aux différents traitements. Ces résistances provoquent l'augmentation de la charge virale et la diminution de la concentration de $CD4^+$, et le décès du patient par perte de sa protection immunitaire. L'évolution de la dynamique virale sous traitement est illustrée sur la figure 1.1 (b).

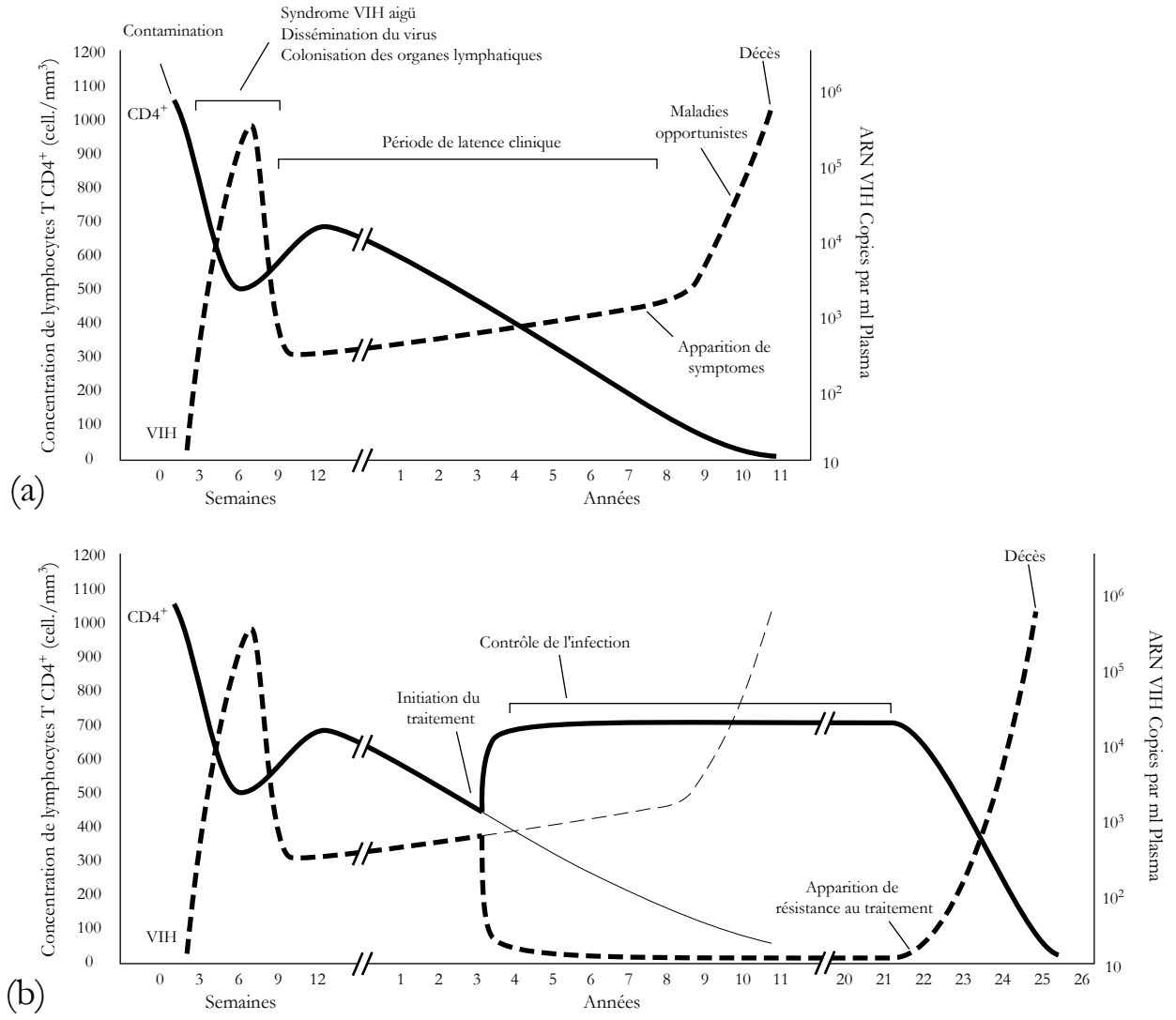


Figure 1.1: Évolution de la dynamique virale VIH en l'absence de traitement (a) ou avec initiation d'une multi-thérapie anti-rétrovirale (b)

L'efficacité des traitements anti-rétroviraux est évaluée à travers deux types d'études complémentaires. La première consiste à évaluer l'évolution de la dynamique virale à l'initiation du traitement, en particulier la rapidité de la décroissance de la charge virale, la capacité du traitement à maintenir en dessous d'un seuil cette charge virale pendant plusieurs semaines/années. Cette première évaluation est détaillée dans la section 1.2. Dans un deuxième temps, il est important d'étudier la pharmacocinétique de ces médicaments anti-rétroviraux, en particulier leur variabilité inter- et intra- patient. De plus, les traitements anti-rétroviraux étant composés de plusieurs médicaments, il est également important d'évaluer les interactions pharmacocinétiques entre ces médicaments co-administrés. Les modèles décrivant cette pharmacocinétique sont décrits dans la section 1.3.

1.2 Modèles de la dynamique du VIH

Comme nous l'avons rappelé précédemment, en l'absence de traitement, l'état stationnaire du système dynamique [virus-CD4$^+$] correspond à une charge virale importante et à une quantité de CD4$^+$ faible, responsable de l'état immunodéprimé rendant le patient sensible à de nombreuses infections opportunistes. A l'initiation du traitement, cette dynamique s'inverse. Les traitements anti-rétroviraux sont donc évalués sur leur capacité à modifier cette dynamique virale.

Les médicaments anti-rétroviraux agissent à différentes étapes du cycle réplicatif du VIH. On peut distinguer quatre classes thérapeutiques : les inhibiteurs de la transcriptase inverse (analogues nucléosidiques ou non), les inhibiteurs de protéase, les inhibiteurs de fusion et les inhibiteurs d'intégrase. Les effets respectifs de ces différentes classes de médicaments sur le processus d'infection d'un lymphocyte CD4$^+$ par le VIH et sur la réplication du virus sont illustrés par la figure 1.2.

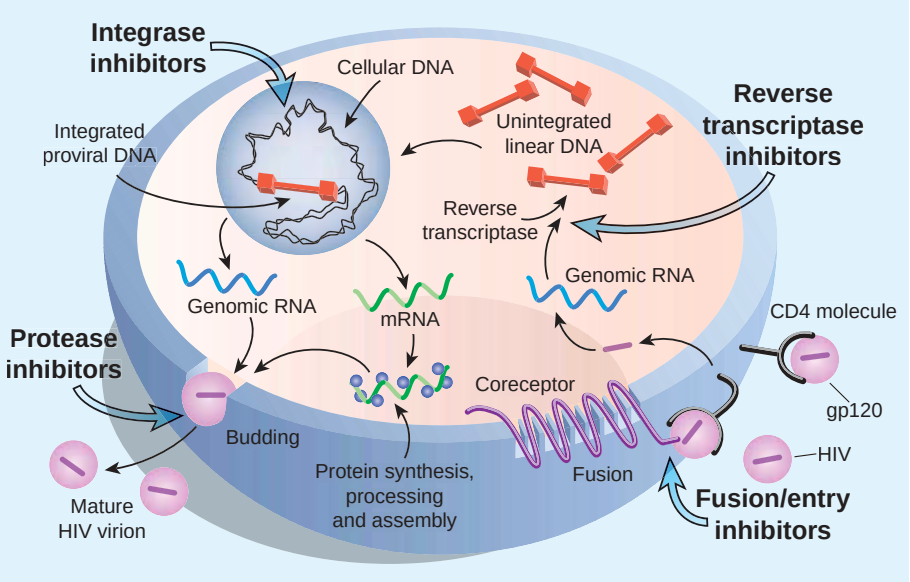


Figure 1.2: Cycle de réplication du VIH et action des anti-rétroviraux
Source : Nature Medicine ©Nature Publishing Group [1]

A l'initiation d'une multi-thérapie classique comprenant un inhibiteur de protéase et deux inhibiteurs de la transcriptase inverse, l'inhibiteur de protéase rend les virus présents non infectieux (les virus libres ne peuvent plus infecter de CD4$^+$) et l'inhibiteur de la transcriptase inverse diminue le nombre de copies du virus libérées suite à l'infection et à la destruction d'un CD4$^+$. Ces effets combinés entraînent donc la décroissance de la charge virale, et la restauration d'un niveau quasi normal de CD4$^+$.

La dynamique combinée du VIH et des CD4$^+$ peut être décrite par un système différentiel représentant les interactions entre ces deux vecteurs. Les différents mécanismes d'action des médicaments sur ces deux marqueurs sont ensuite intégrés au sein des systèmes différentiels décrivant la dynamique virale, reflétant ainsi leur

efficacité [2, 3, 4]. Les paramètres de ces systèmes ont tous une interprétation biologique (taux d'infection des $CD4^+$ par un virus, taux de réplication du virus au sein de la cellule lymphocytaire, durée de vie moyenne des $CD4^+$, du virus, vitesse de régénération des $CD4^+$, efficacité thérapeutique, etc), ce qui en fait leur principal attrait. En particulier, cette modélisation mathématique a joué et joue encore un rôle très important dans la compréhension de cette infection et dans l'amélioration de sa prise en charge thérapeutique. Ces systèmes sont détaillés dans le chapitre 7 de cette thèse.

1.3 Modèles pharmacocinétiques

Comme nous venons de le voir, différentes classes de médicaments sont combinées dans la prise en charge thérapeutique de l'infection par le VIH. L'association de deux médicaments peut être favorable (par exemple co-administration de ritonavir et d'indinavir afin de diminuer l'élimination de ce dernier) ou au contraire conduire à l'apparition de toxicité ou d'effets indésirables. Il est donc crucial d'étudier ces interactions en considérant les quatre grandes phases décrites en pharmacocinétique (l'absorption, la distribution, le métabolisme et l'élimination). Le devenir du médicament dans l'organisme peut s'envisager de manière dynamique : le corps humain est assimilé à un ensemble de compartiments entre lesquels le médicament peut s'échanger et éventuellement se transformer. La figure 1.3 représente un modèle à un seul compartiment schématisant le devenir d'une *Dose* de médicament administrée par voie orale, absorbé depuis l'intestin avec une constante d'absorption k_a vers un unique compartiment de volume V, et éliminé avec une constante d'élimination k_e.

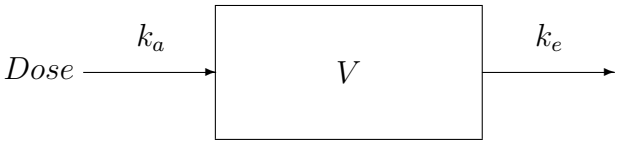


Figure 1.3: Modèle à un compartiment avec absorption et élimination du premier ordre

L'évolution de la quantité Q de médicament dans le sang est alors décrite par une équation différentielle qui dépend de la fonction d'administration du médicament $e(t)$ (orale, intra-veineuse, etc)

$$\frac{dQ(t)}{dt} = -k_e Q(t) + e(t).$$

Pour l'administration orale d'un médicament, cette fonction d'entrée s'écrit $e(t) = Dose \cdot k_a e^{-k_a t}$. Si ϕ est le vecteur de paramètres pharmacocinétiques $\phi = (V, k_a, k_e)$,

la concentration C du médicament dans le sang est alors décrite par l'équation différentielle suivante

$$\frac{dC(\phi,t)}{dt} = -k_e C(t,\phi) + \frac{Dose \cdot k_a}{V} e^{-k_a t}. \tag{1.1}$$

La solution C de cette équation

$$C(t,\phi) = \frac{Dose \cdot k_a}{V(k_a - k_e)} \left(e^{-k_e t} - e^{-k_a t}\right)$$

est une fonction non-linéaire en ϕ. La concentration d'un médicament dans le sang modélisée par cette fonction est représentée en figure 1.4.

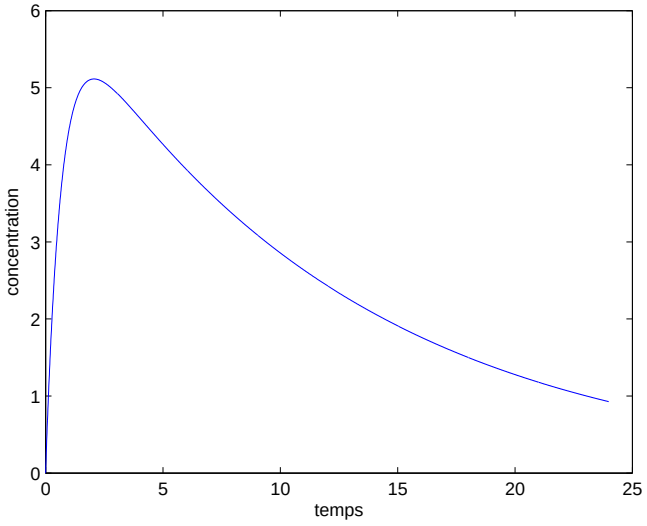


Figure 1.4: Courbe de concentration d'un médicament pour un modèle à un compartiment avec absorption et élimination du premier ordre

Lorsque l'étape d'élimination enzymatique du médicament est saturable, la pharmacocinétique du médicament est décrite par l'équation de Michaelis-Mentens :

$$\frac{dC}{dt}(t,\phi) = -\frac{V_m C(t,\phi)}{k_m + C(t,\phi)} + \frac{Dose \cdot k_a}{V} e^{-k_a t}, \tag{1.2}$$

où $\phi = (V, k_a, V_m, k_m)$ et qui, contrairement à la précédente équation, est sans solution analytique. Des modèles plus complexes à plusieurs compartiments sont également utilisés et permettent par exemple de modéliser la pharmacocinétique d'un médicament et de son métabolite.

Une interaction cliniquement importante entre deux médicaments anti-rétroviraux aura pour conséquence de modifier certains de ces paramètres pharmacocinétiques. Nous y reviendrons dans le chapitre 8 de cette thèse.

1.4 Analyse de données longitudinales

L'analyse de la dynamique du VIH (ou plus généralement de l'évolution d'une maladie) d'une part et l'analyse de l'évaluation des paramètres pharmacologiques d'un médicament d'autre part mettent en jeu des méthodologies statistiques similaires : l'analyse de données longitudinales à partir de modèles différentiels non-linéaires. On peut toutefois remarquer que les échelles de temps des phénomènes en jeu diffèrent de façon importante, les maladies chroniques nécessitant des suivis sur plusieurs semaines voire plusieurs années, alors que les médicaments étant généralement rapidement éliminés par le corps humain, les données d'études pharmacocinétiques sont recueillies sur une période de temps de l'ordre de quelques heures ou jours.

Cependant l'analyse de ces données longitudinales requiert des méthodes statistiques adaptées. En effet, les vecteurs de données de chaque patient sont supposés indépendants les uns des autres, mais les données d'un même sujet sont bien évidemment corrélées dans le temps. Les modèles à effets mixtes ont été développés pour tenir compte de cette corrélation. On peut alors distinguer différentes sources de variabilité : une variabilité entre les individus, dite *inter*-patient, une variabilité des paramètres d'un même sujet au cours du temps, dite *intra*-patient et enfin une variabilité *résiduelle* représentant l'écart par rapport au modèle utilisé. En général, les variabilités intra-patient et résiduelle sont confondues.

Ces modèles mixtes permettent de plus d'évaluer la distribution des paramètres du système biologique au sein de l'ensemble de la population en considérant dans le modèle statistique les paramètres individuels comme des variables aléatoires ("effet aléatoire") centrées autour de leur valeur moyenne ("effet fixe") de population. Il est alors possible d'en déduire une courbe moyenne dite de population, reflétant l'évolution moyenne du processus biologique observé au sein de la population. Pour chaque patient peut être prédite une courbe individuelle reflétant son mécanisme propre. La figure 1.5 illustre ce principe sur les données de pharmacocinétique de la théophylline.

Les modèles mixtes sont de plus en plus utilisés pour analyser l'évolution de maladies chroniques, notamment, la dynamique d'infections virales. Parmi ces infections, l'évolution de la charge virale VIH a été largement décrite par modèles mixtes [5, 6, 7, 8, 9]. Les modèles mixtes constituent également un outil de référence dans le cadre de l'étude des caractéristiques d'un médicament en pharmacocinétique [10, 11] et sont largement utilisés pendant la phase III[2] du développement des médicaments [12].

[2]Le développement d'un médicament est réalisé en plusieurs étapes. Après des études précliniques in vitro ou sur l'animal, le médicament est validé au cours de trois phases. Les essais de phase I étudient la pharmacocinétique et la tolérance du médicament chez le volontaire sain. Les études de phase II explorent l'efficacité thérapeutique chez des malades. L'étude de phase III consiste à prouver son efficacité comparativement au traitement de référence ou à un placebo, chez des malades.

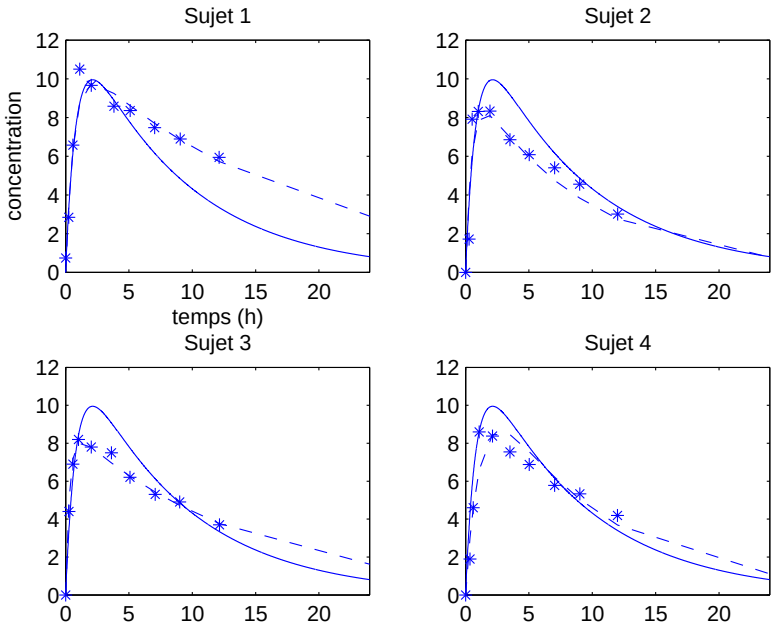


Figure 1.5: Courbes de population (ligne pleine) et individuelles (ligne en pointillée) de la pharmacocinétique de théophylline pour 4 sujets (valeurs observées représentées par des étoiles)

En particulier, ces modèles sont adaptés lorsque le nombre de prélèvements par sujet est faible, ce qui est souvent le cas dans ces essais conduits dans des populations particulières telles que les patients, les enfants ou les personnes âgées.

Cependant l'estimation des paramètres de ces modèles non-linéaires mixtes est complexe, et différentes méthodes d'estimation, présentées dans la partie 2, ont été proposées. Nous détaillerons plus spécifiquement l'algorithme SAEM, une version stochastique de l'algorithme *Expectation Maximisation* (EM), sur lequel les différents développements de cette thèse reposent.

Chapitre 2

Modèles non-linéaires à effets mixtes et méthodes d'estimation

2.1 Modèles et notations

Nous considérons un problème statistique où les données *observées* de l'individu i sont notées $y_i = (y_{i1}, \ldots, y_{in_i})^t$ où y_{ij} est la mesure de la variable y pour le sujet i à l'instant t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Le modèle non-linéaire mixte s'écrit

$$\begin{aligned} &y_{ij} = f(\phi_i, t_{ij}) + g(\phi_i, t_{ij})\, \varepsilon_{ij}, \\ &\varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2 I_{n_i}), \\ &\phi_i = X_i\mu + b_i, \text{où} \;\; b_i \sim_{iid} \mathcal{N}(0, \Omega), \end{aligned} \tag{2.1}$$

où ϕ_i est le vecteur des paramètres individuels, $f(\cdot)$ et/ou $g(\cdot)$ sont des fonctions non-linéaires de $\phi = (\phi_1, \ldots, \phi_N)$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^t$ représente l'erreur résiduelle du sujet i, μ la matrice des effets fixes, X_i le vecteur des covariables, b_i le vecteur des effets aléatoires, supposés indépendants de ε_i, σ^2 la variance résiduelle, I_{n_i} la matrice identité et Ω la matrice de variance inter-individuelle des effets aléatoires. On s'intéresse dans ce cas à l'estimation du vecteur $\theta = (\mu, \Omega, \sigma^2) \in \Theta$ où Θ est un sous-ensemble de $\mathbb{R}^p$.

Ce modèle peut être considéré comme un problème à données *non observées*, les paramètres individuels ϕ étant inconnus. Dans la littérature statistique, ces modèles sont également appelés *incomplete data models* ou *modèles à données manquantes*. Nous éviterons autant que possible dans cette thèse d'employer cette expression car dans la littérature statistique médicale, le terme *données manquantes* est normalement consacré et réservé aux données qui auraient dû être mesurées mais qui n'ont pas pu l'être pour des raisons externes (patients ne se présentant pas à un rendez-vous, sorties d'étude, covariable non renseignée, etc). De plus, il ne faut pas confondre ces données manquantes avec les données censurées traitées dans cette

thèse, et qui sont des données de mesure de concentration censurées par un seuil de détection de dispositif expérimental (concentration de virus trop faible pour être mesurée avec précision). Dans ce cas, la raison de la censure est connue et peut être intégrée au modèle statistique. Le traitement des données manquantes est un axe de recherche important, qui dépasse largement le cadre de cette thèse.

Dans la suite, le vecteur x désigne plus généralement le vecteur des données non observées appartenant à $\mathbb{R}^l$ (dans certains modèles considérés dans cette thèse, x ne sera pas composé uniquement du vecteur ϕ). Le vecteur (y, x) est appelé vecteur des données *complètes*. On suppose que la vraisemblance des données complètes (y, x) appartient à une famille paramétrique notée $\{p(\cdot;\theta), \theta \in \Theta\}$ par rapport à une mesure borélienne ν σ-finie (par exemple la mesure de Lebesgue sur l'espace de définition de x). La vraisemblance des observations est alors définie par

$$p(y;\theta) = \int_{\mathbb{R}^l} p(y, x;\theta)\nu(dx)$$

où $p(y, x;\theta)$ est la vraisemblance des données complètes. Les fonctions de régressions f et/ou g étant non-linéaires, la vraisemblance des données observées n'a pas d'expression analytique. L'estimation des paramètres de ces modèles n'est alors pas directe. En revanche, en général, la vraisemblance des données complètes a une expression analytique.

C'est le cas par exemple pour un modèle non-linéaire mixte simple où les seules données non observées sont les vecteurs des paramètres individuels ($x = \phi$). Pour ce modèle, la vraisemblance des données complètes s'écrit

$$\begin{aligned}\log p(y, \phi;\theta) &= -\frac{\sum_{i=1}^{N} n_i}{2}\log(2\pi\sigma^2 g^2(\phi_i, t_{ij})) - \frac{1}{2\sigma^2 g^2(\phi_i, t_{ij})}\sum_{i,j}(y_{ij} - f(\phi_i, t_{ij}))^2 \\ &\quad -\frac{N}{2}\log(2\pi|\Omega|) - \frac{1}{2}\sum_{i}(\phi_i - X_i\mu)\,\Omega^{-1}(\phi_i - X_i\mu)^t.\end{aligned}$$

Certaines méthodes d'estimation des paramètres des modèles non-linéaires à effets mixtes utilisent cette propriété et sont fondées sur le calcul de la vraisemblance des données complètes au lieu du calcul direct de la vraisemblance des données observées.

Nous présentons dans la section suivante les principales méthodes d'estimation habituellement utilisées, c'est-à-dire celles qui sont mises en oeuvre dans des logiciels à la disposition des utilisateurs. Nous distinguons les méthodes classiques des méthodes par approche bayesienne. Toutefois, les méthodes classiques implémentées dans ces logiciels étant toutes basées sur une approximation de la vraisemblance, aucune ne maximise la vraisemblance des données observées du modèle mixte original. Nous présentons dans la section 2.3 des méthodes "exactes" d'estimation par maxi-

mum de vraisemblance, qui reposent sur des méthodes statistiques plus élaborées et complexes.

2.2 Méthodes d'estimation usuelles pour les modèles non-linéaires mixtes

Nous présentons ici les méthodes d'estimation des paramètres des modèles non-linéaires mixtes qui sont implémentées dans les principaux logiciels à la disposition de l'utilisateur.

2.2.1 Méthodes d'estimation basées sur une approximation du modèle

Plusieurs algorithmes basées sur des approximations du modèle ont été proposées, présentant des estimateurs minimisant des erreurs quadratiques d'un modèle approché. Les plus utilisés sont les algorithmes itératifs First Order (FO) et First Order Conditional Estimate (FOCE), développés notamment par Beal et Sheiner [13] et Lindstrom et Bates [14]. Une linéarisation à l'ordre un de la fonction de régression par rapport aux effets aléatoires permet d'avoir une expression analytique de la vraisemblance, qui est maximisée par un algorithme de Newton-Raphson, les effets aléatoires ϕ étant estimés par maximum a posteriori de la distribution conditionnelle $p(\cdot|y;\theta)$ à chaque itération. Ces deux algorithmes sont implémentés dans le logiciel NONMEM, utilisé par une large majorité des laboratoires pharmaceutiques, et dans la fonction `nlme` de Splus et R. Il existe aussi des méthodes basées sur une approximation de Laplace ou une quadrature de Gauss de la vraisembalnce [15], qui sont par exemple mises en oeuvre dans la procédure NLMIXED du logiciel SAS.

Cependant, aucune de ces méthodes n'est considérée comme établie théoriquement, aucune preuve de convergence vers un maximum de vraisemblance n'ayant été publiée. En particulier, Vonesh [16] donne un exemple pour lequel les estimateurs produits par les algorithmes de linéarisation FO, FOCE sont inconsistants dès que le nombre d'observations par sujet croît moins vite que le nombre de sujets. Ge et al. [17] rapportent des problèmes similaires lorsque la variance des effets aléatoires est trop grande. De plus, il est très fréquent que ces logiciels ne convergent pas, et les tests statistiques utilisés pour la sélection des covariables ont des propriétés mal connues. Par exemple, plusieurs auteurs ont montré par simulation une inflation du risque de première espèce des tests les plus utilisés (Wald, rapport de vraisemblance) avec ces algorithmes [6, 18, 19, 20].

Récemment, une méthode a été proposée comme alternative à la linéarisation du modèle, intégrant numériquement la vraisemblance par échantillonnage préférentiel.

Cependant, comme l'ont souligné Ge et al. [17], la stabilité numérique de cet algorithme s'obtient par des méthodes lourdes, nécessitant un temps de calcul très long par rapport à d'autres méthodes paramétriques.

Il existe donc un réel besoin de méthodes d'estimation par maximum de vraisemblance des paramètres d'un modèle non-linéaire mixte, maximisant la vraisemblance du modèle original et non pas celle d'un modèle approché, et qui soient convergentes et consistantes. Des méthodes répondant à ces critères sont présentées dans la partie 2.3.

2.2.2 Méthodes d'estimation bayesiennes

La deuxième approche d'estimation habituellement utilisée est l'approche bayesienne. Dans ce cadre, une loi a priori, notée $\pi(\theta)$, sur le paramètre θ est introduite dans le modèle mixte. Le problème est alors d'évaluer la loi a posteriori $p(\theta|y) \propto \pi(\theta)p(y;\theta)$, où $p(y;\theta)$ est la vraisemblance du modèle. La fonction de vraisemblance n'ayant pas d'expression analytique dans les modèles non-linéaires mixtes, et les constantes de normalisation de ces différentes distributions ne pouvant pas être calculées, l'évaluation de la distribution a posteriori est difficile. Les méthodes de Monte Carlo par chaînes de Markov (MCMC) ont été développées pour contourner ce problème. A l'origine développées dans un contexte bayésien, ces méthodes sont maintenant utilisées dans des méthodes d'estimation par maximum de vraisemblance. Le principe des méthodes MCMC (algorithmes de Gibbs, de Metropolis-Hastings) est détaillé plus loin.

Gelfand et al. [21] ont proposé d'utiliser un échantillonneur de Gibbs pour évaluer les distributions a posteriori $p(\theta|y)$ et conditionnelle $p(\phi|y)$ simultanément, ces deux distributions $p(\theta|y) = \int p(\theta|y,\phi)\ p(\phi|y;\theta)\ d\phi$ et $p(\phi|y) = \int p(\phi|y;\theta)p(\theta|y)\ d\theta$ étant fortement dépendantes l'une de l'autre. Cependant, pour les modèles non-linéaires mixtes, la simulation selon la distribution $p(\phi|y;\theta)$ n'est pas directe. Des algorithmes de Monte Carlo par chaînes de Markov hybrides ont été proposés par Racine-Poon [22], Wakefield et al. [23, 24] et Bennet et al. [25]. Spiegelhalter et al. [26] ont développé les logiciels BUGS et PK-BUGS mettant en oeuvre ces algorithmes.

Au cours de cette thèse, nous avons développé deux versions de ces algorithmes bayésiens adaptés aux modèles mixtes définis par des systèmes dynamiques. Ces travaux sont détaillés dans les sections 6.1 et 6.2.

2.3 Méthodes d'estimation par maximum de vraisemblance

Nous présentons maintenant les différentes méthodes d'estimation réalisant une maximisation exacte de la vraisemblance du modèle.

Étant donnée l'expression complexe de la vraisemblance d'un modèle non-linéaire mixte, que nous avons rappelée dans la section 2.1, l'estimation des paramètres de ces modèles par maximisation de cette vraisemblance est difficile. Deux alternatives principales sont possibles, fondées soit sur l'utilisation de l'algorithme de Newton-Raphson (NR) ou sur celle de l'algorithme Expectation-Maximization (EM). Les différentes versions proposées pour les modèles mixtes et fondées sur une approximation Monte-Carlo, une approximation stochastique ou autre, de ces algorithmes sont résumées dans le tableau 2.1.

Table 2.1: Méthodes d'estimation par maximum de vraisemblance des modèles non-linéaires à effets mixtes

Algorithme d'estimation	Méthodes d'approximation		
	Monte Carlo	Approximation stochastique	Autre
Newton-Raphson	Mc Culloch [27]	Gu et Kong [28]	Commenges et al. [29]
Expectation-Maximization	Wu [7]	Kuhn et Lavielle [30]	

La plupart de ces méthodes intègrent des procédures de Monte Carlo par chaînes de Markov de simulation des données non-observées. Nous rappelons leur principe dans le paragraphe suivant.

2.3.1 Méthodes de Monte Carlo par chaînes de Markov

On appelle algorithme de Monte Carlo par chaînes de Markov (MCMC) toute méthode produisant une chaîne de Markov ergodique ayant pour loi stationnaire une loi $\pi(\cdot)$. Ces méthodes sont généralement utilisées lorsque la loi $\pi(\cdot)$ n'est pas simulable directement et/ou lorsqu'elle est connue à une constante de normalisation près. Plusieurs algorithmes MCMC ont été proposés, qui sont des versions hybrides ou des généralisations de l'échantillonneur de Gibbs et de l'algorithme de Metropolis-Hastings. Robert et Casella proposent une revue de ces méthodes [31]. Nous rappelons ici leurs principes, plusieurs algorithmes hybrides ayant été développés dans cette thèse.

L'échantillonneur de Gibbs et l'algorithme de Metropolis-Hastings sont basés sur le même principe : ils assurent l'existence de noyaux de transition, noté $\Pi(x, x')dx'$, permettant de générer une chaîne de Markov de loi stationnaire $\pi(\cdot)$. On note $X = \{X_t; t \geq 0\}$, avec $X_t = (X_{t,1}, \ldots, X_{t,d})$, cette chaîne de Markov de dimension d et de loi stationnaire $\pi(\cdot)$.

Les principales propriétés de convergence des chaînes de Markov sont l'*ergodicité*

géométrique et l'*ergodicité uniforme*, que nous rappelons ici.

Une chaîne est *géométriquement ergodique* si elle est ergodique et s'il existe une fonction positive M de mesure finie $\pi(M) < \infty$ et une constante $r < 1$ telles que

$$\left\|\Pi^k(x,\cdot) - \pi\right\| \leq M(x) r^k$$

pour tout x et tout $k \geq 0$.

Une chaîne est *uniformément ergodique* si elle est ergodique et s'il existe une constante positive M et une constante $r < 1$ telles que

$$\left\|\Pi^k(x,\cdot) - \pi\right\| \leq M r^k$$

pour tout $k \geq 0$.

Différentes conditions assurant l'une ou l'autre des ergodicités ont été proposées pour les deux algorithmes.

Échantillonneur de Gibbs

L'algorithme de Gibbs repose sur la probabilité de transition suivante

$$\Pi(X, X') = \prod_{i=1}^{d} q_i(X'_i | X'_1, \ldots, X'_{i-1}, X_{i+1}, \ldots, X_d)$$

où $q_i(x_i|x_j, j \neq i) = \pi(x_i|x_j, j \neq i)$ est la i-ème densité conditionnelle. A partir de la valeur X_{k-1} de la chaîne obtenue à l'itération $k-1$, la valeur X_k est obtenue après la réalisation de d étapes de simulation à l'aide des lois q_i appelées aussi lois instrumentales (*proposal distributions*) :

Etape 1 : Simulation de $X_{k,1}$ suivant $q_1(\cdot|X_{k-1,2}, \ldots, X_{k-1,d})$

Etape i, $i = 2, \ldots, d-1$: Simulation de $X_{k,i}$ suivant $q_i(\cdot|X_{k,1}, \ldots, X_{k,i-1}, X_{k-1,i+1}, \ldots X_{k-1,d})$

Etape d : Simulation de $X_{k,d}$ suivant $q_d(\cdot|X_{k,1}, \ldots, X_{k,d-1})$

La convergence de cet algorithme et l'ergodicité géométrique ou uniforme de la chaîne simulée sont assurées, par exemple, sous une condition de minoration proposée par Tierney [32]. Robert et Casella [33] détaillent d'autres conditions de convergence. La principale limite de l'échantillonneur de Gibbs réside dans le choix restreint de lois instrumentales possibles. En effet, il est nécessaire de connaître π analytiquement ou de savoir simuler chacune des densités conditionnelles pour implémenter cet algorithme. Lorsque ce n'est pas le cas, il est possible d'avoir recours à l'algorithme de Metropolis-Hastings.

Algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings a pour probabilité de transition

$$\Pi(X, X') = q(X, X')\rho(X, X')$$

où $q(X, \cdot)$ est la loi instrumentale et $\rho(X, X')$ est la probabilité d'accepter le candidat X' à partir de X. A partir de la valeur X_{k-1} de la chaîne obtenue à l'itération $k-1$, un candidat X' est généré dans la distribution $q(X_{k-1}, \cdot)$. On choisit ensuite $X_k = X'$ avec la probabilité

$$\rho(X_{k-1}, X') = \min\left(\frac{p(X'|y)}{p(X_{k-1}|y)}\frac{q(X'|X_{k-1})}{q(X_{k-1}|X')}, 1\right)$$

et $X_k = X_{k-1}$ avec la probabilité $1 - \rho(X_{k-1}, X')$.

Cette procédure accepte le candidat proposé chaque fois que le quotient $\frac{p(X'|y)}{q(X|X')}$ est supérieur au précédent. La probabilité d'acceptation ne dépend que du rapport des distributions conditionnelles $\frac{p(X'|y)}{p(X_{k-1}|y)}$, ce qui permet de contourner le problème du calcul, souvent impossible, de la constante de normalisation.

L'universalité de cet algorithme en fait un outil très puissant. Quelque soit la loi instrumentale q, on peut simuler une variable sous n'importe quelle distribution π dès que les supports coïncident. Cependant, cette universalité reste théorique si la loi instrumentale q ne simule que rarement des "bons" candidats, c'est-à-dire des candidats dans la région où π a la plus grande masse. Le choix de q est donc important et dépend fortement de π. Deux lois instrumentales particulières sont en général utilisées.

Dans le premier cas, on considère une loi instrumentale possédant la propriété suivante $q(X, X') = q(X')$. On parle alors d'algorithme de Metropolis-Hastings *indépendant*, la simulation du candidat étant indépendante de la valeur précédente de la chaîne. Dans ce cas, lorsque la loi instrumentale q est proche de la loi π, la convergence est rapide. Cependant dans certains cas, cet algorithme est très sensible aux valeurs initiales et aux états absorbants de la chaîne. Il est donc recommandé de ne pas l'utiliser seul.

La deuxième classe de lois instrumentales utilisées repose sur la propriété de symétrie $q(X, X') = q(X', X) = q(X' - X)$, on parle alors d'une *marche aléatoire*. L'idée sous-jacente consiste à prendre en compte la dernière valeur simulée pour simuler la suivante. La chaîne de Markov est alors géométriquement ergodique sous des conditions générales raisonnables, par exemple lorsque la loi π appartient à la famille exponentielle et est suffisamment régulière. En général, cette marche aléatoire dépend d'un paramètre d'échelle δ dont le choix est délicat, un exemple courant consistant à simuler X' selon la loi $\mathcal{N}(X, \delta)$. Si δ est trop grand, une large proportion de candidats vont être rejetés. Inversement si δ est trop petit, la marche aléatoire

n'accepte que des candidats dans un petit voisinage et se déplace très lentement dans l'espace des paramètres. Lorsque la dimension d de la chaîne de Markov est petite, différents auteurs [34, 35] recommandent d'adapter ce paramètre d'échelle afin d'assurer un taux de candidats *acceptable* proche de 30%. Cependant, dès que d augmente, aucune procédure n'est proposée.

On peut remarquer que l'algorithme de Gibbs est formellement un cas particulier de celui de Metropolis-Hastings : c'est une combinaison de d algorithmes de Metropolis-Hastings appliqués à chaque composante du vecteur X avec un taux d'acceptation de 1.

Algorithme MCMC hybride

Le point faible de l'algorithme de Metropolis-Hastings est le choix de la loi instrumentale. Cette loi peut être très différente de la loi π visée, et l'algorithme risque alors de simuler "grossièrement" π. Ce n'est pas le cas pour l'échantillonneur de Gibbs, dont la loi instrumentale est directement déduite de la loi π. Cependant, la structure composée de l'échantillonneur de Gibbs peut être une faiblesse. Par exemple, pour une loi de mélange π, l'échantillonneur de Gibbs peut n'explorer qu'une seule des deux composantes du support de π. La liberté du choix de q dans l'algorithme de Metropolis-Hastings permet parfois de remédier à ce problème, en particulier au travers de l'optimisation des paramètres d'échelle.

Pour conserver les avantages de ces deux méthodes, Tierney [32] propose une approche hybride. On appelle alors algorithme *MCMC hybride* ou *Gibbs hybride* une méthode MCMC combinant simultanément des étapes d'un échantillonneur de Gibbs avec des étapes d'algorithmes de Metropolis-Hastings. Si $\Pi_1, \ldots, \Pi_n$ sont les noyaux de transition correspondants à ces étapes, on appelle *cycle* de $\Pi_1, \ldots, \Pi_n$ l'algorithme de noyau $\Pi^* = \Pi_1 \circ \ldots \circ \Pi_n$, où $\circ$ est l'opérateur de composition de fonctions. Une composition de noyaux associés à la même loi stationnaire π correspond à un noyau Π^* de loi stationnaire π. L'ergodicité uniforme de la chaîne de Markov est assurée dès que l'un des noyaux la garantit.

Différentes versions de méthodes MCMC hybrides sont développées et proposées au cours de cette thèse, chacune s'adaptant à une problématique rencontrée dans la modélisation de données longitudinales biomédicales (parties 5, 6.2 et 8).

Nous présentons maintenant les méthodes d'estimation par maximum de vraisemblance, utilisant ces méthodes MCMC. Dans la section 2.3.2, les méthodes fondées sur l'algorithme de Newton-Raphson sont exposées, celles reposant sur l'algorithme Expectation-Maximisation sont détaillées dans la section 2.3.3.

2.3.2 Méthodes d'estimation basées sur l'algorithme de Newton-Raphson

L'algorithme de Newton-Raphson est une méthode classique d'estimation par maximum de vraisemblance. C'est un algorithme itératif reposant sur la résolution d'une équation de score. Notons respectivement $J(\theta) = \frac{\partial \log p(y;\theta)}{\partial \theta}$ et $H(\theta) = \frac{\partial^2 \log p(y;\theta)}{\partial \theta \partial \theta'}$ les fonctions de score (jacobien) et hessienne de la fonction de vraisemblance. L'estimateur du maximum de vraisemblance est obtenu itérativement comme la solution de l'équation $J(\theta) = 0$. A l'itération k, l'estimateur de Newton-Raphson est actualisé par la relation

$$\theta_k = \theta_{k-1} + (H(\theta_{k-1}))^{-1} J(\theta_{k-1}).$$

Cependant, cet algorithme nécessite le calcul des fonctions jacobienne $J(\theta)$ et hessienne $H(\theta)$ de la fonction de vraisemblance. Or, lorsque la vraisemblance $p(y;\theta)$ et ces intégrales sont sans expression analytique, l'algorithme de Newton-Raphson ne peut pas être appliqué directement. Des versions stochastiques adaptées aux problèmes à données non observées sont donc proposées, se ramenant au calcul de la vraisemblance analytique des données complètes $p(y, x; \theta)$ grâce au théorème proposé par Louis [36]. Le principe de Louis [36] relie les fonctions jacobienne et hessienne de la vraisemblance observée $p(y;\theta)$ et celles de la vraisemblance des données complètes $p(y, x; \theta)$

$$\begin{aligned} J(\theta) &= E\left[\frac{\partial \log p(y,x;\theta)}{\partial \theta} | y, \theta\right] \\ H(\theta) &= E\left[\frac{\partial^2 \log p(y,x;\theta)}{\partial \theta \partial \theta'} | y, \theta\right] + \mathrm{Var}\left[\frac{\partial \partial_\theta \log p(y,x;\theta)}{\partial \theta} | y, \theta\right]. \end{aligned}$$

L'algorithme Monte Carlo Newton-Raphson (MC-NR)

Des approximations de Monte Carlo de l'algorithme de Newton-Raphson ont été proposées par Mc Culloch [27] et Tanner [37]. Dans ce cas, les calculs des intégrales de score et hessienne sont remplacés à chaque itération par des approximations empiriques de Monte-Carlo basées sur un échantillon simulé de données non observées. A l'itération k, le gradient est approché par

$$\hat{J}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log p(y, x_k^t; \theta_k)}{\partial \theta}$$

où $(x_k^t)_{1 \le t \le T}$ est un échantillon de taille T distribué selon la loi $p(x|y;\theta_k)$. La matrice hessienne $\hat{H}(\theta)$ est estimée de façon similaire. D'après la loi des grands nombres, cette approximation peut être aussi précise que voulue en augmentant simplement la taille T de l'échantillon.

Cette solution élégante soulève de nouvelles questions, en particulier le choix de

T qui doit augmenter à chaque itération pour assurer la convergence de l'algorithme. Ceci peut engendrer des problèmes numériques, sur lesquels nous reviendrons plus en détail lors de la présentation de la version de Monte Carlo de l'algorithme EM (MC-EM). Pour les modèles non-linéaires mixtes, la simulation d'un échantillon selon la loi $p(x|y;\theta_k)$ est complexe puisque cette loi n'est pas connue. L'algorithme MC-NR est alors combiné avec des méthodes de simulation du type MCMC, que nous avons présenté dans la section précédente.

L'algorithme Stochastic Approximation Newton-Raphson (SA-NR)

La principale alternative à l'approche de Monte Carlo, assurant une convergence avec un nombre fixé et petit de simulations, repose sur la méthode d'approximation stochastique (SA) développée par Robbins et Monroe [38]. Le principe de cette méthode est le suivant : plutôt que d'augmenter

la taille de l'échantillon simulé à chaque itération, la méthode d'approximation

stochastique calcule

une moyenne pondérée entre la valeur calculée à l'itération k et les

valeurs calculées aux
itérations précédentes.

En utilisant une suite de poids décroissants (γ_k), cette

ERRATUM

méthode réalise l'approximation stochastique suivante

$$\begin{aligned} J_{k+1}(\theta) &= J_k(\theta) + \gamma_k \left(\frac{\partial \log p(y, x_k; \theta_k)}{\partial \theta} - J_k(\theta)\right), \\ &= (1 - \gamma_k) J_k(\theta) + \gamma_k \frac{\partial \log p(y, x_k; \theta_k)}{\partial \theta} \end{aligned}$$

à partir d'une initialisation déterministe $J(\theta_0)$ et où x_k est une réalisation simulée selon la distribution $p(x|y, \theta_k)$. La matrice hessienne $H_{k+1}(\theta)$ est estimée de façon similaire. Grâce à cette procédure, l'information tirée des premières itérations a de moins en moins d'importance, alors que les dernières itérations, qui apportent une information plus précise de la valeur à évaluer, ont un poids relatif de plus en plus important.

Une version SA de l'algorithme de Newton-Raphson est proposée par Gu et Li [39]. Gu et Kong [28] ont proposé de combiner cet algorithme SA-NR avec une méthode MCMC pour la simulation de ces paramètres individuels en particulier lorsque la simulation dans la loi a posteriori des paramètres individuels $p(x|y; \theta)$ n'est pas directe. Sous des conditions de régularité du modèle et d'ergodicité de la chaîne simulée par méthode MCMC, ils ont montré la convergence de leur estimateur.

Algorithme quasi-Newton

Cependant, une approximation numérique satisfaisante de la fonction hessienne H est souvent difficile à obtenir, en particulier pour les modèles non-linéaires mixtes pour lesquels la loi $p(x|y; \theta)$ est inconnue. Commenges et al. [29] ont récemment proposé un algorithme du type Newton-Raphson reposant sur l'évaluation suivante des gradient et hessien

$$J(\theta) = p(y; \theta)^{-1} E\left[p(y, x; \theta) \frac{\partial \log p(y, x; \theta)}{\partial \theta} | y; \theta\right],$$

$$H(\theta) = -J(\theta)J(\theta)^t + p(y; \theta)^{-1} E\left[p(y, x; \theta)\left(\frac{\partial^2 \log p(y, x; \theta)}{\partial \theta \partial \theta'} + \frac{\partial \log p(y, x; \theta)}{\partial \theta} \frac{\partial \log p(y, x; \theta)}{\partial \theta}^t\right) | y; \theta\right].$$

Plus rapide que l'algorithme original de Newton-Raphson, il est aussi performant, les propriétés de convergence étant conservées. Guedj et al [40] ont appliqué cette méthode à un modèle d'équations différentielles modélisant la dynamique virale VIH, et montrent que dans ce cadre, le calcul de la fonction de score est exact.

2.3.3 Méthodes d'estimation basées sur l'algorithme EM

La principale alternative à l'algorithme de Newton-Raphson est l'algorithme Expectation-Maximisation (EM) proposé par Dempster et al [41]. Comme pour l'al-

gorithme de Newton-Raphson, la vraisemblance $p(y;\theta)$ n'étant pas connue analytiquement, il repose sur le calcul de l'espérance conditionnelle de la log-vraisemblance complète par rapport aux données observées y

$$Q(\theta|\theta') = \int \log p(y,x;\theta) p(x|y;\theta') \nu(dx),$$

utilisant le fait que la vraisemblance des données complètes $p(y,x;\theta)$ est connue analytiquement. Le succès de cet algorithme repose largement sur sa propriété de monotonie

Proposition 1 *Pour tout* (θ,θ') *dans* Θ^2, *si* $Q(\theta|\theta') \geq Q(\theta|\theta)$, *alors* $\log p(y;\theta') \geq \log p(y;\theta)$ *où* $p(y;\cdot)$ *est la vraisemblance des données observées.*

Cette propriété implique que tout accroissement de Q engendre un accroissement de $p(y;\cdot)$. Ainsi, lorsque la maximisation de Q est plus simple que la maximisation de $p(y;\cdot)$, des maximisations successives de Q peuvent permettre d'atteindre un maximum de $p(y;\cdot)$. S'appuyant sur ce principe, Dempster et al. [41] proposent donc l'algorithme itératif EM. A l'itération k, cet algorithme est réalisé en deux étapes :

– Étape E : calcul de l'espérance conditionnelle de la log-vraisemblance complète

$$Q(\theta|\theta_k) = \int \log p(y,x;\theta) p(x|y;\theta_k) \nu(dx).$$

– Étape M : maximisation de cette quantité en θ

$$\theta_{k+1} = \arg\max_{\theta} Q(\theta|\theta_k)$$

La suite $(p(y;\theta_k))_{k\geq 1}$ est une suite de valeurs croissantes de la vraisemblance des données observées. La convergence de la suite $(\theta_k)_{k\geq 1}$ fournie par l'algorithme EM vers le maximum de la vraisemblance a été largement étudiée. Citons les travaux de Dempster et al. [41], complétés quelques années plus tard par ceux de Wu [42]. Nous présentons un résultat de convergence de l'algorithme EM proposé par Delyon et al. [43] qui obtiennent pour le cadre particulier des modèles de type exponentiel des hypothèses plus simples que celles de Wu.

Soient les hypothèses

(M1) L'espace des paramètres Θ est un ouvert de $\mathbb{R}^p$. La fonction de vraisemblance des données complètes s'écrit

$$p(y,x;\theta) = \exp\left\{-\psi(\theta) + \left\langle \tilde{S}(x), \phi(\theta)\right\rangle\right\}, \tag{2.2}$$

où $\langle .,. \rangle$ est le produit scalaire, $\tilde{S}(.)$ est une fonction sur $\mathbb{R}^{\mathbb{J}}$ prenant ses valeurs dans un ouvert $\mathcal{S}$ de $\mathbb{R}^m$. Pour tout $\theta \in \Theta$, $\int \left|\tilde{S}(x)\right| p(x|y;\theta)dx < \infty$.

(M2) Soit la fonction $L : \mathcal{S} \times \Theta \to \mathbb{R}$ définie par $L(s;\theta) = -\psi(\theta) + \langle s, \phi(\theta)\rangle$. Les fonctions $\psi(\theta)$ et $\phi(\theta)$ sont deux fois continuement différentiables sur Θ.

(M3) La fonction $\bar{s} : \Theta \to \mathcal{S}$ définie par $\bar{s}(\theta) = \int \tilde{S}(x)p(x|y;\theta)dx$ est continuement différentiable sur Θ.

(M4) La fonction $l(\theta)$ est continuement différentiable sur Θ et $\partial_\theta \int p(y,x;\theta)dx = \int \partial_\theta p(y,x;\theta)dx$.

(M5) Il existe une fonction $\hat{\theta} : \mathcal{S} \to \Theta$ telle que $\forall\theta \in \Theta, \forall s \in \mathcal{S}, L(s,\hat{\theta}(s)) \geq L(s,\theta)$. De plus, la fonction $\hat{\theta}(s)$ est continuement différentiable sur $\mathcal{S}$.

Théorème 1 ***(Delyon et al [43])*** *Soit $\mathcal{L}$ l'ensemble des points stationnaires de la fonction $\log p(y;\theta)$: $\mathcal{L} = \{\theta \in \Theta; \partial_\theta \log p(y;\theta) = 0\}$. Sous les hypothèses (M1)-(M5), et en supposant que pour tout $\theta \in \Theta$, l'adhérence de $\mathcal{L}(\theta)$ est un sous-ensemble compact de Θ, pour tout point initial $\theta_0 = \theta$, la suite $(\log p(y;\theta_k))_{k\leq 1}$ est croissante et $\lim_{k\to\infty} d(\theta_k, \mathcal{L}(\theta)) = 0$.*

L'implémentation pratique de l'algorithme EM peut être rendue difficile par trois sources de problèmes différents : le calcul de la quantité $Q(\theta|\theta')$ qui n'est souvent pas connue analytiquement, la maximisation en θ de cette fonction qui, lorsque Q n'est pas connue, est complexe, et enfin la convergence de l'algorithme qui peut être relativement lente. Pour le deuxième problème, lorsque la maximisation directe de $Q(\theta|\theta_k)$ est délicate, on peut procéder par des accroissements successifs en choisissant la valeur θ_{k+1} telle que $Q(\theta_{k+1}|\theta_k) \geq Q(\theta_k|\theta_k)$. Une solution est apportée par Lange qui utilise une itération d'un algorithme de Newton-Raphson au cours de l'étape M, permettant de plus l'accélération de la convergence de EM [44].

Les problèmes persistants sont donc principalement liés aux difficultés de calcul de l'intégrale définissant $Q(\theta|\theta_k)$ et qu'ainsi au temps de calcul nécessaire à la convergence de l'algorithme dans les applications pratiques. Ces deux types de problèmes peuvent être résolus par le développement de versions stochastiques de l'algorithme EM.

L'algorithme MCEM

Lorsque l'espérance conditionnelle de la log-vraisemblance complète ne peut pas être calculée, Wei et Tanner [45] proposent l'algorithme MCEM (*Monte Carlo EM*), qui consiste à l'itération k de l'algorithme EM à approcher l'intégrale $Q(\theta|\theta_k)$ par une méthode de Monte Carlo. Sur le même principe que la version de Monte Carlo de l'algorithme de Newton-Raphson, à chaque itération, un nombre T de variables aléatoires $(x_k^t)_{1\leq t\leq T}$ est simulé à partir de la loi a posteriori $p(x|y;\theta_k)$ et la fonction Q est approchée par une moyenne empirique

$$Q(\theta|\theta_k) \approx \frac{1}{T}\sum_{t=1}^{T} \log p(y, x_k^t; \theta_k).$$

Wei et Tanner [45] illustrent leur article par simulation sur deux exemples, mais aucun résultat théorique de convergence n'est présenté. Un algorithme MCEM est aussi proposé par Meng and Rubin [46].

L'algorithme MCEM peut avoir des problèmes numériques, tels qu'une convergence lente ou même inexistante. Comme pour la version MC de l'algorithme de Newton-Raphson, le nombre de réplications T a un impact important sur la convergence et son choix reste un problème ouvert. De plus, la simulation d'un nombre important de réalisations engendre une nette augmentation du temps de calcul. Par exemple Booth et Hobert [47] implémentent un MCEM avec un échantillon de taille $T = 60\,000$ pour les dernières itérations, afin d'assurer la convergence numérique de l'algorithme, ce qui nécessite un temps de calcul particulièrement élevé. D'autres auteurs ont besoin de plus d'un million de simulations au total pour assurer cette convergence.

La simulation sous la loi $p(x|y;\theta_k)$ reste délicate pour de nombreux problèmes, en particulier pour les modèles non-linéaires mixtes. Walker [48] et Wu [49, 50] combinent un algorithme MCEM avec une procédure MCMC de simulation des données non observées. Ils rapportent les mêmes problèmes de convergence numérique et ne proposent aucun résultat théorique. Leary [51] et Guzy et al. [52] ont également présenté des variantes de l'algorithme MCEM adaptées aux modèles non-linéaires mixtes.

L'algorithme SAEM

Une alternative à la fois au problème de convergence presque sûre mais aussi au problème numérique repose sur une approximation stochastique de l'étape E et a été proposée par Delyon et al. et Kuhn et Lavielle [43, 30]. Dans ce cas, l'étape E de cet algorithme, appelé algorithme SAEM (*Stochastic Approximation EM*), est divisée en une étape de simulation et une étape d'approximation stochastique. A l'itération k, l'étape de simulation S consiste à simuler une réalisation x_k des données manquantes x d'après la loi conditionnelle $p(x|y;\theta_k)$. L'étape d'approximation stochastique SA introduit une suite décroissante de pas positifs (γ_k) et réalise l'approximation stochastique suivante à l'itération k

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k(\log p(y, x_k; \theta_k) - Q_k(\theta)),$$

où x_k est simulé sous la loi $p(x|y,\theta_k)$. L'étape M devient

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} Q_{k+1}(\theta).$$

Sous l'hypothèse (M1) d'un modèle du type exponentiel, il suffit de réaliser l'approximation stochastique sur une statistique exhaustive $\tilde{S}$ du modèle.

Le résultat de convergence, obtenu par Delyon et al. [43], nécessite, outre les hypothèses de régularité (M1)-(M5), des hypothèses relatives à la méthode de simulation des données manquantes x et à la suite de pas (γ_k).

On suppose que les variables aléatoires $x_1, \ldots, x_k$ sont définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$. On définit $\mathcal{F} = \{\mathcal{F}_k\}_{k\geq 0}$ la famille de σ-algèbre croissante générée par les variables aléatoires $x_1, \ldots, x_k$.

Soient les hypothèses :

(SAEM 1) Pour tout $k > 0$, $0 \leq \gamma_k \leq 1$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ et $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$

(SAEM 2) $l : \Theta \to \mathbb{R}$ et $\hat{\theta} = \mathcal{S} \to \Theta$ sont m fois différentiables,

(SAEM 3)
1. Pour toute fonction positive borélienne Φ, $E\left[\Phi(x_{k+1})|\mathcal{F}_k\right] = \int \Phi(x)p(x|y;\theta_k)dx$
2. Pour tout $\theta \in \Theta$, $\int \|\tilde{S}(z)\|^2 p(x|y;\theta)dx < \infty$ et la fonction $\alpha(\theta) := \mathrm{Cov}_\theta(\tilde{S}(z))$ est continue par rapport à θ.

Théorème 2 ***(Delyon et al. [43])*** *Sous les hypothèses (M1)-(M5), et (SAEM1)-(SAEM3), si la suite $(s_k)_{k\geq 0}$ prend ses valeurs dans un sous-ensemble compact de $\mathcal{S}$, la suite $(\theta_k)_{k\geq 0}$ générée par SAEM converge vers un point stationnaire de la vraisemblance observée p_y.*

Cet algorithme requiert donc la simulation d'une seule réalisation des données x à chaque itération, ce qui réduit de façon conséquente le temps de calcul par rapport aux algorithmes MCEM. Néanmoins d'un point de vue pratique, la mise en oeuvre de l'algorithme SAEM nécessite de savoir simuler des réalisations de variables aléatoires suivant la loi a posteriori $p(x|y;\theta)$, ce qui n'est pas toujours possible, comme pour les modèles non-linéaires mixtes.

Kuhn et Lavielle ont donc proposé de combiner l'algorithme SAEM avec une méthode de Monte Carlo par chaînes de Markov (MCMC) pour contourner ce problème [30]. Dans ce nouvel algorithme SAEM, seule l'étape S de l'algorithme est modifiée. A l'itération k, l'algorithme SAEM est alors réalisée à travers ces étapes

- Etape E :
  - Etape S : Simulation de x_k par algorithme de Metropolis-Hastings d'après une probabilité de transition Π_{θ_k} correspondant à une chaîne de Markov de loi stationnaire $p(x|y,\theta_k)$
  - Etape SA : Approximation stochastique des statistiques suffisantes du modèle

$$s_{k+1} = s_k + \gamma_k(\tilde{S}(y, x_k) - s_k)$$

– Etape M : Maximisation de la vraisemblance complète

$$\theta_{k+1} = \hat{\theta}(s_{k+1})$$

Le calcul des statistiques suffisantes est détaillé pour les modèles que nous avons considéré dans les chapitres 4, 5, 6.2 et 8.

Kuhn et Lavielle [30] étendent le résultat de convergence de Delyon et al [43], sous l'hypothèse supplémentaire suivante
(SAEM 3')

1. La chaîne $(x_k)_{k\geq 0}$ prend ses valeurs dans un compact $\mathcal{E}$ de $\mathcal{R} \subset \mathbb{R}^d$.
2. Pour tout compact V de Θ, il existe un constante réelle L telle que pour tout (θ, θ') dans V^2

$$\sup_{\{(x),(x')\}\in\mathcal{E}} |\Pi_\theta(x'|x) - \Pi_{\theta'}(x'|x)| \leq L\|\theta - \theta'\|_{\mathbb{R}^q}$$

3. La probabilité de transition Π_θ fournit une chaîne uniformément ergodique dont la probabilité invariante est la distribution conditionnelle $p(\cdot|y;\theta)$

$$\exists K_\theta \in \mathbb{R}^+, \quad \exists \rho_\theta \in]0,1[\quad | \quad \forall k \in \mathbb{N} \quad \|\Pi_\theta^k(\cdot|x) - p(\cdot;|y;\theta)\|_{TV} \leq K_\theta \rho_\theta^k$$

où $\|\cdot\|_{TV}$ est la norme en variation totale, et

$$K = \sup_{\theta\in\Theta} K_\theta < \infty \quad \text{et} \quad \rho = \sup_{\theta\in\Theta} \rho_\theta < 1$$

4. La fonction S_h est bornée sur $\mathcal{E}$.

Théorème 3 ***(Kuhn et Lavielle [30])*** *Sous les hypothèses (M1)-(M5), (SAEM1)-(SAEM2) et (SAEM3'), et si la suite $(s_k)_{k\geq 0}$ prend ses valeurs dans un sous-ensemble compact de $\mathcal{S}$. Alors avec probabilité 1, la suite $(\theta_k)_{k\geq 0}$ générée par SAEM converge vers un maximum (local) de la vraisemblance observée p_y.*

Disposant d'un algorithme performant d'estimation des paramètre des modèles non-linéaires mixtes, nous avons cherché, au cours de cette thèse, à généraliser cet algorithme, et à l'adapter plus particulièrement à la modélisation de la dynamique virale VIH.

Chapitre 3

Principales problématiques issues de l'évaluation de la dynamique virale sous traitement dans l'infection par le VIH

Après avoir présenté dans le chapitre précédent les principales méthodes d'estimation des paramètres des modèles non-linéaires mixtes actuellement proposées, nous présentons brièvement dans cette partie le contexte et les enjeux entourant l'évaluation de la dynamique virale sous traitement dans l'infection par le VIH. Comme nous l'avons rappelé dans le chapitre 1, cette évaluation se réalise autour de deux axes.

Le premier consiste à étudier l'évolution de deux marqueurs biologiques (charge virale et concentration de $CD4^+$) reflétant l'effet du traitement sur la dynamique virale. Autour de ce premier axe se développent plusieurs problèmes. Les tests statistiques étudiant l'influence d'un effet traitement sur la vitesse de la décroissance de la charge virale sont détaillés dans la section 3.1, ainsi que le calcul du nombre de sujets à inclure dans un essai clinique reposant sur l'analyse de données longitudinales. L'une des spécificités de l'étude de la charge virale est l'existence d'une censure des données due à un manque de précision des appareils de mesure lorsque le nombre de copies du virus devient très faible. Ce problème est exposé dans la section 3.2. Enfin, il est indispensable, afin de mieux comprendre la dynamique virale dans toute sa complexité, de modéliser conjointement l'évolution des $CD4^+$ avec celle de la charge virale. Cette dynamique conjointe est décrite par des systèmes différentiels complexes, que nous avons évoqués dans le chapitre 1. Le problème de l'estimation des paramètres d'un système dynamique à partir de l'analyse de données longitudinales est évoqué dans la section 3.3. Son application spécifique à la dynamique virale VIH est détaillée dans la section 3.4.

Enfin, le second axe consiste à étudier la variabilité intra-patient ainsi que les interactions pharmacocinétiques des médicaments impliqués dans cette prise en charge thérapeutique de l'infection par le VIH. Nous présentons dans la section 3.5 le cadre d'étude de ces interactions.

3.1 Tests d'un effet traitement et calcul du nombre de sujets nécessaires

Pour tous les essais cliniques où l'évaluation de l'efficacité du traitement repose sur des mesures répétées au cours du temps d'une variable biologique continue, par exemple la charge virale, une analyse fondée sur un modèle de régression permet de mieux prendre en compte l'ensemble des données disponibles. En particulier, le test de l'effet d'un traitement peut s'effectuer à partir de la modélisation des données longitudinales, ce qui diffère des tests de comparaison classiques qui reposent principalement sur les mesures finales (ajustées ou non sur les valeurs initiales). Récemment, Jonsson et Sheiner ont montré et discuté la pertinence de cette approche par modélisation en s'appuyant sur un exemple provenant du développement d'un médicament [53].

L'analyse de ces données longitudinales étant réalisée à travers l'estimation des paramètres de modèles non-linéaires mixtes, la mise en place de ces tests statistiques de comparaison de groupes dépend, pour le test de Wald et le test du rapport de vraisemblance respectivement, du choix de l'estimateur de la matrice de Fisher et de la vraisemblance d'un modèle non-linéaire mixte. Nous avons montré la complexité de l'estimation des paramètres de ces modèles dans le chapitre 2. L'estimation de la matrice de Fisher et de la vraisemblance de ces modèles est également délicate.

L'estimation de la matrice de Fisher, proposée par Delyon et al. [43] et Kuhn et Lavielle [30] repose sur le principe de Louis [36], que nous avons rappelé dans la section 2.3.2, reliant les fonctions du gradient et hessienne de la vraisemblance observée $p(y;\theta)$ et celles de la vraisemblance des données complètes $p(y,x;\theta)$

$$\begin{aligned}\partial_\theta \log p(y;\theta) &= E\left[\partial_\theta \log p(y,x;\theta)|y,\theta\right] \\ \partial^2_\theta \log p(y;\theta) &= E\left[\partial^2_\theta \log p(y,x;\theta)|y,\theta\right] + \mathrm{Var}\left[\partial_\theta \log p(y,x;\theta)|y,\theta\right].\end{aligned}$$

Ces fonctions du gradient et hessienne de la vraisemblance des données complètes sont alors évaluées par approximation stochastique dans l'algorithme SAEM. Kuhn et Lavielle [30] ont montré la convergence de cet estimateur vers la matrice d'information de Fisher observée.

Concernant l'estimation de la vraisemblance, Kuhn et Lavielle ont proposé une évaluation par méthode de Monte-Carlo. Une étude de simulation, réalisée lors de

mon DEA de biostatistiques en 2003, a montré que cet estimateur avait une grande variance, et n'était donc pas utilisable pour la mise en place du test du rapport de vraisemblance. La méthode classique pour réduire cette variance est d'utiliser un estimateur par échantillonnage préférentiel

$$\widehat{\log p}_T(y;\theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{p(y|x^{(t)};\theta)p(x^{(t)};\theta)}{h(x^{(t)};\theta)}$$

où $x^{(t)} \sim h(\cdot;\theta)$, h étant une loi instrumentale à choisir avec soin en fonction du modèle. A partir d'un estimateur convenable de cette vraisemblance, il est alors possible de mettre en place le test du rapport de vraisemblance.

Nous avons donc développé, pour les modèles mixtes, des tests de Wald et du rapport de vraisemblance reposant sur ces estimations de la matrice de Fisher et de la vraisemblance. Nous avons également proposé une méthode de calcul du nombre de sujets nécessaires dans un modèle mixte, pour la planification d'un test d'une covariable binaire, par exemple un effet traitement. Nous avons illustré par simulation nos résultats sur un modèle de décroissance de charge virale après initiation d'un traitement anti-VIH. Ces deux travaux font l'objet d'un article soumis à *Statistics in Medicine* et présenté dans le chapitre 4.

3.2 Prise en compte des données censurées par une limite de quantification

Comme nous l'avons déjà souligné dans l'introduction de ce chapitre, dans la réalité, les données de charge virale sont sujettes à une limite de détection liée aux dispositifs expérimentaux. En effet les appareils de mesure sont calibrés pour rendre des résultats de mesure de concentration fiables, c'est-à-dire ayant une variabilité inférieure à 20%. Il existe un seuil en dessous duquel deux mesures d'une même concentration varient généralement de plus de 20 %, la concentration étant trop faible pour être mesurée avec précision. Dans ce cas, on sait seulement que la concentration du marqueur est en dessous du seuil (appelée aussi *limite de quantification*, LOQ), mais on ne connaît pas sa valeur. Ce problème est extrêmement fréquent dans de nombreux domaines (pharmacologie, dynamique virale, etc). Par exemple, la mesure d'une charge virale n'est pas disponible dès qu'elle se situe en dessous de la limite de quantification qui s'établit généralement entre 20 et 400 copies/mL selon les appareils (la valeur moyenne en phase d'infection aiguë est de l'ordre de 10^5 copies/mL, et généralement inférieure à 100 copies/mL sous traitement). Les données observées sont alors censurées à gauche, ce qui complique leur analyse statistique.

Ce phénomène est illustré sur la figure 3.1. Les données non observées sont celles mesurées sous la limite de quantification (limite représentée par la ligne en poin-

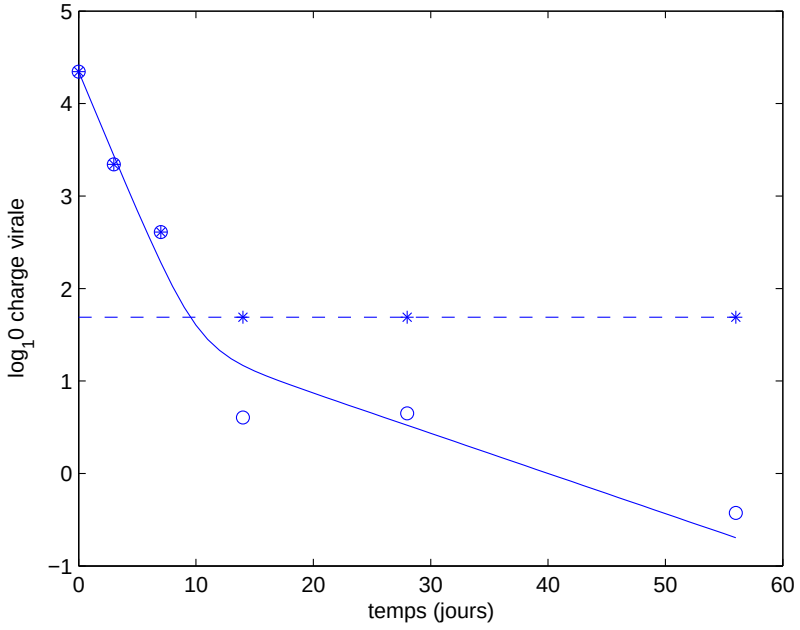


Figure 3.1: Données de décroissance de charge virale : $(*)$ données observées, $(\circ)$ données censurées, $(-)$ courbe individuelle, $(--)$ valeur limite de quantification.

tillé). L'observation des données réelles (représentées par des cercles) permettrait de prédire la courbe individuelle réelle (en trait plein) par des méthodes d'estimation pour modèles mixtes classiques. Dès qu'il y a des données censurées (représentées à la valeur LOQ par des étoiles sur la figure 3.1), des méthodes statistiques spécifiques doivent être développées pour estimer les paramètres de ces modèles.

Dans la pratique on observe les données suivantes $y_{ij}^{obs} = y_{ij}$ si $y_{ij} \geq LOQ$, et $y_{ij}^{obs} = LOQ$ si $y_{ij} \leq LOQ$. Si on note y_i^{cens} les valeurs inconnues du sujet i, les données non observées sont alors $x = (y_i^{cens}, \phi_i)_{1 \leq i \leq N}$, où ϕ_i est le i-ème paramètre individuel. La vraisemblance des données observées du modèle s'écrit dans ce cas

$$L(y^{obs}\,;\,\theta) = \log\left(\prod_{i=1}^{N} \int p(y_i^{obs}, y_i^{cens}, \phi_i; \theta)\, d\phi_i\, dy_i^{cens}\right). \tag{3.1}$$

Plusieurs procédures ont été proposées pour contourner ce problème. La procédure la plus naïve consiste à omettre l'ensemble des données censurées. D'autres méthodes proposent de remplacer ces données par une valeur fixe, habituellement la valeur LOQ ou LOQ/2 mais ces procédures induisent un biais dans l'estimation des paramètres. Pour éviter ce biais, certains auteurs recommandent de ne garder que la première valeur censurée imputée à LOQ/2 [54]. Cependant les propriétés statistiques de ces procédures ne sont pas établies.

Des approches plus habiles ont été proposées comprenant des imputations multiples des données censurées, c'est-à-dire la substitution de valeurs raisonnables pour chaque donnée censurée, en tenant compte du mécanisme de censure.

Différentes méthodes ont été proposées dans le cas des modèles linéaires mixtes

$$y_{ij} = X_i\mu + Z_i b_i + \varepsilon_{ij}$$

Hughes [55] propose un algorithme MCEM. L'étape E de son algorithme consiste à calculer

$$\begin{aligned} E(b_i b_i^t | y_i^{obs}, \theta) &= \int E(b_i b_i^t | y_i^{cens}, y_i^{obs}, \theta)\ p(y_i^{cens} | y_i^{obs}, \theta)\ dy_i^{cens} \\ E(\varepsilon_i \varepsilon_i^t | y_i^{obs}, \theta) &= \int E(\varepsilon_i \varepsilon_i^t | y_i^{cens}, y_i^{obs}, \theta)\ p(y_i^{cens} | y_i^{obs}, \theta)\ dy_i^{cens} \end{aligned}$$

par approximation de Monte Carlo, en simulant les données censurées $(y_i^{cens})_{i=1,\ldots,N}$ dans une loi normale tronquée à droite à l'aide d'un algorithme de Gibbs. Il montre que sa méthode réduit considérablement les biais associés aux autres méthodes d'imputations naïves (imputation à la valeur LOQ, LOQ/2 ou imputation aléatoire) sur une étude de simulation d'un modèle de décroissance de charge virale VIH. Cependant il ne montre pas la convergence de son algorithme.

Jacqmin-Gadda et al. [5] proposent une maximisation directe de la vraisemblance, utilisant un algorithme itératif, combinant deux algorithmes d'optimisation (le simplex et l'algorithme de Marquardt). Le calcul de la vraisemblance, qui comprend l'intégrale multiple d'une distribution multi-normale des données censurées, est évaluée numériquement par une méthode transformant l'intégrale en une intégrale sur un hyper-cube. Leur algorithme est adapté au cas d'un modèle linéaire mixte avec modèle de variance-covariance incluant un processus stochastique gaussien w_i

$$y_{ij} = X_i\mu + Z_i b_i + w_i(t_{ij}) + \varepsilon_{ij}.$$

Sur une étude de simulation d'un modèle de décroissance de charge virale VIH, ces auteurs obtiennent de meilleurs résultats que ceux obtenus par la méthode d'imputation à la valeur LOQ ou par l'algorithme MCEM de Hughes. Ils font état de problèmes de convergence numérique de l'algorithme MCEM, alors que leur algorithme converge à chaque fois.

Concernant les méthodes non-linéaires mixtes, Wu et Wu [56] proposent une méthode d'imputation basée sur l'algorithme FOCE, sans résultat théorique de convergence. Wu propose ensuite un algorithme MCEM approché, basé sur une linéarisation du modèle [49]. A l'itération k, l'algorithme se décompose en deux étapes

– Etape E : approximation de Monte Carlo de

$$E(\phi | y, y^{cens}; \theta_k) \approx \frac{1}{T} \sum_{t=1}^{T} E(\phi | y, y_t^{cens}; \theta_k)$$

et

$$E(\phi\phi^t|y, y^{cens}; \theta_k) \approx \frac{1}{T}\sum_{t=1}^{T} E(\phi\phi^t|y, y_t^{cens}; \theta_k)$$

où $(y_t^{cens})_{1\leq t\leq T}$ est simulé par algorithme de Gibbs dans la distribution $p(y^{cens}|y, \phi; \theta_k)$. Les espérances conditionnelles $E(\phi|y, y_t^{cens}; \theta_k)$ et $E(\phi\phi^t|y, y_t^{cens}; \theta_k)$ sont calculées directement après linéarisation du modèle.

– Etape M : linéarisation de la fonction de régression f en ϕ_k et maximisation directe en θ du modèle linéaire mixte associé.

Il s'agit donc d'une méthode d'estimation à comparer aux méthodes FO et FOCE présentées dans la section 2.2.1 et ne maximisant pas la vraisemblance du modèle initial. Tout comme Hughes, Wu ne propose aucun résultat de convergence.

Nous avons adapté l'algorithme SAEM au problème des données censurées afin de disposer d'une méthode aux propriétés statistiques établies. Nous l'avons utilisée pour l'analyse de la décroissance de charge virale de l'essai clinique TRIANON-ANRS 81 de l'Agence Nationale de Recherche sur le Sida (ANRS). Ce travail est présenté dans la section 5 sous la forme d'un article accepté pour publication dans la revue *Computational Statistics and Data Analysis* (sous réserve de modifications mineures).

3.3 Modèles dynamiques mixtes

Les traitements anti-rétroviraux étant de plus en plus efficaces, la charge virale est dès la deuxième ou la troisième semaine après l'initiation du traitement, ramenée sous la limite de quantification des appareils de mesure. L'évaluation à long terme de l'efficacité de ces traitements ne peut donc pas reposer uniquement sur l'analyse de la décroissance de la charge virale dont la part de censure peut atteindre rapidement 40 ou 50 % du nombre total de données.

Le deuxième marqueur biologique en jeu dans la dynamique virale est la concentration de $CD4^+$ dans le sang. Ce marqueur ayant une grande variabilité, même parmi les sujets sains, il ne peut pas être utilisé seul pour évaluer l'efficacité des traitements anti-rétroviraux. En revanche son analyse conjointe à celle de la charge virale permet de mieux prendre en compte toute la complexité de la dynamique virale et d'augmenter le nombre de données disponibles pour l'analyse, et ainsi de diminuer le pourcentage de données censurées.

Comme nous l'avons vu dans la section 1 de cette thèse, ce processus de la dynamique virale du VIH est décrit par des systèmes dynamiques plus ou moins complexes, ordinaires ou stochastiques, sans solution analytique en général. L'analyse de ces données longitudinales est alors réalisée par un modèle mixte dont la fonction de régression n'est pas connue analytiquement. Le problème de l'estimation des paramètres de tels modèles est présenté dans la section 3.3.1 pour les équations

différentielles ordinaires et dans la section 3.3.2 pour les équations différentielles stochastiques.

3.3.1 Systèmes différentiels ordinaires

Nous considérons le modèle mixte suivant

$$\begin{aligned} y_{ij} &= f(\phi_i, t_{ij}) + \ \varepsilon_{ij}, \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2 I_{n_i}), \\ \phi_i &= X_i\mu + b_i, \text{où} \ \ b_i \sim \mathcal{N}(0, \Omega), \end{aligned}$$

où $f : \mathbb{R}\times\mathbb{R}^k \longrightarrow \mathbb{R}^d$ est défini comme la solution de l'équation différentielle ordinaire suivante

$$\begin{aligned} &\frac{\partial f(t,\phi)}{\partial t} = F(f(t,\phi), t, \phi) \\ &f(t_0, \phi) = f_0(\phi) \end{aligned} \tag{3.2}$$

où la fonction $F : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ est connue ainsi que la condition initiale $f_0(\phi) \in \mathbb{R}^d$, $t \in [t_0, T]$.

Pour l'estimation des paramètres de ces modèles, les logiciels NONMEM et `nlme-ode` [57] intègrent une méthode numérique de résolution d'équations différentielles dans les algorithmes FO, FOCE.

Nous avons adapté l'algorithme SAEM à l'estimation de paramètres de modèles mixtes dont la fonction de régression est solution de systèmes différentiels, en intégrant un schéma de linéarisation locale optimisé pour résoudre le système différentiel. La section 6.1 est consacrée à ces modèles, sous la forme d'un article soumis à *Journal of Statistical Planning and Inference*.

3.3.2 Systèmes différentiels stochastiques

Certains auteurs ont montré les limites de la modélisation par des systèmes différentiels ordinaires [58, 59, 60]. Les systèmes différentiels stochastiques permettent de modéliser les erreurs résiduelles temporelles corrélées, dues par exemple à des erreurs de dosage ou une mauvaise spécification du modèle. Une variabilité supplémentaire est alors introduite dans le système dynamique, qui est distincte à la fois de l'erreur de mesure et de la variabilité inter-individuelle. Jacqmin-Gadda et al. [5] proposent un modèle linéaire combiné à un modèle stochastique auto-régressif pour décrire la décroissance de la charge virale. Tornoe et al. [60] montrent que ce bruit supplémentaire permet de modéliser des variations dans les paramètres pharmacocinétiques au cours du temps et d'obtenir une meilleure adéquation des prédictions aux données observées.

On considère dans ce cas le modèle mixte suivant

$$\begin{aligned} y_{ij} &= Z(t_{ij}, \phi_i) + \ \varepsilon_{ij}, \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2 I_{n_i}), \\ \phi_i &= X_i\mu + b_i, \text{où} \ \ b_i \sim \mathcal{N}(0, \Omega), \end{aligned}$$

où la fonction de régression Z est un processus de diffusion défini par l'équation stochastique suivante

$$\begin{aligned} dZ(t) &= F(Z, t, \phi)\, dt + \gamma dB(t)\ , \\ Z(t_0, \Phi) &= Z_0(\phi), \end{aligned}$$

où B est un mouvement brownien, F est la fonction de *dérive* connue, γ est le coefficient de *volatilité* et $t \in [0, T]$. Les conditions assurant l'existence d'une solution de l'équation différentielle stochastique sont supposées. Le but est d'estimer $\theta = (\mu, \Omega, \sigma^2, \gamma^2)$ à partir du vecteur d'observations discrètes $(y_i)_{1 \leq i \leq N}$.

Pour certaines équations différentielles stochastiques particulières (par exemple un processus de Ornstein-Uhlenbeck), la solution est explicite, mais pour des processus plus complexes, ce n'est plus le cas. Différentes méthodes d'estimation des paramètres d'un processus observé discrètement ont été proposées, reposant sur une approximation de la loi de distribution du processus. Genon-Catalot et al. [61] proposent un estimateur consistant et asymptotiquement normal. Gloter et Jacod [62] ont développé un estimateur de contraste, consistant et asymptotiquement normal. Les autres méthodes proposées reposent sur une approximation d'Euler-Maruyama du processus de diffusion.

Alors que l'estimation des paramètres d'un processus de diffusion observé discrètement à été largement étudiée, très peu d'auteurs se sont intéressés à l'estimation des paramètres d'un processus observé à la fois de façon discrète *et* bruitée. Overgaard et al. et Tornoe et al. [59, 60] ont développé une méthode d'estimation de ces modèles mixtes basée sur l'algorithme FOCE et utilisant un filtre de Kalman pour la reconstruction du processus de diffusion. Toutefois, ils ne proposent aucun résultat statistique de convergence de leur algorithme d'estimation.

Nous avons généralisé l'algorithme SAEM aux équations différentielles stochastiques. La section 6.2 présente en détail cette méthode, sous la forme d'un article soumis à *Scandinavian Journal of Statistics*.

Les travaux présentés dans les sections 6.1 et 6.2 ont été réalisés en collaboration avec Sophie Donnet, doctorante à l'université Paris Sud.

3.4 Modélisation de la dynamique du VIH par modèles mixtes

Comme nous l'avons déjà souligné, il nous a paru intéressant et important de modéliser conjointement l'évolution à long terme de la charge virale et de la concentration de $CD4^+$, afin de mieux appréhender la dynamique complexe de l'infection par le VIH et de mieux évaluer les traitements proposés.

Dans les deux premiers travaux de cette thèse, nous avons analysé la décroissance de la charge virale seule à l'aide d'une fonction bi-exponentielle. Cette fonction a été proposée par Ding et Wu [6] comme solution simplifiée d'un système différentiel décrivant le processus de l'infection des cellules immunitaires par le virus, sous l'hypothèse très forte d'une concentration de $CD4^+$ constante au cours du temps. Cette hypothèse n'est réaliste que sur une courte période d'observation, et ne peut pas être appliquée dès lors qu'est considérée la dynamique virale à long terme, ou que la concentration de $CD4^+$ est modélisée conjointement. Dans ces cas où l'hypothèse de Ding et Wu [6] n'est plus acceptable, le système dynamique n'a plus de solution analytique.

La modélisation conjointe de la décroissance de la charge virale du VIH et de la croissance des lymphocytes $CD4^+$ sous traitement anti-rétroviral est très peu répandue dans la littérature, l'estimation de telles données par modèles mixtes étant complexe. Cette modélisation conjointe a été proposée pour la première fois par Thiebaut et al. [9] avec un modèle linéaire mixte à deux dimensions, prenant en compte les données censurées de charge virale et dont les paramètres sont estimés par maximum de vraisemblance. Ils ont ensuite complété ce travail en modélisant les processus de sorties de l'étude par un modèle de survie [63].

Cependant, comme nous l'avons rappelé plus haut, le processus de l'infection des cellules immunitaires par le virus est décrit par un système différentiel multidimensionnel, aboutissant à une solution bi-dimensionnelle non-linéaire plus adaptée que la fonction linéaire utilisée par Thiebaut et al. [9]. En particulier, l'utilisation de ces systèmes dynamiques permet d'obtenir une estimation des paramètres du système dont l'interprétation biologique est cruciale pour mieux comprendre le mécanisme de l'infection par le VIH et la réponse au traitement. Ces modèles sont en général sans solution analytique, et les méthodes d'estimation adaptées à l'analyse de tels modèles sont peu nombreuses, la convergence de ces méthodes étant extrêmement difficile à obtenir numériquement sur ces exemples. Putter et al. [8] ont proposé l'analyse de données de dynamique virale à partir d'un modèle d'équations différentielles en utilisant une méthode d'estimation bayésienne, soulignant qu'aucune méthode d'estimation par maximum de vraisemblance à disposition des utilisateurs n'est capable de relever ce défi algorithmique. Ils ont utilisé un système différentiel simple, peu réaliste, et se sont donc limités à l'analyse des données lors

des deux premières semaines après l'initiation du traitement. Les articles les plus récents, proposant des systèmes différentiels complexes, ne font état que de la modélisation des données de charge virale, probablement pour des raisons de problème de convergence d'algorithme d'estimation. De plus, à notre connaissance, excepté Thiebaut et al. [9, 63] et Putter et al. [8], les autres auteurs ne se préoccupent pas des données censurées des charges virales. Les résultats actuels ne sont donc pas satisfaisants.

Nous avons combiné les différentes méthodologies développées dans cette thèse pour l'analyse par modèle mixte de l'évolution simultanée de la décroissance de la charge virale VIH et la croissance des lymphocytes $CD4^+$, décrite par système dynamique, et prenant en compte les données censurées de charge virale. Nous avons appliqué cette méthode à l'analyse des données de l'essai COPHAR II-ANRS 111. Les résultats de cette analyse sont présentés dans le chapitre 7 sous forme d'un article en préparation pour *Antiviral Therapy*.

3.5 Méthodologie de l'étude des interactions pharmacocinétiques médicamenteuses

Les traitements anti-rétroviraux dans l'infection par le VIH impliquent actuellement la combinaison d'au moins trois médicaments, en général un inhibiteur de protéase et deux inhibiteurs de la transcriptase inverse. Il est donc important d'étudier les interactions pharmacocinétiques entre ces différentes molécules.

Lors d'essais dits d'*interaction*, la pharmacocinétique du médicament étudié est mesurée au cours de plusieurs périodes d'observations successives, chaque patient étant alors pris comme son propre témoin afin de mieux évaluer l'une des sources principales de variabilité, la variabilité intra-individuelle. Le tableau 3.1 présente un essai permettant d'étudier les interactions pharmacocinétiques du traitement B sur le traitement A (de référence), lors de deux périodes d'observations.

Table 3.1: Essai à deux périodes et une séquence

	Période 1	Période2
Traitement administré	A	A+B

Un exemple fictif de données longitudinales recueillies lors de ces essais est représenté sur la figure 3.2, le phénomène d'interaction y est artificiellement amplifié dans un but didactique. La concentration de la molécule A est mesurée lors des deux périodes d'observations, quand elle est administrée seule ou avec la molécule B. L'interaction entre ces deux molécules permet de ralentir l'élimination de la molécule A et ainsi de prolonger son activité.

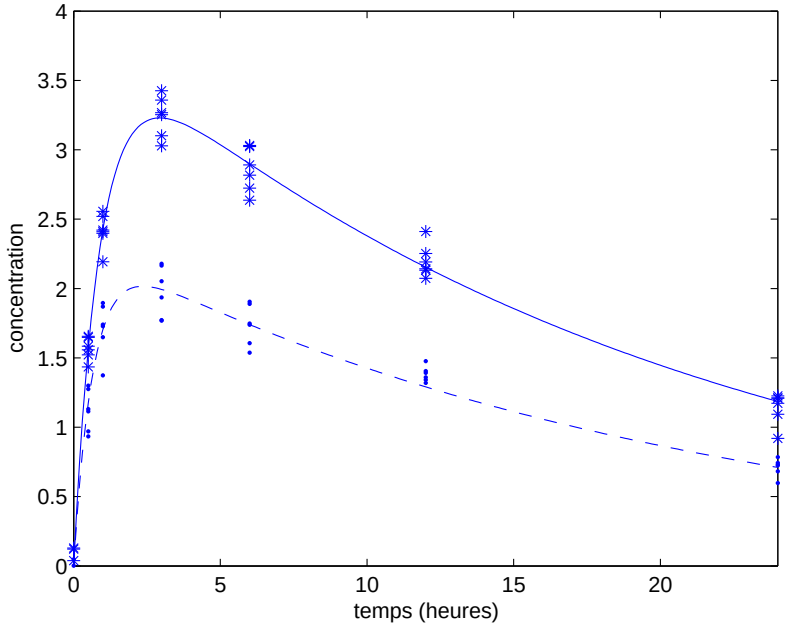


Figure 3.2: Concentrations de la molécule A administrée seule (observations en points et courbe moyenne en trait pointillé) et co-administrée avec la molécule B (observations en étoiles et courbe moyenne en trait plein).

D'autre part, Netlles et al. [64] ont très récemment montré l'existence d'une grande variabilité intra-individuelle de la concentration des inhibiteurs de protéase. Ces essais à plusieurs observations permettent de l'évaluer et de la distinguer de la variabilité résiduelle.

Dans ce contexte, les données sont notées $y_{ik} = (y_{i1k}, \ldots, y_{in_ik})^t$ où y_{ijk} est la mesure de la variable y pour le sujet i à l'instant t_{ij} et à l'occasion $k = 1, \ldots, K$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. On considère le modèle non-linéaire mixte suivant

$$
\begin{aligned}
&y_{ijk} = f(\phi_i, t_{ij}) + g(\phi_{ik}, t_{ij})\ \varepsilon_{ijk}, \\
&\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2), \\
&\phi_{ik} = X_i\mu + b_i + c_{ik}, \text{où}\ \ b_i \sim \mathcal{N}(0, \Omega), c_{ik} \sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

où b_i représente l'effet aléatoire "patient" et c_{ik} l'effet aléatoire "période" pour chaque sujet. Ces modèles permettent de distinguer trois sources de variabilités : la variabilité *inter*-sujet (Ω), la variabilité *intra*-sujet (Ψ), aussi appelée *inter*-occasion (entre les périodes d'observations), et enfin, la variabilité résiduelle liée à l'erreur de mesure de la concentration du médicament (σ^2).

Les logiciels NONMEM et `nlme` proposent l'estimation des paramètres de ces modèles, à partir d'algorithmes de linéarisation de la fonction de régression. Cependant, les utilisateurs font état de problèmes numériques de convergence très fréquents, obligeant en général à réduire l'analyse initialement prévue (diminution du nombre

d'effets aléatoires modélisés, simplification du modèle, etc).

Nous avons proposé une version de l'algorithme SAEM adaptée aux essais à plusieurs périodes de suivi. Nous avons utilisé cette méthode pour étudier l'interaction du ténofovir, un inhibiteur nucléotidique de la transcriptase inverse, sur la pharmacocinétique de l'atazanavir, un inhibiteur de protéase, en analysant des données de concentration recueillies lors de l'essai Puzzle 2- ANRS 107. Ce travail, réalisé en collaboration avec Xavière Panhard, post-doctorante à l'unité INSERM U738, est présenté dans la section 8.

Deuxième partie

Travaux et résultats

Chapitre 4

Tests et calcul de puissance fondés sur l'algorithme SAEM dans les modèles non-linéaires à effets mixtes

Nous proposons deux tests de comparaison, le test de Wald et le test du rapport de vraisemblance, permettant d'évaluer la significativité d'une différence d'efficacité entre deux traitements à partir de l'analyse de données longitudinales du critère de jugement. Ces tests sont fondés sur une analyse statistique par modèles non-linéaires à effets mixtes utilisant l'algorithme d'estimation SAEM.

La mise en place du test de Wald nécessite l'évaluation des erreurs standards des estimateurs des paramètres, calculées à partir de la matrice de Fisher du modèle non-linéaire mixte. Comme l'ont proposé Delyon et al. [43] et Kuhn et Lavielle [30], nous utilisons le principe de Louis [36] pour estimer la matrice de Fisher à partir des gradients et hessiens de la vraisemblance des données complètes. Cette estimation est réalisée par approximation stochastique au cours des itérations de l'algorithme SAEM.

Pour utiliser le test du rapport de vraisemblance, la vraisemblance du modèle doit être évaluée respectivement sous les hypothèses nulle et alternative du test. Nous proposons un estimateur par échantillonnage préférentiel

$$\widehat{\log p_T}(y;\theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{p(y|\phi^{(t)};\theta)p(\phi^{(t)};\theta)}{h(\phi^{(t)};\theta)}$$

où $\phi^{(t)} \sim_{iid} h_(.;\theta)$, pour $t = 1,\ldots,T$ et h est une approximation gaussienne de la distribution a posteriori des paramètres individuels. Cet estimateur a de meilleures propriétés que l'estimateur par approximation de Monte Carlo proposé par Kuhn et Lavielle [30].

Nous avons évalué le risque de première espèce de ces tests par une étude de simulation dans le cadre d'un modèle de décroissance de la charge virale sous traite-

ment dans l'infection par le VIH. Nous avons utilisé un modèle bi-exponentiel pour décrire ce processus

$$f(\phi, t) = P_1 \exp(-\lambda_1 t) + P_2 \exp(-\lambda_2 t)$$

où le vecteur de paramètres individuels est log-paramétré $\phi = (\log P_1, \log P_2, \log \lambda_1, \log \lambda_2)$ pour assurer la positivité des estimations des paramètres $(P_1, P_2, \lambda_1, \lambda_2)$. Les résultats de cette étude de simulation ont illustré les propriétés de convergence de l'algorithme SAEM. Les erreurs de type I des deux tests sont proches du seuil nominal de 5%.

Nous avons également proposé une méthode de calcul du nombre de sujets nécessaires pour assurer une puissance donnée à un test de Wald d'une covariable binaire, par exemple le test d'un effet traitement. Cette méthode permet la planification d'essais cliniques comparant deux traitements et dont l'analyse sera réalisée par modèles mixtes. Étant donnés une fonction de régression, les valeurs des paramètres θ, un plan d'expérience (temps de prélèvements), ainsi que la différence d'efficacité attendue entre les deux traitements qui seront comparés, cette méthode repose sur l'évaluation de la matrice de Fisher attendue $H(\theta)$

$$H(\theta) = E\left(H_{obs}(y;\theta)|y;\theta\right)$$

où $H_{obs}(y;\theta)$ est la matrice de Fisher observée pour un jeu de données y.

Nous avons proposé de simuler un jeu de données sous ces conditions avec un grand nombre d'individus (plusieurs milliers), qui est ensuite analysé avec l'algorithme SAEM. Un estimateur de la matrice de Fisher observée de ce grand jeu de données est ainsi obtenu, qui est une approximation de Monte Carlo de la matrice de Fisher attendue. L'erreur standard (SE) attendue du paramètre d'effet de la covariable binaire est alors calculée à partir de la diagonale de l'inverse de la matrice de Fisher. La SE étant proportionnelle à la racine carré du nombre de sujets, le nombre de sujets nécessaires permettant d'à assurer une puissance donnée est directement déduit du calcul de cette SE. Nous avons illustré cette méthode dans le cadre du modèle bi-exponentiel, pour un test d'effet traitement β sur le paramètre de la première pente de la décroissance de la charge virale $\log \lambda_1$. Ces résultats sont présentés dans un article soumis à *Statistics in Medicine*.

STATISTICS IN MEDICINE

Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model

Adeline Samson[1,*] Marc Lavielle[2], France Mentré[1]

[1] *INSERM U738, Paris, France; University Paris 7, CHU Bichat-Claude Bernard, Biostatistics unit, Paris, France*

[2] *University Paris-Sud, Bat. 425 Orsay, France*

SUMMARY

Nonlinear mixed-effects models (NLMEM) are used to improve information gathering from longitudinal studies and are applied to treatment evaluation in disease evolution studies, such as HIV or HBV viral dynamics, prostate cancer evolution, etc. Statistical tests especially in NLMEM for comparing different treatment groups are critical issues. We propose two group comparison tests, the Wald test and the likelihood ratio test (LRT), based on the Stochastic Approximation EM estimation algorithm (SAEM). SAEM is an alternative method to approximate estimation methods of which asymptotic convergence is not proved. We propose to estimate the Fisher information matrix by using stochastic approximation and the likelihood by using importance sampling, both required to perform the tests. We evaluate these SAEM-based methods on a simulation study in the context of HIV viral load decrease after initiation of an antiretroviral treatment. Results from this simulation illustrate the theoretical convergence properties of SAEM. Lastly, we propose a method based on the SAEM algorithm and on the estimate of the Fisher information matrix by stochastic approximation to compute the minimum sample size required to carry out a Wald test of a treatment effect on a NLMEM.

KEYWORDS: Likelihood ratio test; Wald test; longitudinal data; Nonlinear mixed effects models; SAEM algorithm; sample size.

*Correspondence to: Adeline Samson, INSERM U738, Biostatistics unit, CHU Bichat-Claude Bernard, 46 rue Huchard, 75018 Paris, France

Received

1. Introduction

Most clinical trials aim at comparing the efficacy of two different treatments randomly assigned to two groups of patients. To assess whether one treatment achieve a better reduction of the disease than the other, several biological endpoints are repeatedly measured along the trial extent. The statistical approaches usually used to compare the two treatment groups are classically based only on the final measurements of this longitudinal data. As these methods do not exploit the richness of the dynamics seized by repeated measurements, mixed-effects models are developed to improve information extraction yield from longitudinal studies. However, due to the complexity of the observed treatment responses, these dynamics are frequently nonlinear with respect to the parameters and this leads to the use of nonlinear mixed-effects models (NLMEM). Such models are developed for disease evolution studies when a biological marker can be used as a surrogate endpoint. Several applications can be stated, for instance, the efficacy of anti-viral treatments in HIV [1, 2, 3, 4] or HBV [5] infections may be evaluated through measures of viral load evolution, or prostate cancer treatment may be assessed from prostate specific antigen dosage [6]. NLMEM are also used to model the evolution of functional markers, measured by clinical examination or patient interview, as for instance the functional capacity decay in rheumatoid arthritis-suffering patients [7], or the evolution of the ventilatory function in patients with asthma [8].

Comparison of treatment effects based on longitudinal data analysis is critical. The properties of the statistical tests used to perform this comparison are based on the maximum likelihood theory. However, because of the nonlinearity of the regression function in the random effects, the likelihood of NLMEM cannot be expressed in a closed form. This leads to the development of several widely used likelihood approximation methods. Linearization algorithms realize a first order linearization of the conditional mean and a zero order linearization of the conditional variance with respect to the random effects as in the First Order and First Order Conditional Estimate algorithms [9, 10] implemented in NONMEM software and in the nlme function of Splus and R software. Other methods are based on the Laplacian or Gaussian quadrature algorithms which implement the corresponding classical numerical quadrature methods as in NLMIXED Macro of SAS software [11]. However none of these methods can be considered as fully established theoretically. Vonesh

gives an example of a specific design resulting in inconsistent estimates, such as when the number of observations per subject does not increase faster than the number of subjects [12] or when the variability of random effects is too large [13]. Particularly, convergence assumptions on which the statistical tests are based, are not fulfilled. For instance, several authors show an inflation of the type I error of the most used group comparison tests, the Wald test and the Likelihood Ratio Test (LRT) by simulation [14, 15, 16, 17]. Thus, methods with proved convergence and consistency for finding the maximum likelihood estimate in NLMEM are required.

Recently, several estimation methods are proposed as alternatives to linearization algorithms. Importance sampling is a common method to handle numerical integrations. However, as emphasized by Ge el al. [13], in order to achieve satisfactory numerical stability, this method could be computationally intensive, and hence numerically less efficient than many other parametric methods. The most adapted tool to estimate models with missing or non-observed data such as random effects is the Expectation-Maximization (EM) algorithm [18]. The widespread popularity of the EM is largely due to its monotonicity: the likelihood is always increasing. Furthermore, the convergence of the EM algorithm is widely studied [18]. Because of the nonlinearity of the model, stochastic versions of the EM algorithm are proposed. Wei et al. [19]; Walker [20] and Wu [21, 22] propose MCEM algorithms, with a Monte Carlo approximation of the expectation of the sufficient statistics in the E-step. This Monte-Carlo implementation is based on samples independently and identically distributed from the conditional density, requiring MCMC procedures. However the MCEM algorithm may have computational problems, such as slow or even no convergence. The replications choice of the Monte Carlo sample is a central issue to guarantee convergence and this remains an open problem. Furthermore, simulations of the large samples at each iteration are time consuming. Celeux and Diebolt propose a SEM algorithm, the first stochastic version of EM, which can be viewed as a special case of the MCEM [23]. However the estimate sequence generated by this SEM algorithm does not converge pointwise. As an alternative to address both the pointwise convergence and the computational problem, stochastic approximation versions of EM are proposed [24, 25, 26]. This algorithm requires a simulation of only one realization of the missing data at each iteration. In addition, pointwise almost sure convergence of the estimate

sequence to a local maximum of the likelihood is proved by Delyon et al. [24] under conditions satisfied by models from an exponential family. Kuhn and Lavielle [27] propose to combine the SAEM algorithm with a Monte Carlo Markov Chain (MCMC) procedure adapted to the NLMEM, and they prove that the produced estimates are convergent and consistent.

The first objective of this paper is to propose statistical tests for NLMEM based on the cited above SAEM approach. The Wald test statistic requires the computation of the standard errors (SE) of the parameter. Since the diagonal of the inverse of the Fisher information matrix provides an upper bound of the SE, we propose an estimate of this Fisher matrix based on the Louis' principle [28] and on the stochastic approximation procedure. The LRT requiring the computation of the likelihood, we propose to estimate it using an importance sampling procedure. We then implement these methods and evaluate them on a simulation study of the HIV infection dynamics.

Group comparison tests for NLMEM requires the availability of easy to perform methods of minimum required sample size determination. Most clinical trials aim at observing differences in parameter values between two different treatment groups. To be able to identify statistically significant difference between the two groups, the number of subjects required in each group has to be computed to ensure a given power. Sample sizes greater than required waste research resources while inadequate sample sizes do not allow definitive conclusions or may lead to an improper conclusion. Kang et al. [29] propose a method to compute sample sizes given a test hypothesis with NLMEM using the Wald test statistic. To evaluate the expected Fisher information matrix which is required to perform this computation, they use a first-order linearization of the model as proposed by Retout et al [30].

The second objective of this paper is to propose an alternative to these linearization-based approaches with no accurate statistical properties, to compute the minimum sample size for a given power using the Wald test statistic with NLMEM. We propose a method to estimate the expected Fisher information matrix without linearization, based on the SAEM approach.

After describing the model and notations (section 2), section 3 describes the SAEM algorithm. Especially, we detail how to estimate the Fisher information matrix and the likelihood. At last, we propose a sample size computation method for NLMEM. Section 4 reports the simulation study

and its results. We simulate datasets from the bi-exponential model for HIV dynamics proposed by Ding and Wu [14], and evaluate the statistical properties of the SAEM parameters estimates, and the standard errors and the likelihood estimates. In a second time, we evaluate the type I error of the comparison tests for a treatment effect on one parameter. Finally, the sample size computation method is illustrated on the same HIV dynamics example. Section 5 concludes the article with some discussion.

2. Models and notations

Let us define $y_i = (y_{i1}, \ldots, y_{in_i})^t$ where y_{ij} is the response value for individual i at time t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, and let us define $y = (y_1, \ldots, y_N)$. Let define a NLMEM as follows:

$$\begin{aligned} y_{ij} &= f(\phi_i, t_{ij}) + g(\phi_i, t_{ij})\, \varepsilon_{ij}, \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2 I_{n_i}), \\ \phi_i &= \mu X_i^t + b_i, \text{with} \;\; b_i \sim \mathcal{N}(0, \Omega), \end{aligned}$$

where $f(\cdot)$ and/or $g(\cdot)$ are nonlinear functions of ϕ, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^t$ represents the residual error, ϕ_i is a p-vector of individual regression parameters, μ is the $p \times k$-matrix of fixed effects, X_i is the k-vector of known covariates, b_i is a p-vector of random effects independent of ε_i, σ^2 is the residual variance, I_{n_i} the identity matrix of size n_i and Ω quantifies the variance matrix of the inter-individual random effects.

The maximum likelihood estimation in NLMEM is based on the log-likelihood function $L(y\,;\,\theta)$ of the response y, with $\theta = (\mu, \Omega, \sigma^2) \in \Theta$ the vector of all the parameters of the model. This function is equal to:

$$L(y\,;\,\theta) = \sum_{i=1}^{N} L(y_i\,;\,\theta) = \sum_{i=1}^{N} \log\left(\int p(y_i, \phi_i; \theta)\, d\phi_i\right), \tag{1}$$

where $p(y_i, \phi_i;\, \theta\,)$ is the likelihood of the complete data (y_i, ϕ_i) of the i subject and is equal to $p(y_i, \phi_i; \theta) = \prod_{j=1}^{n_i} p(y_{ij}|\phi_i; \theta) p(\phi_i; \theta)$. As the random effects ϕ_i are unobservable and the regression functions are nonlinear, the foregoing integral (1) has no closed form.

3. Estimation algorithm and statistical tests

3.1. The SAEM algorithm

The EM algorithm is a classical approach for estimating parameters of model with non-observed or incomplete data [18]. For NLMEM, the non-observed vector is the individual parameter vector $\phi = (\phi_1, \ldots, \phi_N)$ and the complete data of the model is (y, ϕ). Let us define the function $Q(\theta|\theta') = E(L_c(y, \phi; \theta)|y; \theta')$, where $L_c(y, \phi; \theta)$ is the log-likelihood of the complete data. At the m^{th} iteration of the EM algorithm, the E step is the evaluation of $Q_m(\theta) = Q(\theta \,|\, \widehat{\theta}_m)$, while the M step updates $\widehat{\theta}_m$ by maximizing $Q_m(\theta)$. For cases where the E step has no closed form, Delyon et al. [24] introduce a stochastic version of the EM algorithm which evaluates the integral $Q_m(\theta)$ by a stochastic approximation procedure. They prove the convergence of this SAEM algorithm under general conditions if $L_c(y, \phi; \theta)$ belongs to a regular curved exponential family:

$$L_c(y, \phi; \theta) = -\Lambda(\theta) + \langle S(y, \phi), \Phi(\theta) \rangle$$

where $\langle ., . \rangle$ is the scalar product and $S(y, \phi)$ is known as the minimal sufficient statistics of the complete data model. The E step is then divided into a simulation step (S step) of the non-observed data $\phi^{(m)}$ under the conditional distribution $p(\phi|y; \widehat{\theta}_m)$ and a stochastic approximation step (SA step) of $\mathbb{E}\left[S(y, \phi)|\widehat{\theta}_m\right]$:

$$s_{m+1} = s_m + \gamma_m (S(y, \phi^{(m)}) - s_m), \tag{2}$$

where $(\gamma_m)_{m \geq 0}$ is a sequence of positive numbers decreasing to 0. The M step is thus the update of the estimate $\widehat{\theta}_m$:

$$\widehat{\theta}_{m+1} = \arg\max_{\theta \in \Theta} \left(-\Lambda(\theta) + \langle s_{m+1}, \Phi(\theta) \rangle \right).$$

However, the simulation step can be complex when the posterior distribution $p(\phi|y; \theta)$ has no analytical form, such as with NLMEM. Therefore a Monte Carlo Markov Chain procedure such as the Metropolis-Hastings algorithm can be used to simulate $\phi^{(m)}$. At the m-th iteration of the SAEM algorithm, the S step is thus the simulation of $\phi^{(m)}$ using a Metropolis-Hastings algorithm which constructs a Markov Chain with $p(\phi|y; \widehat{\theta}_m)$ as the unique stationary distribution (see [26]

for more details).

Kuhn and Lavielle [26] present the details of the SAEM implementation and prove that under general hypotheses, the sequence $(\widehat{\theta}^{(m)})_{m\geq 0}$ obtained by this algorithm converges almost surely towards a (local) maximum of the likelihood $L(y;\cdot)$.

3.2. Estimation of the Fisher Information matrix using Louis' principle

To perform the Wald test and to compute the required minimum sample size, we have to evaluate the standard errors (SE) of the parameters. The diagonal of the inverse of the information Fisher matrix provides an upper bound of the variance of the parameter estimates, namely their SE.

We adapt the estimate, proposed by Delyon et al. [24], of the Fisher information matrix using the fact that the gradient and the Hessian of the log-likelihood function L can be obtained almost directly from the simulated non-observed data ϕ. Using the Louis' missing information principle [28], the Hessian of $L(y;\theta)$ may be expressed as

$$\partial^2_\theta L(y;\theta) = E\left[\partial^2_\theta L_c(y,\phi;\theta)\right] + \mathrm{Var}\left[\partial_\theta L_c(y,\phi;\theta)\right].$$

The Jacobian of $L(y;\theta)$ is the conditional expectation of the complete data likelihood:

$$\partial_\theta L(y;\theta) = E\left[\partial_\theta L_c(y,\phi;\theta)|y,\theta\right].$$

For NLMEM, these derivatives $\partial_\theta L_c(y,\phi;\theta)$ and $\partial^2_\theta L_c(y,\phi;\theta)$ have analytical forms. Therefore we implement this SE estimate using stochastic approximation during the SAEM algorithm as proposed by Kuhn et Lavielle [26]. At the m^{th} iteration of the algorithm, we evaluate the three following quantities

$$\begin{aligned}
\Delta_{m+1} &= \Delta_m + \gamma_m\left(\partial_\theta L_c(y,\phi^{(m)};\theta) - \Delta_m\right),\\
G_{m+1} &= G_m + \gamma_m\left(\partial^2_\theta L_c(y,\phi^{(m)};\theta) + \partial_\theta L_c(y,\phi^{(m)};\theta)\partial_\theta L_c(y,\phi^{(m)};\theta)^t - G_m\right),\\
H_{m+1} &= G_{m+1} - \Delta_{m+1}\Delta^t_{m+1}.
\end{aligned}$$

As the SAEM algorithm converges, $(-H_{m+1})$ is an estimate of the Fisher information matrix.

3.3. Estimation of the likelihood using importance sampling

To perform the LRT, we have to evaluate the likelihood of the observations. For any $\theta \in \Theta$, Kuhn and Lavielle [26] propose a simple Monte Carlo procedure to estimate $L(y;\theta)$. The estimate of the likelihood of the ith subject is thus:

$$\hat{L}_T(y_i;\theta) = \frac{1}{T}\sum_{t=1}^{T} p(y_i|\phi_i^{(t)};\theta),$$

with $\phi_i^{(t)} \sim_{iid} \mathcal{N}(\mu,\Omega)$, for $t = 1\cdots T$. By the strong law of large numbers, this estimate $\hat{L}_T(y;\theta)$ converges almost surely towards $E[L(y;\theta)]$. However, this Monte Carlo estimate may be susceptible to numerical instabilities and to computational precision issues [31].

To avoid these numerical problems, we propose to estimate the likelihood using an importance sampling procedure. The likelihood function is rewritten as:

$$L(y_i;\theta) = \int \frac{p(y_i|\phi_i;\theta)p(\phi_i;\theta)}{h_i(\phi_i;\theta)} h_i(\phi_i;\theta)d\phi_i$$

where $h_i(.;\theta)$ is any instrumental distribution. The importance sampling estimates of the likelihood is thus

$$\hat{L}_T(y_i;\theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{p(y_i|\phi_i^{(t)};\theta)p(\phi_i^{(t)};\theta)}{h_i(\phi_i^{(t)};\theta)}$$

with $\phi_i^{(t)} \sim_{iid} h_i(.;\theta)$, for $t = 1\cdots T$. This estimate converges for the same reason that the regular Monte Carlo estimate converges, whatever the choice of the distribution h_i. However, there are obviously some choices of h_i that are better than others, especially to reduce the variance of the estimate. Among the distributions h leading to finite variances for the estimate of the likelihood, it is possible to exhibit the optimal distribution that minimizes the estimates variance [32]. For NLMEM, this distribution is the individual conditional distribution $p(\phi|y_i;\theta)$. However, as this distribution has no closed form in NLMEM, we propose for h_i a Gaussian approximation of the i^{th} individual posterior distribution, i.e., $\phi_i^{(t)} \sim \mathcal{N}(\mu_i^{post},\Omega_i^{post})$. For each i, the posterior individual mean μ_i^{post} and the posterior individual variance Ω_i^{post} are estimated by the empirical mean and empirical variance of the $\phi_i^{(m)}$ simulated by the MCMC procedure during the last 250 iterations of the SAEM algorithm.

3.4. Statistical tests for group comparison

When NLMEM are used to compare two treatment groups, a treatment effect on some of the some of the fixed effects of the model is tested. Let $G_i = 0$ denote a control treatment group subject and $G_i = 1$ an experiment treatment group subject. To simplify the explanation, let us assume that we test a scalar treatment effect β on the k^{th} fixed effect which is modeled on the k^{th} component ϕ_{ik} of ϕ_i by:

$$\phi_{ik} = \mu_k + \beta G_i + b_{ik}.$$

This notation can easily be extended to a vector β of treatment effects. Let q denote the length of the vector β.

Hence, the null hypothesis to test is H_0: $\{\beta = 0\}$ while the alternative hypothesis is H_1: $\{\beta \neq 0\}$. Both the Wald test and the LRT can be performed to assess this difference between the two groups.

For the Wald test, we estimate with SAEM the parameter β and its standard error $SE(\beta)$ under H_1 and compare the statistic $\beta^2/SE(\beta)$ to a χ^2_q distribution with q degrees of freedom. For the LRT, we evaluate the log-likelihoods L_0 and L_1 by the importance sampling procedure respectively under H_0 and H_1 and compare the $2(L_1 - L_0)$ statistic with a q degrees of freedom χ^2_q distribution.

3.5. Sample size computation

This section aims at proposing a method to compute the minimum sample size required for a Wald test using NLMEM. This sample size computation requires to proceed through the following steps: to specify the regression function and the NLMEM to be used; to identify values for the parameter θ; to specify an experimental design $(t_{ij})_{i,j}$; to identify the minimum difference to test, i.e. the alternative hypothesis H_1; to evaluate the standard errors SE of θ, and finally, given the power and the type I error, to compute the required number of subjects N. These last two steps are detailed below.

To simplify the explanation, let the tested parameter β be a scalar treatment effect on one fixed effect. Assume the null hypothesis is H_0: $\{\beta = \beta_0\}$. Thereby, for a clinical trial aiming at detecting a difference between the two treatment groups of at least $(\beta_1 - \beta_0)$ on this fixed effect, the alternative hypothesis is H_1: $\{\beta \geq \beta_1\}$.

The statistic of the Wald test is $S_W(\hat{\beta}) = (\hat{\beta} - \beta_0)^2/SE(\hat{\beta})$, where $\hat{\beta}$ is the estimate of the β parameter. To ensure a type I error α for the Wald test under H_0, the rejection region is $\{S_W > \chi^2_{1;1-\alpha}\}$, where $\chi^2_{1;1-\alpha}$ is the critical value of the centered χ^2_1 distribution. Under H_1, the statistic $S_W(\hat{\beta})$ is asymptotically distributed with a non-centered χ^2_1 distribution with a non-centrality parameter $(\beta - \beta_0)^2/SE(\beta)$, with $\beta \geq \beta_1$. Therefore, the power of the Wald test is defined as

$$p(\beta) = \int_{\chi^2_{1;1-\alpha}}^{\infty} d(x; 1, (\beta - \beta_0)^2/SE(\beta))dx \tag{3}$$

where $d(x; 1, c)$ is the probability density function of the non-centered χ^2_1 distribution with a non-centrality parameter c. This function $p(\beta)$ is a non-decreasing function of β. Therefore, to ensure a given power under H_1, the minimum sample size required is evaluated from $p(\beta_1)$. This method can be extended to a vector β, see Kang et al. for more details [29].

To compute the expected standard error $SE(\beta)$ for a NLMEM, Kang et al. propose a method based on the linearization of the regression function [29]. As an alternative to this linearization, we propose to use the estimate of the Fisher information matrix provided by the SAEM algorithm detailed in section 3.2. For explanation's sake, all the patients are subjected to the same sampling design. A dataset with a treatment effect $\beta = \beta_1$ is then generated with this sampling design and with a number N_{sim} of subjects large enough to ensure a fine approximation of the expected SE by the observed SE. The estimation of the parameters and of the observed Fisher information matrix is performed on this simulated dataset using the SAEM algorithm. Given the hypothesis of an identical sampling design for each subject hypothesis, the Fisher information matrix of the complete dataset is the sum of the individual Fisher information matrices. Last, the SE of N subjects can be evaluated from the SE of the simulated dataset using $SE_N(\beta) = SE_{N_{sim}}(\beta) \cdot \sqrt{N_{sim}/N}$.

4. Simulation study

4.1. Simulation settings

This simulation study aims at illustrate some statistical properties of the SAEM algorithm in the context of HIV viral dynamics. We evaluate the accuracy of the parameters estimates, the

SE and the likelihood estimates, and at last we perform the two group comparison tests. We use the bi-exponential model for initial HIV dynamics proposed by Ding and Wu [14] to simulate the datasets:

$$f(\phi_i, t_{ij}) = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{ij}} + P_{2i}e^{-\lambda_{2i}t_{ij}}).$$

This model is a simplified analytical solution of a differential system describing HIV viral load decrease during antiretroviral treatment [33]. It has p=4 individual parameters: P_{1i}, P_{2i} are the baseline values and λ_{1i}, λ_{2i} represent two-phase viral decay rates. To prevent these parameters from taking unrealistic negative values, they are assumed to be distributed using a log-normal distribution. Thus, ϕ_i and μ are defined as: $\phi_i = (\ln \mathrm{P}_{1i}, \ln \mathrm{P}_{2i}, \ln\lambda_{1i}, \ln \lambda_{2i})$ and $\mu = (\ln \mathrm{P}_1, \ln \mathrm{P}_2, \ln\lambda_1, \ln \lambda_2)$. We assume identical sampling times for all subjects: for all i in $1, \ldots, N$, $t_{ij} = t_j$ for $j = 1, \ldots, n$. Additive Gaussian random effects are assumed for each parameter with a diagonal variance-covariance matrix Ω. Let denote $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2)$ the vector of the variances of the random effects. Additive Gaussian error is assumed with a constant variance σ^2, i.e, $g(\phi_i, t_j) = 1$, for all i, j.

For the fixed effects, we choose the values proposed by Ding and Wu [14]: $\ln\mathrm{P}_1 = 12$, $\ln\mathrm{P}_2 = 8$, $\ln\lambda_1 = \ln(0.5)$, $\ln\lambda_2 = \ln(0.05)$. We set inter-subject variability to be identical for the four parameters: $\omega_1^2 = \omega_2^2 = \omega_3^2 = \omega_4^2 = 0.3$ corresponding to a variation coefficient of 55%, which is a realistic inter-subject variability in the context of HIV dynamics. A variance $\sigma = 0.065$ is chosen, this corresponds to a constant variation coefficient of 15% for the viral load. We generate twice 1000 trials with respectively $N = 40$ and $N = 200$ total number of subjects and with n=6 blood samples per patient, taken on days 1, 3, 7, 14, 28 and 56. A simulated dataset with $N = 40$ subjects is represented on figure 1, the mean decrease is overlaid on the individual data.

[Figure 1 about here.]

To evaluate the group comparison tests, we consider that the subjects of each simulated trial belong to two different treatment groups of equal size, i.e 20 or 100 subjects per group when respectively $N = 40$ or $N = 200$ subjects. We perform both the Wald test and the LRT to assess a difference between the treatment groups on the viral load decrease, with a scalar treatment effect

β on the first viral decay rate, $\ln\lambda_1$, as proposed by Ding and Wu [14]. The parameter vector θ is $\theta = (\mu, \omega^2, \sigma^2)$ under H_0 and $\theta = (\mu, \beta, \omega^2, \sigma^2)$ under H_1.

Let us detail on this example the sufficient statistics to be evaluated at the SA step of the SAEM algorithm under H_1. The sufficient statistics are the two vectors $S^{(1)} = \sum_{i=1}^N \phi_i$ and $S^{(2)} = \sum_{i=1}^N \phi_i^2$ and the two scalars $S^{(3)} = \sum_{i=1}^N \phi_{i3} G_i$ and $S^{(4)} = \sum_{i,j} (y_{ij} - f(\phi_i, t_j))^2$. At the m-th iteration of SAEM, the M-step is reduced to

$$\begin{aligned}
\widehat{\mu}_{k,m+1} &= \frac{s^{(1)}_{k,m+1}}{N}, \quad \text{for} \quad k = 1, 2, 4 \\
\widehat{\mu}_{3,m+1} &= \frac{s^{(1)}_{3,m+1} - s^{(3)}_{m+1}}{N/2}, \\
\widehat{\beta}_{m+1} &= \frac{s^{(3)}_{m+1}}{N/2} - \widehat{\mu}_{3,m+1}, \\
\widehat{\omega^2}_{k,m+1} &= \frac{s^{(2)}_{k,m+1}}{N} - (\widehat{\mu}_{k,m+1})^2, \quad \text{for} \quad k = 1, 2, 4 \\
\widehat{\omega^2}_{3,m+1} &= \frac{s^{(2)}_{3,m+1}}{N} - \frac{(\widehat{\mu}_{3,m+1})^2}{2} - \frac{(\widehat{\beta}_{m+1} + \widehat{\mu}_{3,m+1})^2}{2}, \\
\widehat{\sigma^2}_{m+1} &= \frac{s^{(4)}_{m+1}}{Nn}.
\end{aligned}$$

The M-step of SAEM under hypothesis H_0 can be deducted from the foregoing equations.

4.2. Simulation results

To evaluate the accuracy of the parameter estimates provided by the SAEM algorithm, the simulation model (ie under H_0) is fitted on the simulated datasets and the relative bias and relative root mean square error (RMSE) for each component of θ are computed. The relative bias and RMSE on the 1000 data sets obtained for $N = 40$ and $N = 200$ subjects are presented in table 1.

[Table 1 about here.]

For $N = 40$ subjects, the estimates have very low bias (<0.5% for the fixed effects, <5% for the variance parameters). The RMSE are really satisfactory for the fixed effects (<13%) as well as for the variance parameters (<30%). As expected with $N = 200$ subjects, both the bias and the RMSE decrease with the increase of the subjects number.

We compare the SE estimated for each component of $\widehat{\theta}$ by SAEM and the procedure detailed in 3.2, with the 'true' SE evaluated by the empirical standard deviation of the 1000 parameter estimates obtained on the 1000 simulated datasets. On figure 2, for each component of θ, the 1000 estimated SE (%) and the true SE with $N = 40$ subjects datasets are plotted.

[Figure 2 about here.]

For all parameters, the SE estimated by SAEM are very close to the true SE. Similar results are observed with $N = 200$ subjects datasets. Thus the SE estimated by the stochastic approximation procedure are very accurate. In order to evaluate the likelihood by importance sampling as detailed in section 3.3, the size T of the random samples must be determined. We study the estimate variability in function of the value of T on one dataset with $N = 200$ subjects. We evaluate 10 times the log-likelihood successively for different sample sizes $T = 1,000$ or $5,000$ or $10,000$ or $50,000$, using the Gaussian approximation of the individual posterior distribution.

[Figure 3 about here.]

Results are reported in figure 3 and show that the variability of the approximation is reduced by increasing the sample size. Therefore we evaluate the likelihood using the importance sampling procedure with a sample size T = 10,000, as a balance between estimate accuracy and time consuming.

As the simulation is performed with $\beta = 0$ ie without any treatment effect, the type I error of both the Wald test and the LRT is evaluated by the proportion of trials for which H_0 is rejected. The estimations of the type I errors for a nominal value of 5% are given in table 2 for datasets with 20 or 100 subjects per group.

[Table 2 about here.]

the estimated type I errors are closed to 5% and given the dataset number of replications, do not differ significantly from the expected 5% value. Once more, as these tests are based on the SE and the likelihoods, this illustrastes the accuracy of the estimates of the SE using stochastic approximation and of the likelihoods using importance sampling.

4.3. Sample size computation example

The method proposed in section 3.5 is applied to the model, the parameter values and the sampling design detailed above. We assume that the clinical trial aim is to detect a difference of at least 30% between the two treatment groups on the parameter $\ln \lambda_1$, namely H_0: $\{\beta = 0\}$ and H_1: $\{\beta \geq \beta_1\}$ with $\beta_1 = 0.262$. A dataset with $N = 10,000$ subjects is simulated, split up into two groups of equal size and with a treatment effect $\beta_1 = 0.262$ on $\ln \lambda_1$. This dataset is then analyzed using the SAEM algorithm to evaluate the observed Fisher information matrix. A $SE(\hat{\beta}) = 0.011$ is obtained for $5,000$ subjects per group. Applying the equation (3), a sample size of 20 subjects per group ($N = 40$) provides a power of 32%, while a sample size of 72 subjects per group ($N = 144$) ensures a power superior to 80%. Complementary results are provided in table 3.

[Table 3 about here.]

These results illustrate the ability of the SAEM approach to compute the minimum sample size required for a Wald test.

5. Discussion

This paper substantially improves the accuracy of group comparison tests adapted to longitudinal data analysis. These tests are based on NLMEM and take into account all the data, while classical tests use only the final measurements. We also propose a sound method adapted to NLMEM to compute the minimum required sample size for the Wald test.

The comparison tests that we propose in this paper are an extension of the SAEM algorithm. SAEM is a maximum likelihood approach of NLMEM that we develop and of which the theoretical convergence is proved [26]. This algorithm and the comparison tests proposed in this paper are implemented in a Matlab function called MONOLIX and free on http://mahery.math.u-psud.fr/∼lavielle/monolix. This simulation study illustrates the accuracy of the SAEM algorithm to fit nonlinear longitudinal data in the context of HIV viral load decrease, the parameter estimates being unbiased and with small RMSE.

The two comparison tests (Wald test and LRT) require the evaluation of the standard errors and

the likelihood of the estimates. The standard errors are deduced from the evaluation of the Fisher information matrix. We propose to estimate this matrix using a stochastic approximation procedure based on the Louis' principle [28]. We implement this estimate in the case of NLMEM, the first and second derivatives of the complete data log-likelihood being known analytically. Results of the simulation study show that the SE estimates are very close to the true SE values evaluated on 1000 datasets. Kuhn and Lavielle [26] propose to estimate the likelihood using a simple Monte Carlo procedure, but this method provides poor estimates, prone to computational instabilities. To avoid this problem, we propose here an importance sampling approach, using an approximation of the conditional posterior distribution to sample the individual parameter. Hence, the Wald test and the LRT based on these two estimates (Fisher matrix and likelihood estimates) have accurate properties, especially the obtained type I errors are very close to the expected threshold of 5%.

Another critical issue of NLMEM is the computation of the minimum sample size required to observe a significant treatment effect on a fixed effect parameter using the Wald test. This requires the evaluation of the expected SE of the treatment effect. Kang et al. [29] propose an evaluation based on the linearization of the model, a method to be compared with the pfim function proposed by Retout et al. [30]. As an alternative to linearization, we propose to simulate a large dataset under the alternative hypothesis and to estimate the standard errors using the SAEM approach. The evaluation of the SE through a simulated dataset is irrelevant when the approach is based on linearization, since methods such as those proposed by Retout et al. or Kang et al are analytical. But in order to avoid the linearization step, this simulation is mandatory. We illustrate this method on the HIV dynamics model used for the previously introduced simulation study. Kang et al. find a small underestimation of power when the inter-subject variability increases and justify this as being a consequence of the linearization approach based on a first-order approximation of the model they use. Since our SAEM-based approach sidesteps this linearization, it probably stays clear from this underestimation. Furthermore, Kang et al. show the influence of the experimental sampling design on the estimated sample size. As an example, the use of a D-optimal design computed by the pfim function [30] may significantly reduce the required sample size.

At last, this paper emphasizes the capacity of the SAEM algorithm to provide excellent estimates

when working with NLMEM and supports the idea that it may be applied to even more difficult issues. Namely, when measuring a biological response such as a concentration or a viral load, these observations may be left-censored, due to the limit of quantification of measuring equipment. We already successfully extend the SAEM algorithm to this case (Samson et al., submitted results), the left-censored data being considered as non-observed data as well as the random effects. Similarly, the dynamic model describing longitudinal data such as HIV dynamics can be defined through a differential system with generally no analytical solution. We already successfully extend the SAEM algorithm to this case (Samson et al., submitted results), a numerical solving method being inserted to evaluate the regression function.

references

[1] Wu H, Wu L. A multiple imputation method for missing covariates in nonlinear mixed-effects models with application to HIV dynamics. *Statistics in Medicine* 2001; **20**:1755–1769

[2] Jacqmin-Gadda H, Thiebaut R, Chene G, Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* 2000; **1**:355–68

[3] Wu H, Wu L. Identification of significant host factors for HIV dynamics modelled by non-linear mixed-effects models. *Statistics in Medicine* 2002; **21**:753–71

[4] Wu H, Zhang JT. The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine* 2002; **21**:3655–75. doi:10.1002/sim.1317

[5] Wolters L, Hansen B, Niesters H, de Man R. Viral dynamics in chronic hepatitis B patients treated with lamivudine, lamivudine-famciclovir or lamivudine-ganciclovir. *European Journal of Gastroenterology and Hepatology* 2002; **14**:1007–1011

[6] Law N, Taylor J, Sandler H. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* 2002; **3**:547–563

[7] Welsing P, Van Gestel A, Swinkels H, Kiemeney L, Van Riel P. The relationship between

disease activity, joint destruction, and functional capacity over the course of rheumatoid arthritis. *Arthritis and Rheumatism* 2001; **44**:2009–2017

[8] Lange P, Parner J, Vestbo J, Schnohr P, Jensen G. A 15-year follow-up study of ventilatory function in adults with asthma. *New England Journal of Medicine* 1998; **17**:1194–200

[9] Beal S, Sheiner L. Estimating population kinetics. *Critical Review of Biomedical Engineering* 1982; **8**:195–222

[10] Lindstrom M, Bates D. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**:673–87

[11] Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika* 1993; **80**:791–795

[12] Vonesh EF. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* 1996; **83**:447–452

[13] Ge Z, Bickel P, Rice J. An approximate likelihood approach to nonlinear mixed effects models via spline approximation. *Computational Satistics and Data Analysis* 2004; **46**:747–776

[14] Ding A, Wu H. Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics* 2001; **2**:13–29

[15] Wahlby U, Jonsson E, Karlsson M. Assessment of actual significance levels for covariate effects in NONMEM. *Journal of Pharmacokinetics and Pharmacodynamics* 2001; **28**:231–252

[16] Comets E, Mentré F. Evaluation of tests based on individual versus population modeling to compare dissolution curves. *Journal of Biopharmaceutical Statistics* 2001; **11**:107–123

[17] Panhard X, Mentré F. Evaluation by simulation of tests based on non-linear mixed-effects models in interaction and bioequivalence cross-over trials. *Statistics in Medicine* 2005; **24**:1509–24

[18] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 1977; **39**:1–38

[19] Wei GCG, Tanner MA. Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika* 1990; **77**:649–652

[20] Walker S. An EM algorithm for non-linear random effects models. *Biometrics* 1996; **52**:934–944

[21] Wu L. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of the American Statistical Association* 2002; **97**:955–964

[22] Wu L. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association* 2004; **99**:700–709

[23] Celeux G, Diebolt J. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly* 1985; **2**:73–82

[24] Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 1999; **27**:94–128

[25] Gu MG, Zhu HT. Maximum likelihood estimation for spatial models by Markov Chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society: Series B* 2001; **63**:339–355

[26] Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* 2005; **49**:1020–1038

[27] Kuhn E, Lavielle M. Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S* 2005; **8**:115–131

[28] Louis TA. Finding the observed information matrix when using the EM algorithm ;

[29] Kang D, Schwartz J, Verotta D. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. *Statistics in Medicine* 2004; **23**:2251–2566

[30] Retout S, Mentré F, Bruno R. Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. *Statistics in Medicine* 2002; **21**:2623–39

[31] Robert C, Casella G. *Monte Carlo statistical methods.* Springer-Verlag 2002

[32] Geweke J. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 1989; **57**:1317–1339

[33] Perelson A, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho D. Decay characteristics of HIV-1 infected compartments during combination therapy. *Nature* 1997; **387**:188–191

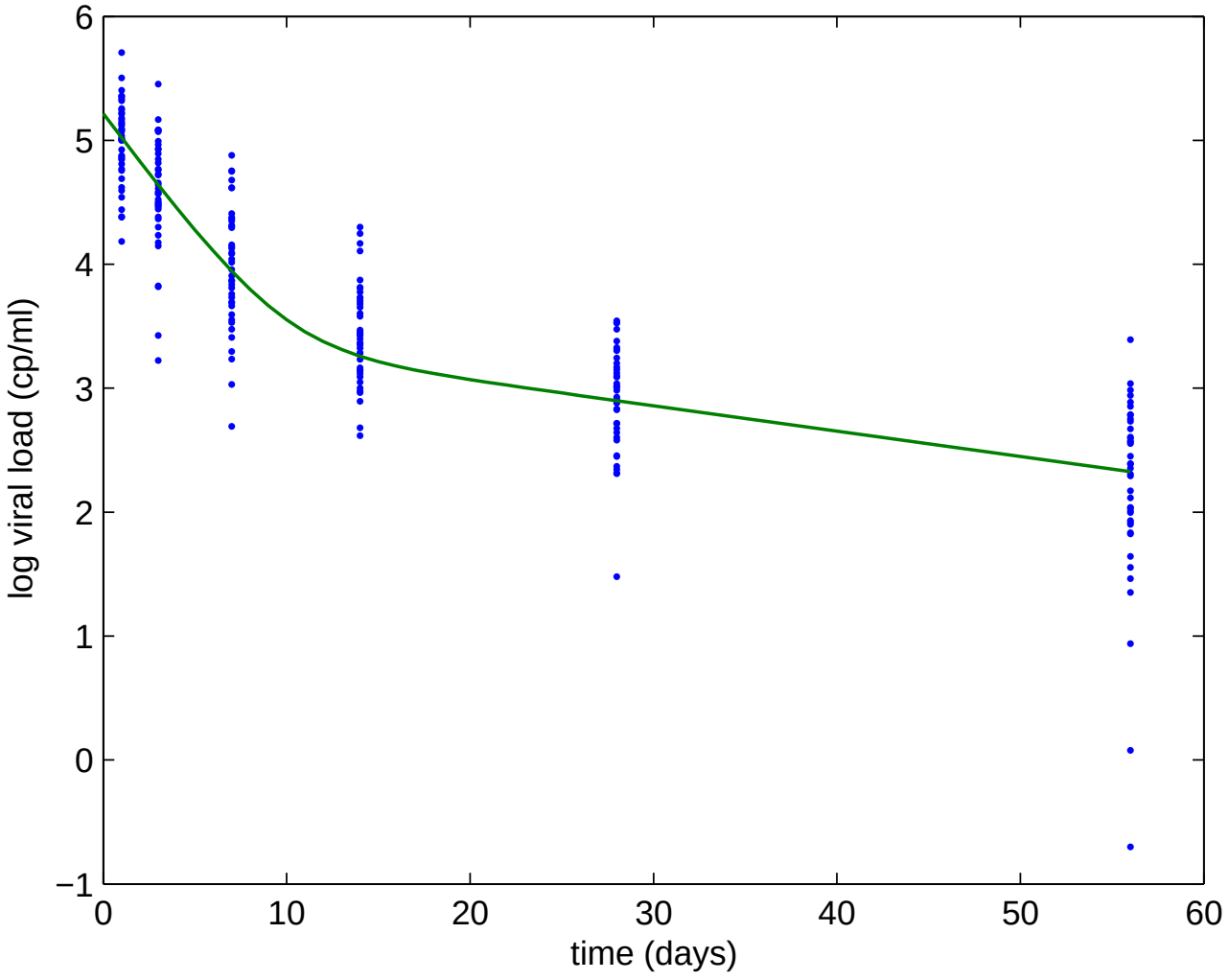


Figure 1. Simulated dataset with $N = 40$ subjects of the biexponential model describing the HIV viral load decrease under treatment: individual observations in dot and the mean decrease in full line.

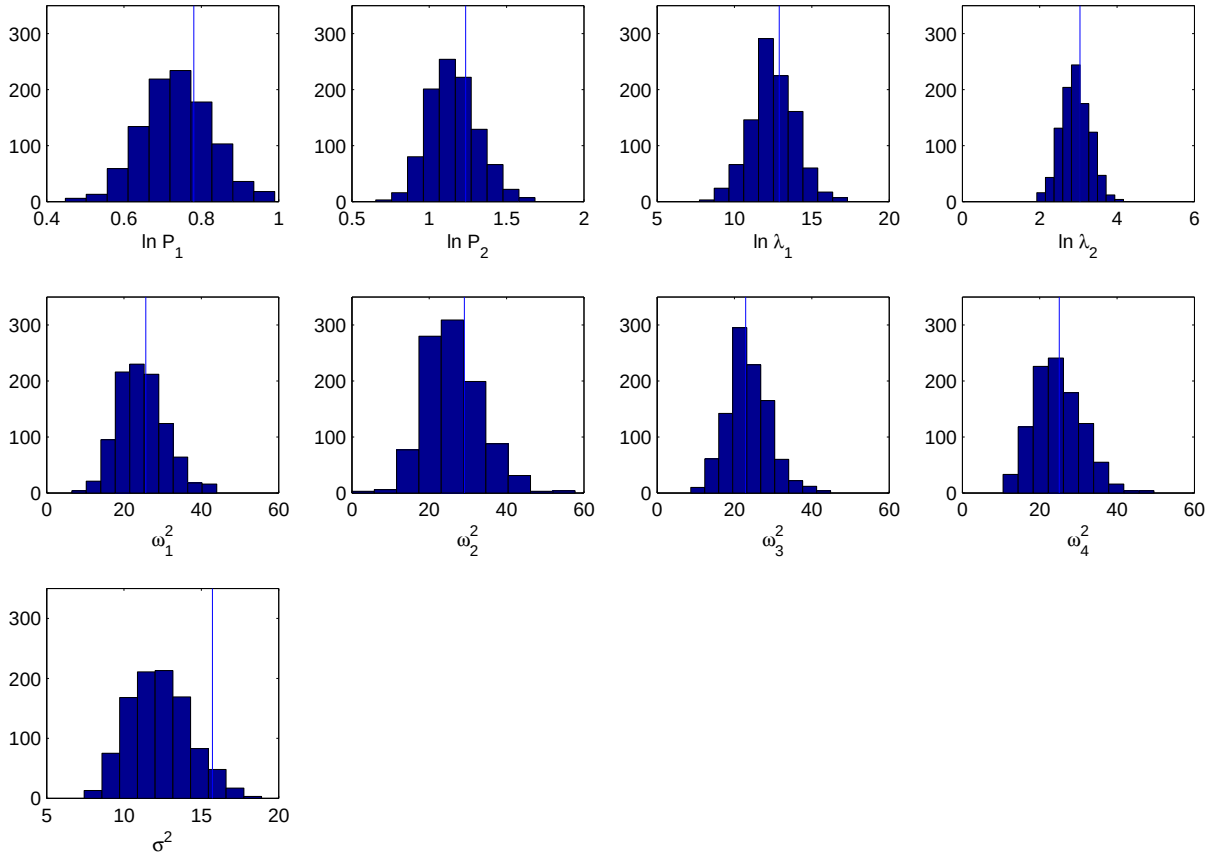


Figure 2. Histograms of the 1000 relative standard errors estimated by SAEM for datasets with N=40 subjects. The full line represents an estimate of the 'true' standard error estimated on the 1000 replications.

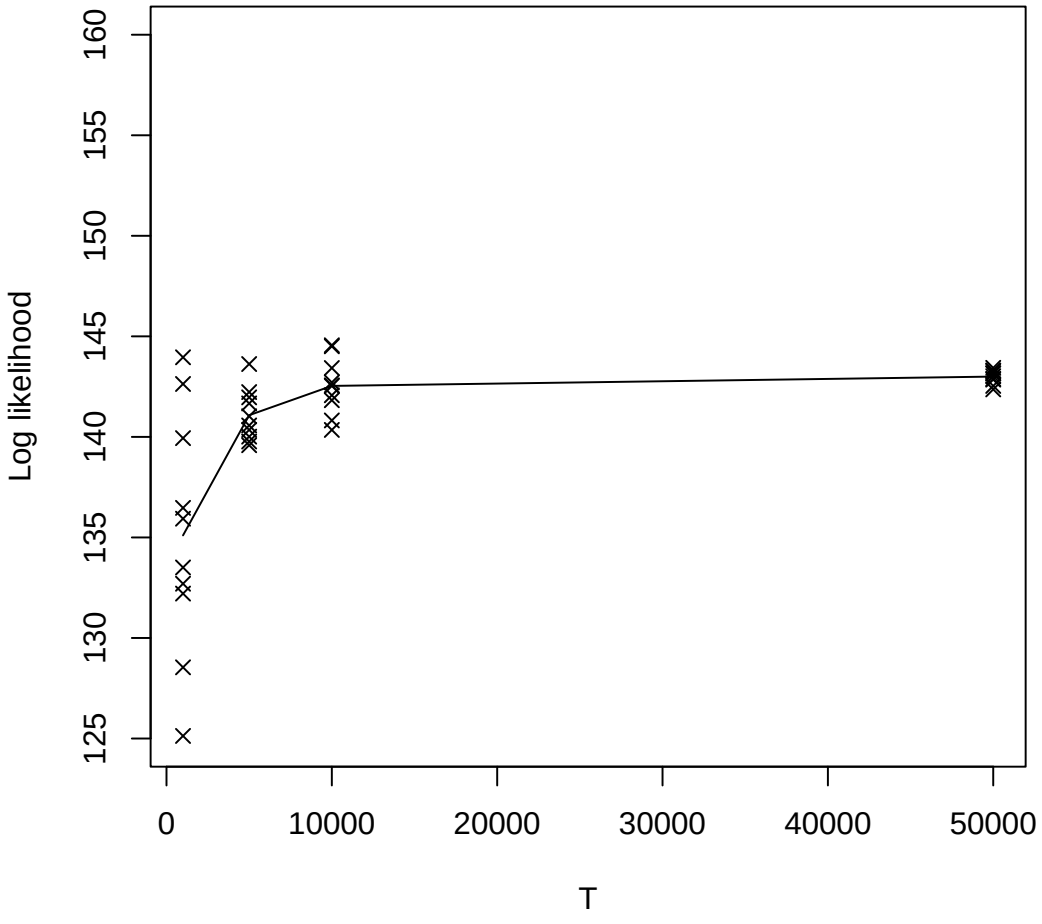


Figure 3. Log-likelihood estimates as a function of the sample size T used in the importance sampling procedure with 10 replications for each T, for one dataset with $N = 200$ subjects.

Table I. Relative bias (%) and relative root mean square error (RMSE) (%) of the estimated parameters evaluated by the SAEM algorithm from 1000 simulated trials with $N = 40$ and $N = 200$ subjects.

Parameters	Bias (%)		RMSE (%)	
	$N = 40$	$N = 200$	$N = 40$	$N = 200$
ln P_1	0.004	-0.004	0.781	0.348
ln P_2	0.011	-0.004	1.229	0.552
ln λ_1	-0.006	-0.247	12.923	5.753
ln λ_2	0.010	0.042	3.032	1.362
ω_1^2	-2.214	-0.418	25.767	10.884
ω_2^2	-3.395	-1.184	29.239	12.194
ω_3^2	-1.761	-0.182	22.874	10.587
ω_4^2	-1.207	0.206	25.156	11.643
σ^2	0.120	0.157	15.816	6.898

Table II. Evaluation on 1000 simulated datasets with 20 or 100 subjects per group of the type I errors of the Wald test and LRT for a treatment effect on the first decay rate.

	number of subjects per group	
	20	100
Wald test	4.0%	4.5%
LRT	5.8%	5.6%

Table III. Power of the Wald test for a treatment effect on the first decay rate according to the alternative hypothesis and the sampling design.

number of subjects per group	Alternative hypothesis H_1	
	$\beta_1 = 0.262$	$\beta_1 = 0.405$
20	32%	63%
40	55%	90%
100	92%	99%

Résultats complémentaires

Le calcul des erreurs standards attendues pour un modèle non-linéaire à effets mixtes a déjà été proposé par Retout et al [65]. Cette méthode, implémentée dans la fonction PFIM du logiciel R/Splus, est fondée sur une linéarisation de la fonction de régression f du modèle, de façon similaire à l'algorithme FO.

Nous avons comparé les erreurs standards attendues obtenues par la fonction PFIM avec celles obtenues par notre méthode fondée sur un calcul exact (sans linéarisation) de la matrice de Fisher sur l'exemple de décroissance de charge virale décrit dans l'article précédent. Nous avons également comparé ces estimateurs aux écarts-types observés sur 100 jeux de données simulés avec le même plan d'expérience et analysés avec l'algorithme SAEM. Les résultats, présentés dans le tableau 4.1, montrent la précision des SE prédites par l'approche SAEM par rapport aux déviations standards observées. Sur cet exemple simple (fonction bi-exponentielle),

Paramètre	SE prédites (%)		SD observés (%)
	Pfim	SAEM	
$\ln P_1$	0.34	0.34	0.35
$\ln P_2$	0.52	0.57	0.59
$\ln \lambda_1$	7.90	7.97	8.00
β *	0.079	0.078	0.086
$\ln\ \lambda_2$	1.30	1.30	1.53
ω_1^2	10.90	10.81	10.66
ω_2^2	11.33	12.66	12.33
ω_3^2	10.66	10.66	9.66
ω_4^2	10.42	10.66	11.33
σ^2	5.44	5.68	6.86

* SE et SD bruts pour ce paramètre

Table 4.1: Comparaison des erreurs standards (SE) prédites (%) par la fonction `pfim` ou l'algorithme SAEM pour le modèle bi-exponentiel avec 200 sujets avec les écarts-type (SD) observés (%) sur 100 jeux de données analysé par SAEM avec le même design.

la fonction PFIM et la méthode reposant sur l'algorithme SAEM fournissent des SE similaires. La fonction PFIM étant construite sur l'hypothèse simplificatrice d'une matrice de Fisher bloc-diagonale, la comparaison des matrices complètes de Fisher obtenues par ces deux méthodes n'est pas possible. Une comparaison similaire sur un exemple avec une fonction de régression plus complexe et des variabilités plus grandes devrait permettre de montrer l'avantage d'utiliser une méthode de calcul exact de la matrice de Fisher par rapport à une méthode fondée sur une linéarisation de la fonction de régression.

Chapitre 5

Prise en compte des données censurées

Dans le chapitre 4, nous avons illustré les capacités de l'algorithme SAEM, en termes de précision des estimateurs et des tests de comparaison, à analyser des données longitudinales de décroissance de la charge virale dans l'infection par le VIH. Cependant, les dispositifs expérimentaux actuels ne permettant pas de mesurer la charge virale en dessous d'un certain seuil (appelé LOQ pour *limit of quantification*) avec une précision suffisante, la valeur exacte de la charge virale sous cette limite n'est pas disponible pour l'analyse statistique. Différentes approches ont été proposées pour traiter ce problème de censure dans le cadre d'une analyse par modèle linéaire mixte, mais à ce jour, les seules méthodes proposées pour les modèles non-linéaires mixtes ont des propriétés statistiques insatisfaisantes et/ou ont des problèmes de convergence.

Nous avons proposé dans cet article une méthode d'estimation adaptée à ce problème et fondée sur l'algorithme SAEM. Un algorithme de Gibbs hybride comprenant la simulation des données censurées à partir d'une distribution gaussienne tronquée à droite est combiné à SAEM. Une estimation de l'espérance des données censurées conditionnellement aux données observées est également proposée. Nous avons comparé par simulation cette méthode à deux approches classiquement utilisées dans ce domaine : l'omission complète des données censurées, ou l'imputation pour la première valeur censurée de la valeur $LOQ/2$ et l'omission des valeurs suivantes. Nous avons montré la supériorité en termes de biais et de précision de notre algorithme sur ces deux méthodes dans l'article 2.

Nous avons également comparé ces résultats avec ceux obtenus avec la méthode naïve consistant à imputer pour chaque valeur censurée la valeur $LOQ/2$. Les biais des estimateurs obtenus par cette approche sont encore plus grands que ceux obtenus par les deux méthodes « classiques »présentées dans l'article. Dans un souci de clarté, nous n'avons pas présenté ces derniers résultats dans l'article et nous nous sommes

limités à la comparaison des résultats obtenus par l'extension de l'algorithme SAEM aux deux premières approches classiques d'imputation.

Lorsque des tests d'hypothèses sont réalisés au cours d'une analyse statistique de ce type, une imputation multiple est recommandée, c'est à dire l'utilisation de plusieurs chaînes de Markov en parallèle, afin d'affiner l'estimation des valeurs imputées. Nous avons montré dans cet article qu'en effet, en utilisant trois chaînes de Markov parallèles dans l'algorithme SAEM, les tests de Wald et du rapport de vraisemblance conservent leurs propriétés statistiques.

Nous avons utilisé cette méthode pour une nouvelle analyse des données de l'essai TRIANON-ANRS 81. Dans cette étude, l'efficacité de deux traitements (lamivudine, d4T et indinavir d'une part et nevirapine, d4T et indinavir d'autre part) était comparée chez 144 patients infectés par le VIH. L'analyse de ces données par notre méthode a permis de montrer une différence significative d'efficacité entre les deux traitements concordante avec les résultats originaux [66], alors que l'utilisation de la méthode par modèle mixte actuellement recommandée pour l'analyse de telles données longitudinales (l'imputation pour la première valeur censurée de la valeur $LOQ/2$ et l'omission des valeurs suivantes) ne permet pas de trouver cette différence.

Ce travail fait l'objet d'un article conditionnellement accepté par *Computational Statistics and Data Analysis*.

Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model: application to HIV dynamics model

Adeline Samson [a,b,*] Marc Lavielle [c] France Mentré [a,b]

[a] *INSERM, U738, Paris, F-75018 France*
[b] *Universit Paris 7, Facult Xavier Bichat, Paris, F-75018 France*
[c] *Universit Paris-Sud, Bat. 425, Orsay, F-91000 France*

Abstract

The reduction of viral load is frequently used as a primary endpoint in HIV clinical trials. Non-linear mixed-effects models are thus proposed to model this decrease of the viral load after initiation of treatment and to evaluate the intra- and inter-patient variability. However, left censoring due to quantification limits in the viral load measurement is an additional challenge in the analysis of longitudinal HIV data. An extension of the Stochastic Approximation Expectation-Maximization (SAEM) algorithm is proposed to estimate parameters of these models. This algorithm includes the simulation of the left-censored data in a right-truncated Gaussian distribution. Simulation results show that the proposed estimates are less biased than the usual naive methods of handling such data: omission of all censored data points, or imputation of half the quantification limit to the first point below the limit and omission of the following points. The viral load measurements obtained in the TRIANON-ANRS81 clinical trial are analyzed with this method and a significant difference is found between the two treatment groups of this trial.

Key words: Accept-reject algorithm, HIV dynamics, Left-censored data, MCMC algorithm, Non-linear mixed-effects models, SAEM algorithm

* Adeline Samson, Biostatistics Department, Bichat Hospital, 46 rue Huchard, 75018 Paris, France, tel: +33 140 25 62 57, fax: +33 140 25 67 73
Email address: adeline@e-samson.org (Adeline Samson).

Preprint submitted to Elsevier Science *29 March 2006*

1 Introduction

HIV viral load is a widespread marker of the evolution of HIV infected patients (1); the reduction in HIV viral load is frequently used as the primary endpoint in clinical trials to evaluate the efficacy of anti-viral treatments (see for example 2; 3; 4; 5; 6; 7; 8). Non-linear mixed-effects models (NLMEM) can be used in these longitudinal studies to exploit the richness of the dynamics seized by repeated measurements and to account for inter- and intra-patient variability in viral load measurements. In addition, understanding the mechanism of the large inter-patient variability may help in making appropriate clinical decisions and providing individualized treatment. Unfortunately, all available assays of viral load measurements have a low limit of quantification (LOQ), generally between 20 and 400 copies/ml. Besides, the proportion of subjects with a viral load below LOQ has increased with the introduction of highly active antiretroviral treatments. Working with such left-censored data complicates the study of longitudinal viral load data. This issue is common in other longitudinal studies with LOQ, such as pharmacokinetics or pharmacodynamics, which also widely use NLMEM.

This paper aims to develop a reliable inference based on maximum likelihood (ML) theory for HIV dynamics models with left-censored viral load and NLMEM. It is indeed important to obtain reliable estimates of the viral dynamic parameters, that can be used to evaluate antiviral therapies through comparison of treatment groups.

To address the estimation problem in longitudinal data analysis containing censored values, naive procedures such as omitting the censored data or imputing a fixed value (e.g., the quantification limit or half the limit) are combined with usual estimation methods of mixed models (see Beal (9) for a comparison of classical procedures in NLMEM). However, the statistical properties of such procedures are unclear. More inventive approaches propose multiple imputations of the censored values, by substituting a reasonable guess for each missing value. For example, in linear mixed models, Hughes (10) proposes a Monte-Carlo version of the Expectation Maximization (EM) algorithm (11), taking into account the censored values as missing data. Hughes (10) shows that his approach significantly reduces the bias associated with naive imputation procedures. Jacqmin-Gadda et al. (5) propose a direct maximization of the likelihood using an iterative process for linear mixed models as well, including an autoregressive error model. They combine two optimization algorithms, the Simplex and the Marquardt algorithms.

For non-linear mixed models, the problem is more complex, the estimation of such model parameters being difficult even without censored observations. Indeed, because of the non linearity of the regression function in the random

effects, the likelihood of NLMEM cannot be expressed in a closed form. Consequently, several authors propose some widely used likelihood approximation methods, such as linearization algorithms, which are implemented in the NONMEM software and in the nlme function of Splus and R software (12; 13); or Laplacian or Gaussian quadrature algorithms, which are implemented in the NLMIXED Macro of SAS (14). Wu and Wu propose a multiple imputation method for missing covariates in NLMEM based on a linearization algorithm (15). However, none of these algorithms based on likelihood approximation can be considered as fully established theoretically. A different point of view can be taken, the individual parameters and the censored values being considered as non-observed data. The EM algorithm is then the most adapted tool to estimate incomplete data models. Because of the nonlinearity of the model, stochastic versions of the EM algorithm are proposed. Wu (16; 17) introduces MCEM algorithms, with a Monte-Carlo approximation of the Expectation step, adapted to both NLMEM and the censoring problem of observations and covariates. This Monte-Carlo implementation is based on samples independently and identically distributed from the conditional density, requiring Markov Chain Monte-Carlo (MCMC) procedures. The replication choice of the Monte-Carlo sample is a central issue to guarantee convergence and this remains an open problem. Wu proposes an "exact" MCEM (17) but emphasizes that this MCEM algorithm is very slow to converge. Indeed simulations of these large samples at each iteration are time consuming. To address this computational problem, Wu also proposes approximate MCEM (16; 17) using a linearization of the model leading to an approximate ML method.

As an alternative to address both the point-wise convergence and the computational problem, stochastic approximation versions of EM (SAEM) are proposed for NLMEM with no censored-values (18; 19). This algorithm requires a simulation of only one realization of the missing data at each iteration, avoiding the computational difficulty of independent sample simulation occurring in the MCEM and shortening the time to simulate. In addition, point-wise almost sure convergence of the estimate sequence to a local maximum of the likelihood is proved by Delyon et al. (18) under conditions satisfied by models from the exponential family. Girard and Mentré (20) propose a comparison of these estimation methods in NLMEM using a blind analysis, showing the accuracy of the SAEM algorithm in comparison with other methods. Especially, the computational convergence of the SAEM algorithm is clearly faster than those of the MCEM algorithm. However, this current SAEM algorithm is only appropriate for NLMEM without censored-values.

The first objective of the present paper is thus to extend the SAEM algorithm to handle left-censored data in NLMEM as an exact ML estimation method. We include in the extended SAEM algorithm the simulation of the left-censored data with a right-truncated Gaussian distribution. We prove the convergence of this extended SAEM algorithm under general conditions. The

second objective of this paper is to illustrate this algorithm with a simulation study in the HIV dynamics context. Furthermore, we compare the extended SAEM algorithm with more classical approaches to handle left-censored data such as omission or imputation of the censored data, on the same simulation study.

After describing the model and the notations (Section 2), Section 3 describes the extended SAEM algorithm. Section 4 reports the simulation study and its results. We simulate datasets using the bi-exponential model for HIV dynamics proposed by Ding and Wu (6), and evaluate the statistical properties of the extended SAEM parameter estimates and the classical approaches. Particularly, we evaluate 2 comparison group tests, the Wald test and the likelihood ratio test, provided by the SAEM algorithm. We then apply the extended SAEM algorithm to the TRIANON-ANRS 81 clinical trial of HIV treatment in Section 5. The aim of this new analysis of the TRIANON data is to show the ability of NLMEM to describe the evolution of the viral load and to test a treatment's effect between the 2 treatment groups, in the presence of left-censored observations. Section 6 concludes the article with some discussion.

2 Models and notations

Let us define $y_i = (y_{i1}, \ldots, y_{in_i})^t$ where y_{ij} is the response value for individual i at time t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, and let $y = (y_1, \ldots, y_N)$. Let us define an NLMEM as follows

$$
\begin{aligned}
&y_{ij} = f(\phi_i, t_{ij}) + g(\phi_i, t_{ij})\, \varepsilon_{ij}, \\
&\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}) \quad \text{and} \\
&\phi_i = X_i \mu + b_i, \text{with} \;\; b_i \sim \mathcal{N}(0, \Omega),
\end{aligned}
$$

where $f(\cdot)$ and/or $g(\cdot)$ are non-linear functions of ϕ_i, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^t$ represents the residual error, ϕ_i is a p-vector of individual parameters, μ is the $k \times p$-matrix of fixed effects, X_i is the k-vector of known covariates, b_i is a p-vector of random effects independent of ε_i, σ^2 is the residual variance, I_{n_i} the identity matrix of size n_i and Ω quantifies the covariance of the inter-individual random effects.

Because of assay limitation, when data y_{ij} are inferior to the limit of quantification (LOQ), we do not observe y_{ij} but only the censored value LOQ. These data are usually named left-censored data. Let denote $I_{obs} = \{(i,j) | y_{ij} \geq LOQ\}$ and $I_{cens} = \{(i,j) | y_{ij} \leq LOQ\}$ the index sets of the uncensored and censored observations respectively. For $(i,j) \in I_{cens}$, let $y_{ij}^{cens} = y_{ij}$ denote the unknown value of the censored observation j of subject i. Let denote y_i^{cens} the

vector of censored observations of subject i. Finally, we observe

$$y_{ij}^{obs} = \begin{cases} y_{ij} & \text{if} \quad (i,j) \in I_{obs}, \\ LOQ & \text{if} \quad (i,j) \in I_{cens}. \end{cases}$$

We denote $y_i^{obs} = (y_{i1}^{obs}, \ldots, y_{in_i}^{obs})$ as the observations of subject i and $y^{obs} = (y_1^{obs}, \ldots, y_N^{obs})$ the total observations dataset.

The maximum likelihood estimation is based on the log-likelihood function $L(y^{obs}\,;\,\theta)$ of the response y^{obs} with $\theta = (\mu, \Omega, \sigma^2)$ the vector of all the parameters of the model

$$L(y^{obs}\,;\,\theta) = \log\left(\prod_{i=1}^{N} \int p(y_i^{obs}, y_i^{cens}, \phi_i; \theta)\, d\phi_i\, dy_i^{cens}\right), \tag{1}$$

where $p(y_i^{obs}, y_i^{cens}, \phi_i;\, \theta\,)$ is the likelihood of the complete data $(y_i^{obs}, y_i^{cens}, \phi_i)$ of the i-th subject. Because the random effects ϕ_i and the censored observations y_i^{cens} are unobservable and the regression functions are non-linear, the foregoing integral has no closed form. The complete likelihood of the i-th subject is equal to:

$$p(y_i^{obs}, y_i^{cens}, \phi_i; \theta) = \prod_{(i,j)\in I_{obs}} p(y_{ij}^{obs}|\phi_i;\theta)p(\phi_i;\theta) \prod_{(i,j)\in I_{cens}} p(y_{ij}^{cens}|\phi_i;\theta)p(\phi_i;\theta),$$

with

$$p(y_{ij}^{obs}|\phi_i;\theta) = \pi(y_{ij}^{obs}; f(\phi_i, t_{ij}), \sigma^2 g^2(\phi_i, t_{ij}))\ \mathbb{1}_{y_{ij}\geq LOQ}, \quad \text{if} \quad (i,j) \in I_{obs} \quad \text{and}$$
$$p(y_{ij}^{cens}|\phi_i;\theta) = \pi(y_{ij}^{cens}; f(\phi_i, t_{ij}), \sigma^2 g^2(\phi_i, t_{ij}))\ \mathbb{1}_{y_{ij}\leq LOQ}, \text{ if} \quad (i,j) \in I_{cens},$$

where $\pi(x; m, v)$ is the probability density function of the Gaussian distribution with mean m and variance v, evaluated at x.

3 Estimation algorithm

3.1 The SAEM algorithm

The EM algorithm introduced by Dempster et al. (11) is a classical approach to estimate parameters of models with non-observed or incomplete data. Let us briefly cover the EM principle. Let z be the vector of non-observed data. The complete data of the model is (y, z). The EM algorithm maximizes the $Q(\theta|\theta') = E(L_c(y, z; \theta)|y; \theta')$ function in 2 steps, where $L_c(y, z; \theta)$ is the log-likelihood of the complete data. At the m-th iteration, the E step is the evaluation of $Q_m(\theta) = Q(\theta \,|\, \widehat{\theta}_{m-1})$, whereas the M step updates $\widehat{\theta}_{m-1}$ by maximizing

$Q_m(\theta)$. For cases in which the E step has no analytic form, Delyon et al. (18) introduce a stochastic version SAEM of the EM algorithm which evaluates the integral $Q_m(\theta)$ by a stochastic approximation procedure. The authors prove the convergence of this algorithm under general conditions if $L_c(y,z;\theta)$ belongs to the regular curved exponential family

$$L_c(y,z;\theta) = -\Lambda(\theta) + \langle S(y,z), \Phi(\theta)\rangle,$$

where $\langle .,. \rangle$ is the scalar product, Λ and Φ are 2 functions of θ and $S(y,z)$ is the minimal sufficient statistic of the complete model. The E step is then divided into a simulation step (S step) of the missing data $z^{(m)}$ under the conditional distribution $p(z|y;\widehat{\theta}_{m-1})$ and a stochastic approximation step (SA step) using $(\gamma_m)_{m\geq 0}$ a sequence of positive numbers decreasing to 0. This SA step approximates $E\left[S(y,z)|\widehat{\theta}_{m-1}\right]$ at each iteration by the value s_m defined recursively as follows

$$s_m = s_{m-1} + \gamma_m(S(y,z^{(m)}) - s_{m-1}).$$

The M step is thus the update of the estimates $\widehat{\theta}_{m-1}$

$$\widehat{\theta}_m = \arg\max_{\theta\in\Theta}\left(-\Lambda(\theta) + \langle s_m, \Phi(\theta)\rangle\right).$$

Let us detail the sufficient statistics needed for evaluation at the SA step of the extended SAEM algorithm for the non-linear mixed models previously presented. The sufficient statistics are $S^{(1)} = \sum_{i=1}^N \phi_i$, $S^{(2)} = \sum_{i=1}^N \phi_i^2$ and $S^{(3)} = \sum_{i,j}(y_{ij} - f(\phi_i, t_{ij}))^2$, where $y_{ij} = y_{ij}^{obs}$ if $(i,j) \in I_{obs}$ and $y_{ij} = y_{ij}^{cens}$ if $(i,j) \in I_{cens}$. Therefore, at the m-th iteration of SAEM the M-step reduces to

$$\begin{aligned}
\widehat{\mu}_m &= \frac{1}{N} s_m^{(1)}, \\
\widehat{\omega^2}_m &= \frac{1}{N} s_m^{(2)} - (s_m^{(1)})^2 \quad \text{and} \\
\widehat{\sigma^2}_m &= \frac{1}{NJ} s_m^{(3)}.
\end{aligned}$$

In cases in which the simulation of the non-observed vector z cannot be directly performed, Kuhn and Lavielle (19) propose to combine this algorithm with a Markov Chain Monte-Carlo (MCMC) procedure. The convergence of this SAEM algorithm is ensured under general conditions; the 2 main conditions are presented below (see Kuhn and Lavielle 21, for technical conditions)

(SAEM 1) For any $\theta \in \Theta$, the Gibbs algorithm generates a uniformly ergodic chain which invariant probability is $p(z|y;\theta)$.
(SAEM 2) For all m in the integer set $\mathbb{N}^*$, $\gamma_m \in [0,1], \sum_{m=1}^\infty \gamma_m = \infty$ and $\sum_{m=1}^\infty \gamma_m^2 < \infty$.

For NLMEM with left-censored data, the nonobserved vector is $z = (\phi, y^{cens})$, with $\phi = (\phi_1, \ldots, \phi_N)$ being the individual parameters vector and $y^{cens} = (y_1^{cens}, \ldots, y_N^{cens})$ the left-censored data vector. The S step of the SAEM algorithm is the simulation of the missing data (ϕ, y^{cens}) under the posterior distribution $p(\phi, y^{cens}|y^{obs}; \theta)$. This step can be performed by use of a Gibbs sampling algorithm. At the m-th iteration of the SAEM algorithm, the Gibbs algorithm is thus divided into 2 steps

(1) Simulation of $\phi^{(m)}$ by use of a Metropolis-Hastings (M-H) algorithm constructing a Markov Chain $\phi^{(m)}$ with $p(\cdot|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})$ as the unique stationary distribution,
(2) Simulation of $y^{cens(m)}$ with the posterior right-truncated Gaussian distribution $p(\cdot|y^{obs}, \phi^{(m)}; \widehat{\theta}_{m-1})$.

Consequently, under assumptions (SAEM1) and (SAEM2) and general additional conditions, by applying the convergence theorem of Kuhn and Lavielle (19) the estimate sequence $(\widehat{\theta}_m)_{m\geq 0}$ produced by the extended SAEM algorithm converges towards a (local) maximum of the likelihood $L(y^{obs}; .)$.

Samson et al. (22) propose to estimate the likelihood function with use of an importance sampling procedure and detail its implementation. They estimate the Fisher information matrix combining a stochastic approximation approach and the Louis' missing information principle (23): the Hessian of the log-likelihood of the observed data can be obtained almost directly from the simulated missing data (see Kuhn and Lavielle 19, for more implementation details). We adapt their estimates of the likelihood and the Fisher information matrix to the extended SAEM algorithm, to implement the 2 comparison group tests, the Wald test and the likelihood ratio test.

3.2 Computational aspects

The convergence of the SAEM algorithm is ensured under the 2 assumptions (SAEM 1) and (SAEM 2), which require careful choices of the implementation of the Gibbs algorithm and the stochastic approximation step size respectively.

3.2.1 Gibbs algorithm

The convergence of the Gibbs algorithm depends on the M-H algorithm generating ϕ and the simulation method generating y^{cens}.

At the m-th iteration of the SAEM algorithm, the M-H algorithm proceeds as follows: a candidate ϕ^c is simulated with a proposal distribution $q_{\widehat{\theta}_{m-1}}$. The

candidate is accepted (i.e. $\phi^{(m)} = \phi^c$), with the acceptation probability ρ

$$\rho = \min\left(\frac{p(\phi^c|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})}{p(\phi^{(m-1)}|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})} \frac{q_{\widehat{\theta}_{m-1}}(\phi^c|\phi^{(m-1)})}{q_{\widehat{\theta}_{m-1}}(\phi^{(m-1)}|\phi^c)}, 1\right),$$

and the candidate is rejected (i.e. $\phi^{(m)} = \phi^{(m-1)}$), with probability $1 - \rho$.

We propose the 3 following proposal distributions $q_{\widehat{\theta}_{m-1}}$ for the M-H procedure

(1) $q^{(1)}_{\widehat{\theta}_{m-1}}$ is the prior distribution of ϕ, that is, the Gaussian distribution $\mathcal{N}(\widehat{\mu}_{m-1}, \widehat{\Omega}_{m-1})$,
(2) $q^{(2)}_{\widehat{\theta}_{m-1}}$ is the multidimensional random walk $\mathcal{N}(\phi^{(m-1)}, \lambda\widehat{\Omega}_{m-1})$, where λ is a scaling parameter chosen to ensure a sufficient acceptation rate,
(3) $q^{(3)}_{\widehat{\theta}_{m-1}}$ is a succession of p unidimensional Gaussian random walks: each component of ϕ is successively updated.

With the proposal distributions detailed below, this M-H algorithm converges and generates a uniformly ergodic chain with $p(\phi|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})$ as the stationary distribution, thus fulfilling the assumption (SAEM 1).

Then, an efficient simulation method has to be implemented to generate $y_{ij}^{cens(m)}$ for all $(i, j) \in I_{cens}$ with the right-truncated Gaussian distribution with mean $f(\phi_i^{(m)}, t_{ij})$, variance equal to $\widehat{\sigma^2}_{m-1}\, g^2(\phi_i^{(m)}, t_{ij})$ and truncated at the right by the value LOQ. We implement the accept-reject algorithm proposed by Robert (24) because of its simplicity and because it slightly improves upon previous algorithms developed by Gelfand and Smith (25). This algorithm is composed of the following steps

(1) compute $\alpha = \frac{C+\sqrt{C^2+4}}{2}$,
(2) simulate x with the translated exponential distribution $\mathcal{E}(\alpha, C)$ with density $p(x|\alpha, C) = \alpha \exp(-\alpha(x - C))\mathbb{1}_{x \geq C}$,
(3) compute $\rho(x) = \exp(-(x - \alpha)^2/2)$,
(4) simulate u with $\mathcal{U}_{[0,1]}$,
(5) if $u \leq \rho(x)$, then keep x and compute $y_{ij}^{cens(m)} = f(\phi_i^{(m)}, t_{ij}) - x\widehat{\sigma}_{m-1} g(\phi_i^{(m)}, t_{ij})$, else return to step (2).

The simulation of x with the translated exponential distribution $\mathcal{E}(\alpha, C)$ in step (2) is performed by simulating u with a uniform distribution $\mathcal{U}[0, 1]$ on the unit interval and then by computing $x = -\frac{1}{\alpha}\ln(1 - u) + C$.

With the proposal distributions detailed below, this Gibbs algorithm converges and generates a uniformly ergodic chain with $p(\phi, y^{cens}|y^{obs}; \theta)$ as the stationary distribution, thus fulfilling the assumption (SAEM 1).

3.2.2 Stochastic approximation step size sequence

The sequence $(\gamma_m)_{m\geq 0}$ has to fulfill the assumption (SAEM 2). We recommend the use of $\gamma_m = 1$ during the first M_1 iterations $1 \leq m \leq M_1$, and $\gamma_m = (m - M_1)^{-1}$ during the last M_2 iterations. Indeed, the initial guess θ_0 may be far from the maximum likelihood value and the first iterations with $\gamma_m = 1$ allow for converging to a neighborhood of the maximum likelihood estimate. Furthermore, the inclusion of a hybrid Gibbs procedure (instead of a Metropolis-Hastings procedure in the classic SAEM algorithm) slows up the convergence of the extended SAEM algorithm. The convergence is monitored from graphical criterion. The choice of M_1 and M_2 values and $(\gamma_m)_{m\geq 0}$ are adapted according to the graphical convergence of all the parameter estimates.

4 Simulation study

4.1 Simulation settings

The first objective of this simulation study is to illustrate the main statistical properties of the extended SAEM algorithm in the context of HIV viral dynamics (bias, root mean square errors, group comparison tests). The second objective is to compare the extended SAEM algorithm to some of the classical approaches proposed to take into account a censoring process.

We use the bi-exponential model for initial HIV dynamics proposed by Ding and Wu (6) to simulate the datasets

$$f(\phi_i, t_{ij}) = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{ij}} + P_{2i}e^{-\lambda_{2i}t_{ij}}).$$

This function is a simplified analytical solution of a differential system describing HIV viral load decrease during anti-retroviral treatment proposed by Perelson et al. (1). It has p=4 individual parameters: P_{1i}, P_{2i} are the baseline values and λ_{1i}, λ_{2i} represent 2-phase viral decay rates. These parameters are positive and distributed according to a log-normal distribution. Thus, ϕ_i and μ take the following values: $\phi_i = (\ln P_{1i}, \ln P_{2i}, \ln \lambda_{1i}, \ln \lambda_{2i})$ and $\mu = (\ln P_1, \ln P_2, \ln \lambda_1, \ln \lambda_2)$. We assume identical sampling times for all subjects: for all i in $1, \ldots, N$, $t_{ij} = t_j$ for $j = 1, \ldots, n$. Additive Gaussian random effects are assumed for each parameter with a diagonal covariance matrix Ω. Let $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2)$ denote the vector of the variances of the random effects. Additive Gaussian error is assumed with a constant variance σ^2 (i.e. $g(\phi_i, t_j) = 1$ for all i, j).

For the fixed effects, the values are those proposed by Ding and Wu (6): $\ln P_1 = 12$, $\ln P_2 = 8$, $\ln \lambda_1 = \ln(0.5)$, $\ln \lambda_2 = \ln(0.05)$. The inter-subject variability is

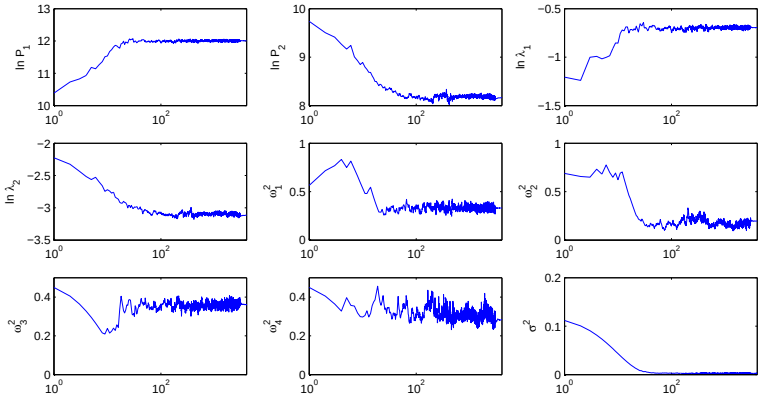


Fig. 1. Convergence of the SAEM parameter estimates for one simulated dataset with $N = 40$ subjects (semi-log scale)

identical for the 4 parameters: $\omega_1^2 = \omega_2^2 = \omega_3^2 = \omega_4^2 = 0.3$ corresponding to a variation coefficient of 55%, which is a realistic inter-subject variability in the context of HIV dynamics. We chose a variance $\sigma = 0.065$, which corresponds to a constant variation coefficient of 15% for the viral load. With the Matlab software, we generate N=40 total number of subjects and, with n=6 blood samples per patient, taken on days 1, 3, 7, 14, 28 and 56. We consider the same limit of quantification as Ding and Wu: $LOQ = \log_{10}(400)$.

The convergence of the SAEM algorithm on a simulated dataset is illustrated in Figure 1. The initial estimates are arbitrarily chosen for all the parameters. During the first $M_1 = 3000$ iterations, the estimates converge to a neighborhood of the maximum likelihood. Then, smaller step sizes during $M_2 = 1000$ additional iterations ensure the almost sure convergence of the algorithm to the maximum likelihood estimate. We implement the extended SAEM algorithm in a Matlab function. It takes about 120 s for the extended SAEM algorithm to converge with 4000 iterations on a conventional Intel Pentium IV 2.8 GHz workstation.

The conditional expectation $E(y^{cens}|y^{obs})$ of the censored values can be evaluated from the posterior mean of the y^{cens} simulated during the last iterations of the extended SAEM algorithm. Figure 2 illustrates this evaluation on a simulated dataset: $E(y^{cens}|y^{obs})$ evaluated by SAEM is plotted as a function of the true simulated values y that are below the LOQ for this simulated dataset. The extended SAEM algorithm provides satisfactory expectation of these censored values.

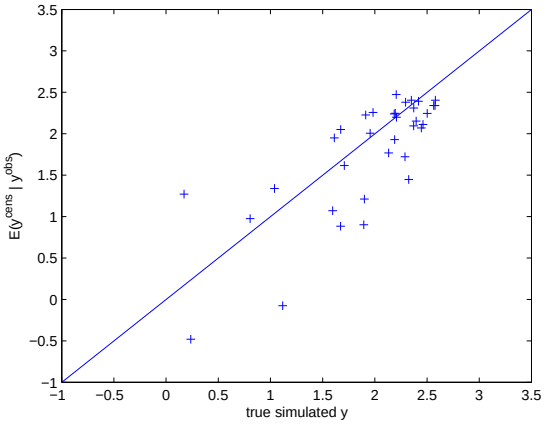


Fig. 2. Expectation of the censored values $E(y^{cens}|y^{obs})$ evaluated by the extended SAEM algorithm as a function of the true simulated values y that are below the LOQ (2.6) on a simulated dataset.

4.2 Evaluation of estimates

Our aim is to evaluate and compare the estimates produced by the extended SAEM algorithm with those produced by 2 estimation approaches recommended in the presence of left-censored data. We fit the simulation model and compute the relative bias and relative root mean square error (RMSE) for each component of θ from 1000 replications of the trial described below.

We first assume that no censoring is present in the viral load. We estimate the datasets using the classical SAEM algorithm; this bias and RMSE are considered the benchmark for the comparison of the 3 methods on the censored datasets described below.

We then censor the simulated datasets by censoring observations that are below the LOQ. The censoring represents, on average, 0% of the observations at days 1 and 3, 0.07% at day 7, 2.81% at day 14, 26.96% at day 28 and 71.57% at day 56. First, we implement 2 classical approaches omitting or an imputing arbitrary value to the censored data. We name M_1 the naive approach, which omits all censored data. We then name M_2 the method recommended by several authors (9; 6; 26); for each patient, the first data below the LOQ is kept and imputed to LOQ/2, and then all the following censored data are omitted. We use the standard SAEM algorithm to fit the datasets for both the M_1 and M_2 methods. Second, we also apply the extended SAEM algorithm presented in Section 3; this gives us the maximum likelihood (ML) estimates of the parameter θ from the original dataset y^{obs}.

The relative bias and RMSE obtained under the simulation model on the uncensored datasets with the classical SAEM algorithm are presented in Table

Table 1
Relative bias (%) and relative root mean square error (RMSE) (%) of the estimated parameters evaluated from 1000 simulated trials on the uncensored datasets (all data) with the SAEM algorithm and the left-censored datasets with 2 classic methods (M_1 and M_2) and with the extended SAEM algorithm (ML).

Parameters	Bias (%)				RMSE (%)			
	all data	left censored data			all data	left censored data		
		M_1	M_2	**ML**		M_1	M_2	**ML**
ln P_1	0.01	0.03	0.32	**0.03**	0.77	0.78	0.93	**0.77**
ln P_2	0.01	2.64	10.71	**0.23**	1.29	3.22	10.88	**1.63**
ln λ_1	0.98	2.67	12.94	**0.57**	12.47	12.55	19.76	**12.36**
ln λ_2	0.04	10.46	22.88	**0.62**	3.09	11.45	23.36	**3.98**
ω_1^2	0.28	3.69	37.51	**4.26**	24.17	26.55	49.60	**26.30**
ω_2^2	2.20	12.67	24.81	**6.21**	26.65	37.15	58.31	**37.70**
ω_3^2	1.97	6.85	12.53	**1.67**	22.48	23.03	31.01	**23.05**
ω_4^2	0.88	47.13	98.331	**6.59**	25.66	55.98	113.53	**36.85**
σ^2	0.51	10.31	440.77	**0.63**	16.34	26.24	453.24	**19.34**

1 and referred as the "all data" estimates. These estimates have very small bias (<0.5% for the fixed effects, <5% for the variance parameters). The RMSE is really satisfactory for the fixed effects (<13%) and the variance parameters (<30%).

The relative bias and RMSE obtained on the censored datasets are presented in Table 1. Three of the fixed effects are estimated with bias by the M_2 method, especially ln λ_2 (23%). The M_1 method reduces the bias for all the fixed effects but ln λ_2 still has a larger bias (10.5%) than before the censor. The bias of the variance parameters is increased with both the M_1 and M_2 methods, especially for ω_4^2 and σ^2 (47% and 98% for ω_4^2 and 10% and 440% of bias for σ^2 respectively). In contrast with these 2 methods, the extended SAEM algorithm provides estimates of all the parameters with small bias. The M_1 method gives a satisfactory RMSE except for λ_2 (11%) and ω_4^2 (56%). The M_2 method increases all the RMSE, especially for ω_4^2 (113%) and σ^2 (453.2%). The mean bias of the σ M_2-estimates is 1.31, which corresponds to the value of $LOQ/2$. The RMSE is satisfactory with the extended SAEM algorithm.

Every dataset is almost censored at days 28 or 56 during the second decay phase of the viral load decrease. This finding explains that the parameter estimates corresponding to this second decay rate (ln λ_2 and its variance ω_4^2) are the most affected by the censoring process. However, even with 71% of

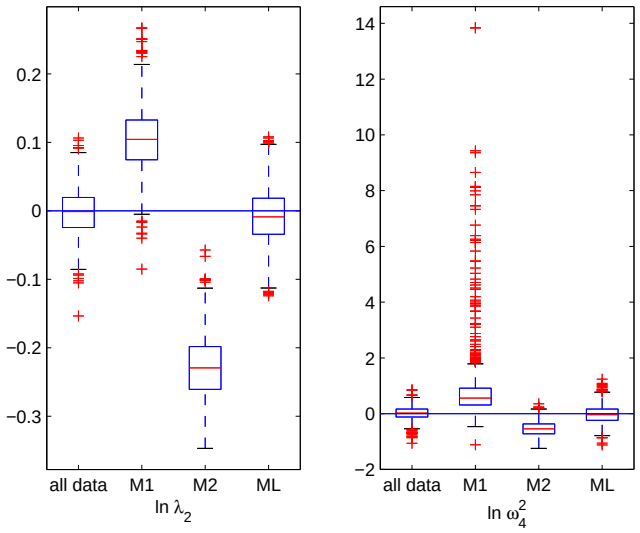


Fig. 3. Boxplot of the second decay rate parameter estimates (ln λ_2 and ln ω_4^2) for the 4 methods on the 1000 replications: the all-data method, the M_1 and M_2 methods, and ML, the extended SAEM algorithm.

censor at day 56, the bias and RMSE of the extended SAEM algorithm almost reach the uncensored dataset benchmark. This accuracy is also illustrated in Figure 3, which presents the distribution of the second decay rate parameters estimates (ln λ_2 and ln ω_4^2) for the 4 methods from the 1000 replications. This figure again points out the bad properties of the M_1 and M_2 methods, and reemphasizes that the extended SAEM algorithm reaches the exactness level of the estimation method applied to the uncensored datasets. The difference in the other parameters distributions between the 4 methods are similar.

The M_1 method provides estimates that are less biased than the M_2 method for all parameters. This finding can be explained by the design used for the simulation. The number of uncensored measurements is large enough to estimate quite accurately all the parameter by omitting all the censored data. Contrary to the M_1 method, which considers a partial dataset from the original dataset, the M_2 method is based on a modified partial dataset. Because the modification affects the data measured during the second decay, the parameter estimates of this second decay rate (ln λ_2 and its variance ω_4^2) are noticeably biased.

4.3 Application to group comparison

We consider that the subjects of each simulated trial belong to 2 different treatment groups of equal size (i.e. 20 subjects per group). We performed a Wald test and Likelihood ratio test (LRT) to test a difference between the

treatment groups on the viral load decrease, especially on the first viral decay rate, $\ln\lambda_1$, as proposed by Ding and Wu (6). We apply these tests using SAEM on the uncensored datasets and the extended SAEM algorithm on the censored datasets and evaluate their type I errors. We do not evaluate the type I errors obtained with the M_1 and M_2 methods on the censored datasets because the previous simulation study already illustrates their bad properties.

Let $G_i = 0$ denote a control treatment group subject and $G_i = 1$ an experiment treatment group subject. In this example, the vector of covariates X_i is (1, G_i). Let β denote the treatment effect parameter on $\ln\lambda_1$.

$$\begin{aligned} f(\phi_i, t_{ij}) &= \log_{10}(P_{1i}\exp(-\lambda_{1i}t_{ij}) + P_{2i}\exp(-\lambda_{2i}t_{ij})) \quad \text{and} \\ \ln\lambda_{1i} &= \ln\lambda_1 + \beta G_i. \end{aligned}$$

In this case, the matrix of fixed effects is

$$\mu = \begin{pmatrix} \ln P_1 & \ln P_2 & \ln\lambda_1 & \ln\lambda_2 \\ 0 & 0 & \beta & 0 \end{pmatrix}.$$

We test by LRT or Wald test the hypothesis that the 2 treatments are equal, H_0: $\{\beta = 0\}$, versus the alternative hypothesis H_1: $\{\beta \neq 0\}$. For the LRT, we evaluate the log-likelihoods L_0 and L_1 by importance sampling under H_0 and H_1, respectively. Because the likelihood function is differentiable for every θ and the H_0 is equivalent locally to a linear space, the LRT statistic is asymptotically chi-squared distributed. We thus compare the $2(L_1 - L_0)$ statistic with a 1 degree of freedom χ^2_1 distribution. The importance sampling procedure is implemented by simulating a sample of size 5000 of the individual parameters ϕ_i with the Gaussian approximation of the posterior distribution, using estimates of the individual posterior mean $E(\phi_i|y_i^{obs})$ and the posterior variance $\mathrm{Var}(\phi_i|y_i^{obs})$ evaluated by the empirical mean and variance of the ϕ_i simulated by the SAEM algorithm during the last 500 iterations.

For the Wald test, the SAEM algorithm provides an estimate of the information Fisher matrix, whose inverse matrix's diagonal corresponds to the variance of the parameter estimates. We estimate the parameter $\hat{\beta}$ and its standard error $SE(\hat{\beta})$ under H_1. Under the hypothesis that likelihood is twice continuously differentiable for every θ, the Wald statistic is asymptotically chi-squared distributed. Therefore, we compare the statistic $\hat{\beta}^2/SE^2(\hat{\beta})$ with a χ^2_1 distribution. For both tests, the type I error is estimated by the proportion of trials for which H_0 is rejected as these datasets are simulated without any treatment effect.

The type I error of the Wald test is 4% for the classical SAEM algorithm on the uncensored datasets, and 5.9% for the LRT. We find similar results on

the left-censored datasets using the extended SAEM algorithm. The type I error of the Wald test and the LRT are 4.1% and 5.4%, respectively using this algorithm. These again illustrate the good statistical properties of this extended SAEM algorithm.

5 Application to the Trianon (ANRS81) trial

We illustrate the extended SAEM algorithm on viral load data from the clinical trial TRIANON supported by the French Agence National de Recherche sur le Sida (ANRS). In this study, 144 patients infected with HIV-1, who were randomized into 2 treatment groups, undergo treatment for 72 weeks: 71 patients receive treatment A (lamivudine, d4T and indinavir) and 73 patients treatment B (nevirapine, d4T and indinavir). Viral load is measured at weeks 4 and 8 and every 8 weeks thereafter up through week 72. The HIV RNA assay used in this study has a limit of detection of 20 cp/ml. The comparison of the log reduction of the viral load from baseline to week 72 between the 2 groups with use of a standard statistical approach involving intention to treat shows treatment A to be superior, although the authors expected a superiority of the 3-class regimen (treatment B). See Launay et al. (27) for a more complete description of the study design and results. The data are presented in Figure 4.

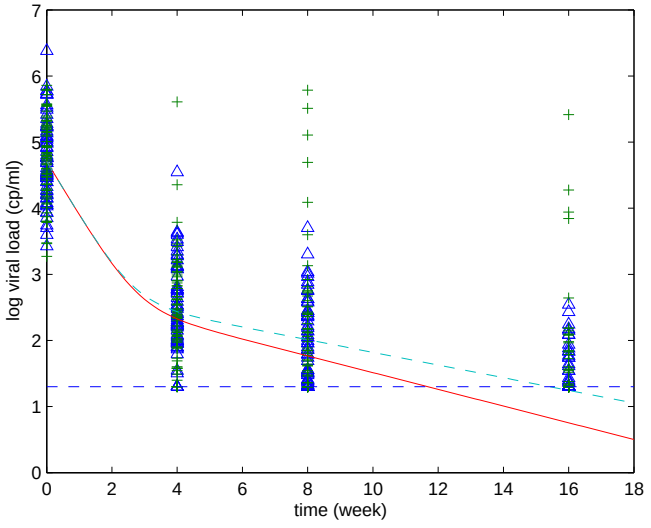


Fig. 4. Observed individual viral load decreases in the 2 groups of patients of the TRIANON trial, with the predicted mean curves obtained with the extended SAEM algorithm in the 2 groups: (Δ), group A observations; (+), group B observations; plain line, group A prediction; dashed line, group B prediction; dotted line, LOQ level.

This new analysis of TRIANON data aims to evaluate the treatment effects on the evolution of the initial viral load decrease. We use the bi-exponential model presented in section 4 to fit the $\log_{10}$ viral load measurements until week 16. There are 64 (out of 275) and 65 (out of 281) observations, respectively, below the LOQ in group A and group B. We compare the extended SAEM algorithm with the usual M_2 method, the one recommended by Ding and Wu (6) to handle left-censored data. For both methods, we analyze the model under the null hypothesis (i.e. without treatment effect). We analyze then the 3 alternative hypotheses proposed by Ding and Wu: AH_1: $\beta_{\lambda_1} \neq 0$; AH_2: $\beta_{\lambda_2} \neq 0$ and AH_3: $\beta_{\lambda_1} \neq 0$ $\beta_{\lambda_2} \neq 0$. In group B, β_{λ_1} and β_{λ_2} are treatment effects added to $\ln\lambda_1$ and $\ln\lambda_2$ in group A

$$\ln \lambda_{1i} = \ln \lambda_1 + \beta_{\lambda_1} G_i \quad \text{and}$$
$$\ln \lambda_{2i} = \ln \lambda_2 + \beta_{\lambda_2} G_i,$$

where $G_i = 0$ denotes a group A subject and $G_i = 1$ a group B subject. We use the one-dimensional Wald test to assess the AH_1 and AH_2 alternative hypotheses. We use a bi-dimensional Wald test for the two-dimensional vector $\beta = (\beta_{\lambda_1}, \beta_{\lambda_2})$ to assess the AH_3 hypothesis. We use the LRT to test all the nested models.

Using the M_2 method, the log-likelihoods are estimated at -617.24, -617.18, -617.0 and -616.92 under H_0, AH_1, AH_2 and AH_3, respectively. None of the 4 LRT is significant at 5%, and we find the same conclusions using the Wald test. With the extended SAEM algorithm, the log-likelihoods are estimated at -472.11, -467.59, -467.28 and -466.09 under H_0, AH_1, AH_2 and AH_3, respectively. The LRT are significant at 5%, except for the test of AH_2 vs AH_3. We find similar results using the Wald tests. Unsurprisingly, the likelihoods are not of the same order with both methods because they come from datasets with different numbers of observations: with the M_2 method, the left-censored data are omitted except for the first ones; with the extended SAEM algorithm, all the left-censored data are kept. The population parameter estimates (and their standard errors) of the final model under AH_2 for the extended SAEM algorithm are $\ln P_1$=10.8 (0.05), $\ln P_2$=6.39 (0.17), $\ln\lambda_1$=-1.30 (0.02), $\ln\lambda_2$=-3.18 (0.05), β_{λ_2}=-0.277 (0.08), ω_1^2=0.106 (0.03), ω_2^2=2.76 (0.46), ω_3^2=0.012 (0.01), ω_4^2=0.059 (0.02), and σ^2=0.38 (0.03). Figure 4 presents the curves predicted by this model, overlaid on the data. The censored data are plotted at the value LOQ. The predicted curves are below the LOQ at week 16 as the extended SAEM algorithm handles the censored data.

In conclusion, we find a significant difference between treatments using the extended SAEM algorithm but not with the recommended M_2 method. The superiority of the treatment A ($\beta_{\lambda_2} < 0$) is in concordance with the previous analysis of the TRIANON dataset (27). In addition, we are able to describe the evolution of the viral load and the treatments' effects. In our example, we

find a trend for a faster viral load decrease under treatment A in the second phase.

6 Discussion

To analyse longitudinal data with left-censored responses, we propose a maximum likelihood estimation method that may be preferred over methods classically used with NLMEM. We extend the SAEM algorithm developed by Kuhn and Lavielle (19) and the monolix 1.1 Matlab function (http://mahery.math.u-psud.fr/∼lavielle/monolix) by including in the simulation step of the SAEM algorithm a simulation of the left-censored data with the right-truncated Gaussian distribution using an accept-reject algorithm proposed by Robert (24). This extended SAEM algorithm is available on the same web address. At the same time, the convergence of the algorithm is monitored from graphical criterion. An automatic implementation of a stopping criterion to optimize both the iterations number and the stochastic approximation step will be considered in the next extension.

We apply this extended SAEM algorithm to model the HIV viral load decrease. The simulation study illustrates the accuracy of our approach. We show that the bias and RMSE obtained by the extended SAEM algorithm are highly satisfactory for all parameters. They almost reach the benchmark obtained before censoring the datasets, although for the last observation time, 72% of the observations are below the LOQ. We consider 2 classical methods obtained either by omitting the data points below the limit or by imputing half the LOQ to left-censored data. We show that the bias and RMSE obtained by the extended SAEM algorithm are much reduced compared to these 2 approaches.

The analysis of the TRIANON dataset also demonstrates the ability of the extended SAEM algorithm to detect differences between 2 treatment groups. This example illustrates the necessity to handle carefully the left-censored data, as the usual approach fails to detect statistical difference between treatment groups. The bi-exponential model that we use is deduced from a differential equation model proposed by Perelson et al. (1) describing the global HIV dynamics with both the viral load decrease and the CD4 increase under treatment. Ding and Wu (6) show that this differential system has an analytic solution under the assumption that the non-infected CD4 cells concentration is constant. Because this assumption is not valid after several week's treatment, the authors recommend using this model only during the first weeks after beginning a treatment, before any rebound of the viral load due to multiple virus mutations. Thus, we consider only the first weeks of the HIV dynamics of TRIANON data. After several weeks, the differential system has no more analytical solution. The exact SAEM algorithm could also be extended to this

case but is out of the scope of this paper.

To take into account the censored-data problem with NLMEM, Wu (7; 17) proposes MCEM algorithms. In his first paper, he proposes a MCEM with an M-step based on the linearization of the model, leading to an approximate maximum likelihood estimation method. In the second paper, he proposes an exact MCEM. However, he emphasizes computational problems, such as slow or even no convergence, especially when the dimension of the random effects is not small. Because the main problem of the MCEM is the simulation of large independent samples of the random effects at each iteration, Wu proposes complex sampling methods for the E-step. As an alternative, he also proposes an approximate MCEM, based on the linearization of the model for both the E- and M- steps, leading again to an approximate maximum likelihood estimation method. To avoid both the linearization step and the computational problem, the SAEM algorithm is a more adapted tool to estimate models with missing or non-observed data such as random effects or censored observations. Indeed, only one realization of the random effects has to be simulated at each iteration, sidestepping the computational problem of the E-step of the MCEM. The extended SAEM requires more iterations to reach the convergence than the standard SAEM, because of the inclusion of a more complex Gibbs algorithm. However, the extended SAEM is still less time consuming than the MCEM. As an example, Wu uses the same bi-exponential HIV dynamic model in his simulation study (i.e. a model with a random effect vector of size $p = 4$). Wu explains that it takes about 1 hour for the exact MCEM algorithm to converge, whereas the extended SAEM algorithm takes about 120 s to converge. The extended SAEM, which is a true maximum likelihood estimation method, is about 10 times faster than the approximate MCEM algorithm proposed by Wu (17). Wu proposes a PX-EM (28) version of its MCEM, which converges faster. The extended SAEM could also be combined with the PX-EM, gaining a similar rate of convergence. The method proposed by Jacqmin-Gadda et al. (5) for censored data analysed with linear mixed models could also be extended to the non-linear case.

We only focus on the left-censored data problem in the context of log viral load observations, but SAEM can be extended to other missing processes such as missing covariates, which Wu (17) includes in his MCEM. This requires making distribution assumptions for the incompletely observed covariates, conditional on the completely observed covariates. This problem is beyond the scope of this article, but it may be solved by a highly similar approach.

In conclusion, the extended SAEM algorithm combines the statistical properties of an exact method together with computational efficiency. We thus recommend the use of this method in NLMEM with left-censored data.

Acknowledgements

The authors thank the scientific committee of the TRIANON-ANRS81 trial for giving us access to their patients' viral load measurements and especially Dr O. Launay, main investigator of TRIANON (Hospital Cochin University, Paris, France) and Dr J.P. Aboulker, methodological coordinator (INSERM SC10, Villejuif, France).

References

[1] A. Perelson, P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, D. Ho, Decay characteristics of HIV-1 infected compartments during combination therapy, Nature 387 (1997) 188–191.

[2] H. Wu, A. Ding, V. De Gruttola, Estimation of HIV dynamic parameters, Stat. Med. 17 (1998) 2463–85.

[3] A. Ding, H. Wu, Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics., Math. Biosci. 160 (1) (1999) 63–82.

[4] A. Ding, H. Wu, A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies., Biometrics 56 (1) (2000) 293–300.

[5] H. Jacqmin-Gadda, R. Thiebaut, G. Chene, D. Commenges, Analysis of left-censored longitudinal data with application to viral load in HIV infection, Biostatistics 1 (2000) 355–68.

[6] A. Ding, H. Wu, Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models, Biostatistics 2 (2001) 13–29.

[7] H. Wu, L. Wu, Identification of significant host factors for HIV dynamics modeled by non-linear mixed-effects models, Stat. Med. 21 (2002) 753–71.

[8] H. Wu, J.-T. Zhang, The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models, Stat. Med. 21 (2002) 3655–75.

[9] S. Beal, Ways to fit a PK model with some data below the quantification limit, J. Pharmacokinet. Pharmacodyn. 28 (2001) 481–504.

[10] J. Hughes, Mixed effects models with censored data with applications to HIV RNA levels, Biometrics 55 (1999) 625–629.

[11] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.

[12] S. Beal, L. Sheiner, Estimating population kinetics, Crit. Rev. Biomed. Eng. 8(3) (1982) 195–222.

[13] M. Lindstrm, D. Bates, Nonlinear mixed-effects models for repeated measures data, Biometrics 46 (1990) 673–687.

[14] R. Wolfinger, Laplace's approximations for non-linear mixed-effect models, Biometrika 80 (1993) 791–795.
[15] H. Wu, L. Wu, A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics, Stat. Med. 20 (2001) 1755–1769.
[16] L. Wu, A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies, J. Am. Stat. Assoc. 97 (2002) 955–964.
[17] L. Wu, Exact and approximate inferences for nonlinear mixed-effects models with missing covariates, J. Am. Stat. Assoc. 99 (2004) 700–709.
[18] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27 (1999) 94–128.
[19] E. Kuhn, M. Lavielle, Maximum likelihood estimation in nonlinear mixed effects models, Comput. Statist. Data Anal. 49 (2005) 1020–1038.
[20] P. Girard, F. Mentré, A comparison of estimation methods in nonlinear mixed effects models using a blind analysis, PAGE 14Abstr 834 [www.page-meeting.org/?abstract=834].
[21] E. Kuhn, M. Lavielle, Coupling a stochastic approximation version of EM with a MCMC procedure, ESAIM: P & S 8 (2005) 115–131.
[22] A. Samson, M. Lavielle, F. Mentré, The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model, submitted.
[23] T. A. Louis, Finding the observed information matrix when using the EM algorithm, J. R. Stat. Soc. B 44 (1982) 226–233.
[24] C. Robert, Simulation of truncated normal variables, Stat. Comput. 5 (1995) 121–125.
[25] A. E. Gelfand, A. F. M. Smith, Sampling-based approaches to calculating marginal densities, J. Am. Stat. Assoc. 85 (1990) 398–409.
[26] V. Duval, M. Karlsson, Impact of omission or replacement of data below the limit of quantification on parameter estimates in a two-compartment model, Pharm. Res. 19 (2002) 1835–40.
[27] O. Launay, L. Grard, L. Morand-Joubert, P. Flandre, S. Guiramand-Hugon, V. Joly, G. Peytavin, A. Certain, C. Lvy, S. Rivet, C. Jacomet, J.-P. Aboulker, P. Yni, A. N. de Recherches sur le SIDA (ANRS) 081 Study Group Yni, Nevirapine or lamivudine plus stavudine and indinavir: examples of 2-class versus 3-class regimens for the treatment of human immunodeficiency virus type 1., Clin. Infect. Dis. 35 (9) (2002) 1096–105.
[28] C. Liu, D. B. Rubin, Y. N. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.

Chapitre 6

Extension de l'algorithme SAEM aux modèles définis par systèmes dynamiques

6.1 Modèles définis par équations différentielles ordinaires

Dans les chapitres précédents, la fonction de régression utilisée pour l'analyse de la décroissance de la charge virale est une solution analytique d'un système différentiel simplifié décrivant le processus de la dynamique virale. Dans de nombreux cas, les hypothèses simplificatrices du système dynamique permettant d'en obtenir une solution analytique ne sont toutefois pas satisfaisantes. Les systèmes dynamiques n'ont alors pas de solution analytique, rendant plus délicate leur utilisation dans le cadre d'une analyse par modèle mixte.

Nous avons proposé deux algorithmes d'estimation adaptés à ces modèles, un algorithme SAEM pour l'approche par maximum de vraisemblance et un échantillonneur de Gibbs pour l'approche bayesienne. En particulier, nous avons développé un nouveau schéma de linéarisation locale pour la résolution d'équations différentielles, fondé sur une linéarisation par rapport au temps et aux paramètres de l'équation, permettant d'optimiser le temps de calcul de l'algorithme de marche aléatoire de Metropolis-Hastings intégré dans ces deux algorithmes d'estimation. Nous avons montré la convergence de ce nouveau schéma de linéarisation locale et celle des deux algorithmes d'estimation. Nous avons ensuite proposé une borne de l'erreur d'estimation due à la méthode numérique de résolution de système différentiel en fonction du pas de cette approximation. Ce résultat n'avait jamais été proposé dans la littérature.

Nous avons illustré cette méthode sur l'exemple simple d'une équation différentielle à une dimension issue d'une problématique réelle de pharmacocinétique. Cette

méthode est appliquée à un système de dimension cinq dans le domaine de la dynamique virale VIH dans le chapitre 7.

Ce travail fait l'objet d'un article soumis à *Journal of Statistical Planning and Inference*.

Estimation of parameters in incomplete data models defined by dynamical systems

Sophie Donnet

sophie.donnet@math.u-psud.fr, Université Paris-Sud, Laboratoire de Mathématiques, Bat 425, 91400 Orsay, France

Adeline Samson

adeline.samson@bch.aphp.fr, INSERM U738, Paris, France; University Paris 7, CHU Bichat-Claude Bernard, Biostatistics unit, Paris, France

Abstract

Parametric incomplete data models defined by ordinary differential equations (ODEs) are widely used in biostatistics to describe biological processes accurately. Their parameters are estimated on approximate models, of which regression functions are evaluated by a numerical integration method. Accurate and efficient estimations of these parameters are critical issues. This paper proposes parameter estimation methods involving either a stochastic approximation EM algorithm (SAEM) in the maximum likelihood estimation, or a Gibbs sampler in the Bayesian approach. Both algorithms involve the simulation of non-observed data with conditional distributions using Hastings-Metropolis (H-M) algorithms. A modified H-M algorithm, including an original Local Linearization scheme to solve the ODEs, is proposed to reduce the computational time significantly. The convergence on the approximate model of all these algorithms is proved. The errors induced by the numerical solving method on the conditional distribution, the likelihood and the posterior distribution are bounded. The maximum likelihood estimation method SAEM coupled with a H-M algorithm is illustrated on a simulated pharmacokinetic nonlinear mixed-effects model defined by an ODE. Simulation results illustrate the ability of this algorithm to provide accurate estimates.

Key words: Bayesian estimation, Incomplete data model, Local linearization scheme, MCMC algorithm, Nonlinear mixed-effects model, ODE integration, SAEM algorithm

Preprint submitted to Elsevier Science *25 August 2005*

1 Introduction

Dynamical systems specified by differential equations are widely used to describe dynamic processes following physical, physiological or biological principles. Thus, when observed data are assumed to follow such a biological process, a statistical model can be introduced, of which regression function is the solution of the corresponding differential system. Difficulties arise frequently with the absence of any analytical solution of the differential system. In these cases, the estimation of the biological parameters requires the use of numerical integration methods, for which accuracy and speed are critical issues. Moreover, a second difficulty is often raised in biomedical situations: these data are frequently described by statistical models involving incomplete data problems. Let us detail below two biological examples combining both previously mentioned problems, the second situation being further exemplified to illustrate the algorithms proposed in this paper.

The first example takes place in the field of brain activation research. Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging technique using the vascular oxygenation contrast as an indirect measure of cerebral activity. Despite the extensive development of such techniques, the coupling mechanisms between neuronal activity and cerebral physiological changes -such as vascular changes- are still poorly understood. Recently, dynamical models such as the Balloon model have been introduced to explain these interactions (Buxton et al., 1998), leading to define statistical models defined by ODEs. Furthermore, some additional hypothesis such as the variability in the hemodynamic response explored by Donnet et al. (2005) can imply the use of statistical incomplete data models. A second relevant example can be found in pharmacokinetics which consist in the study of drug concentration dynamics after its absorption by an individual. These dynamic problems are modeled by differential systems of compartment interactions, the human body being simplified as a system of compartments with nonlinear transfers. The transfer parameters of a specific drug are estimated from repeated measurements of the drug concentration in a large population of patients. Furthermore, since the drug concentration in each individual body follows the same dynamical system but with slightly different parameters, mixed-effects models are widely used in this context, the individual dynamical parameters being considered as non-observed data.

This paper aims at providing a general answer to such statistical problems defined by differential equations, which can be treated as incomplete data models.

Let y be the noised observations of a biological process measured at instants $(t_1, \cdots, t_J)$. The biological process is described by the solution g of an or-

dinary differential equation (ODE), depending on a stochastic non-observed parameter ϕ:

$$y_j = g(t_j, \phi) + \varepsilon_j \qquad \text{for } j = 1 \cdots J.$$

We consider that the observable vector Y is part of a so-called complete vector (Y, ϕ). We assume that both Y and (Y, ϕ) have density functions, $p_Y(y; \theta)$ and $p_{Y,\phi}(y, \phi; \theta)$ respectively, depending on a parameter θ belonging to some subset Θ of the Euclidean space $\mathbb{R}^q$. The estimation of the parameter θ has been widely studied when the regression function g has an explicit form. Two approaches can be followed to tackle this challenge, respectively the maximum likelihood and the Bayesian estimations.

Generally, the maximization of the likelihood of the observations cannot be done in a closed form. Dempster et al. (1977) propose the iterative Expectation-Maximization (EM) algorithm for incomplete data problems. At the k^{th} iteration, the E-step of EM algorithm computes $Q(\theta|\theta_k) = E(\log p_Y(y; \theta)|y; \theta_k)$ while the M-step determines θ_{k+1} maximizing $Q(\theta|\theta_k)$. For cases where the E-step has no closed form, stochastic versions of EM have been introduced. Wei and Tanner (1990) suggest the Monte-Carlo EM (MCEM) estimating $Q(\theta|\theta_k)$ by the averaging of m Monte-Carlo replications. Celeux and Diebolt (1985) propose the Stochastic EM algorithm (SEM), the first stochastic version of EM, which is a special case of MCEM in which $m = 1$. Recently, Wu (2004) emphasizes that MCEM is computationally intensive. As an alternative, Delyon et al. (1999) propose the Stochastic Approximation EM algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q(\theta|\theta_k)$. These methods require the simulation of the non-observed data ϕ. For cases where this simulation can not be performed in a closed form, Kuhn and Lavielle (2004) suggest to resort to iterative methods such as Monte Carlo Markov Chain algorithms (MCMC).

The Bayesian approach estimates the posterior distribution $p_{\theta|Y}(\cdot|y)$ of θ, a prior $p_\theta(\cdot)$ being given. Because of the conditional independence structure of $p_{\theta|Y} = \int p_{\theta|Y,\phi}\ p_{\phi|Y}\ d\phi$ and $p_{\phi|Y} = \int p_{\phi|Y,\theta}\ p_{\theta|Y}\ d\theta$, Gelfand and Smith (1990) propose a Gibbs sampling to evaluate these two integrals simultaneously. At iteration k, $\phi^{(k)}$, a realization of ϕ, is simulated with $p_{\phi|Y}(\cdot, \theta^{(k-1)})$ followed by $\theta^{(k)}$, a realization of θ with $p_{\theta|Y,\phi}(\cdot, \phi^{(k)})$. Consequently, as in maximum likelihood estimation, difficulties arise when the simulation of the conditional distribution can not be performed in a closed form. For these cases, a Hastings-Metropolis (H-M) algorithm can be included in the Gibbs sampler.

The use of the H-M algorithm in estimation algorithms requires the evaluation of the regression function g at each iteration. When g is a non-analytical solution of a dynamical system, it is evaluated using a numerical integration method. Thus a trade-off between accuracy, stability and computational cost is required. In this paper, we detail the Local Linearization scheme (see e.g.

Biscay et al., 1996; Ramos and García-López, 1997; Jimenez, 2002) not only because of its stability performances but also because this scheme can be extended to a so-called modified Local Linearization scheme. This modified scheme is particularly adapted to situations encountered with the H-M algorithm, when the dynamical system has to be solved successively for different sets of the ϕ parameter. In such situations, it allows a significant decrease in the computational time. The estimation algorithms are then applied to an approximate model of which regression function is an approximate solution of the ODE.

The objective of this research is to quantify the error induced by the numerical approximation of the regression function g. The paper is organized as follows. Section 2 defines the original statistical model; the Local Linearization scheme and its modified version are detailed; the approximate statistical model resulting from the numerical approximation is introduced. Section 3 focuses on the H-M algorithm to simulate the non-observed data ϕ with the conditional distribution. A modified version of the H-M algorithm including the extended Local Linearization scheme is proposed to reduce the computational time. Their convergence is proved on the approximate model, and the error induced by the numerical approximation of g is quantified on the conditional distribution. Section 4 is dedicated to the parameter estimation algorithms. Concerning maximum likelihood and Bayesian estimations, the standard algorithms are adapted to solve the approximate model. The error induced by the use of the numerical solving method is bounded respectively on the likelihood and the posterior distribution. This error is distinct from the error on the estimates induced by the estimation algorithm which is evaluated by their standard errors. Finally, the SAEM algorithm is illustrated on a nonlinear mixed-effects model deriving from pharmacokinetics in section 5.

2 Models and notations

2.1 An incomplete data model defined by ODEs

Let $y = (y_j)_{j=1..J}$ denote the observations measured at times $(t_1, \cdots, t_J)$. We consider the incomplete data model, called model $\mathcal{M}$, defined as follows:

$$\begin{aligned} y_j &= g(t_j, \phi) + \varepsilon_j \qquad 1 \leq j \leq J \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma^2) \\ \phi &\sim \pi(\cdot; \beta) \end{aligned} \tag{$\mathcal{M}$}$$

where $g(.)$ is a nonlinear function of ϕ, ε_j represents the error of the measurement j, σ^2 is the residual variance, ϕ is a non-observed random parameter,

distributed with the density $\pi(\cdot, \beta)$, depending on the parameter β. The parameter $\theta = (\beta, \sigma^2)$ belongs to some open subset $\Theta \subset \mathbb{R}^q$.

Let g be written $g = H \circ f$, where $H : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a known function and, $f : \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ is defined as the solution of the following ODE:

$$\begin{aligned} &\frac{\partial f(t,\phi)}{\partial t} = F(f(t,\phi), t, \phi) \\ &f(t_0, \phi) = f_0(\phi) \end{aligned} \tag{1}$$

with a known function $F : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ and the initial condition $f_0(\phi) \in \mathbb{R}^d$, $t \in [t_0, T]$.

We make the following additional assumptions:

- Assumption **H1**: π has a compact support $K_1 \subset \mathbb{R}^k$, and there exist two constants a and, b such that

$$0 < a < \pi(\phi; \beta) < b \text{ for all } \phi \in K_1.$$

- Assumption **H2**: $F : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ is $\mathcal{C}^2$ on its definition domain, and $\phi \longrightarrow f_0(\phi) \in \mathbb{R}^k$ is $\mathcal{C}^1$ on K_1.
- Assumption **H3**: H is an L_H-lipschitzian function.

2.2 Approximation of the regression function

A great variety of numerical schemes have been proposed to solve ODEs (see e.g. Hairer et al., 1987). The accuracy of such numerical methods is qualified by the order and the step size of these schemes. The numerical scheme is applied on sub-intervals $[t_n, t_{n+1}[$, $n = 0, \ldots, N-1$, of the time interval $[t_0, T]$, with $t_N = T$. The maximal length or step size of the sub-intervals is denoted h. The order of a numerical scheme is defined as follows:

Definition 1 *Let f_h be the resulting approximate function obtained by a numerical integration scheme of step size h. This scheme is of order p if there exists a constant C such that*

$$\sup_{t \in [t_0, T]} |f(t, \phi) - f_h(t, \phi)| \leq C h^p.$$

Of all the numerical schemes, the Local Linearization (LL) scheme provides a good trade-off between computational cost and numerical stability, as exposed in Biscay et al. (1996). It derives from the local linearization of the right term of the ODE (1) with respect to time t, and the exact integration of the deduced linear differential equation. Its implementation requires matrix

exponential computations using new algorithms such as Pade or Schur methods, which have proved their efficiency and stability. The LL scheme has the additional advantage of preserving stability properties on stiff systems (see e.g. Ramos and García-López, 1997; Jimenez et al., 2002). In cases where the ODE depends on parameter ϕ, we extend this scheme using a Taylor expansion with respect to time t and parameter ϕ. More precisely, let the solution at ϕ_0 be computed using the LL scheme. Let ϕ be in a neighborhood of ϕ_0. On each sub-interval $[t_n, t_{n+1}[$, $n = 0, \ldots, N-1$, the solution f_{h,ϕ_0} of the following linear equation

$$\begin{aligned}\frac{\partial f(t,\phi)}{\partial t} = & F(f(t_n,\phi_0),t_n,\phi_0) + \frac{dF}{df}\left(f(t_n,\phi_0),t_n,\phi_0\right)\left(f(t,\phi) - f(t_n,\phi_0)\right) \\ & + \frac{dF}{dt}\left(f(t_n,\phi_0),t_n,\phi_0\right)(t-t_n) + \frac{dF}{d\phi}\left(f(t_n,\phi_0),t_n,\phi_0\right)(\phi-\phi_0).\end{aligned}$$

is evaluated. Details of this scheme (called LL2 scheme) are given in appendix B. This LL2 scheme does not involve any additional computation of matrix exponentials: the required matrix exponential has already been computed with the LL scheme at ϕ_0. This reduces the computational time, which is a key issue in iterative processes. For instance, during H-M algorithm implementation, the ODE has to be integrated at ϕ contained in a neighborhood of ϕ_0, for which the ODE has already been solved by LL. This leads us to propose a modified version of the H-M algorithm, taking advantage of these schemes (see part 3.1.2).

Convergence properties of the two previous numerical schemes are given in the following lemma. Let us consider the following assumption:

- Assumption **H1′**: ϕ remains in the compact set $K_1 \in \mathbb{R}^k$.

Lemma 2 *Let ϕ_0 and ϕ be in K_1. Let $f(.,\phi)$ be the exact solution of the ODE (1), $f_h(.,\phi)$ the one obtained by the LL scheme with step size h, and $f_{h,\phi_0}(.,\phi)$ the LL2 solution. Assume that **H1′** and **H2** hold. Then:*

(1) there exists a constant C independent of ϕ, such that, for any $t \in [t_0, T]$ and for any ϕ,

$$|f(t,\phi) - f_h(t,\phi)| \leq Ch^2,$$

(2) there exist constants C_1 and C_2 such that, for any $t \in [t_0, T]$ and for any ϕ,

$$|f(t,\phi) - f_{h,\phi_0}(t,\phi)| \leq \max(C_1 h^2, C_2 \left\|\phi - \phi_0\right\|^2_{\mathbb{R}^k}).$$

Part 1 is proved in Ramos and García-López (1997), part 2 is proved in appendix B.

2.3 An approximate incomplete data model

In practice, estimation algorithms require the numerical approximation of the regression function and are thus applied to an approximate version of the model $\mathcal{M}$. Let f_h be the approximate solution of the ODE (1), obtained by a numerical integration method of step size h and order p. Let the approximate statistical model $\mathcal{M}_h$ be defined by:

$$\begin{array}{ll} y_j = g_h(t_j, \phi) + \varepsilon_j & 1 \leq j \leq J \\ \varepsilon_j \sim \mathcal{N}(0, \sigma^2) & \\ \phi \sim \pi(\cdot; \beta) & \end{array} \qquad (\mathcal{M}_h)$$

where $g_h = H \circ f_h$. Subsequently, the different distributions of the model $\mathcal{M}_h$ are subscripted with h.

3 Simulation of non-observed data with the conditional distribution

In this paper, the Hastings-Metropolis (H-M) algorithm is combined successively with the SAEM algorithm in the maximum likelihood estimation, and the Gibbs sampler in the Bayesian estimation. The H-M algorithm updates ϕ in the target distribution $p(\phi|y;\theta)$. The computation of its acceptance probabilities requires an explicit expression of the regression function. As a consequence, this H-M algorithm can only be applied to the approximate statistical model $\mathcal{M}_h$.

3.1 Two Hastings-Metropolis algorithms

To implement the H-M algorithm, it is necessary to specify a suitable proposal density. Several possible choices of proposal density are discussed in the literature. In the following implementation, the proposal density is chosen to be either the prior density π or a symmetric distribution centered at the current value of ϕ. The standard H-M algorithm is briefly presented, followed by its modified version including the LL2 scheme with the second proposal density.

3.1.1 The standard H-M algorithm

The iterative H-M algorithm is outlined as follows. At step $r+1$, given $\phi^{(r)}$, move the chain to a new value ϕ^c, generated from the proposal density $q(.|\phi^{(r)})$.

Evaluate the acceptance probability of the move, $\rho(\phi^{(r)}, \phi^c)$ given by

$$\rho(\phi^{(r)}, \phi^c) = \min\left\{1, \frac{p_{h,Y|\phi}(\phi^c)\pi(\phi^c)}{p_{h,Y|\phi}(\phi^{(r)})\pi(\phi^{(r)})}\frac{q(\phi^{(r)}|\phi^c)}{q(\phi^c|\phi^{(r)})}\right\}.$$

If the move is accepted, choose $\phi^{(r+1)} = \phi^c$. If it is not accepted, $\phi^{(r+1)} = \phi^{(r)}$ and the chain does not move. The choice of the proposal density q is essentially arbitrary, although in practice a careful choice will help the algorithm to move quickly inside the parameter space. Two proposal densities are combined. First the prior density π allows to move inside the parameters space efficiently. Second, a symmetric distribution centered at the current value of ϕ such that ϕ remains in K_1, results in the so-called random-walk H-M algorithm (see e.g. Bennet et al., 1996).

Hence, for each iteration, we need to derive $g(t, \phi)$ by solving the ODE (1) so that the acceptance probability can be evaluated. Thus, choosing the most time-saving computational solving method is fundamental. To that purpose, a slightly modified H-M, including the LL2 scheme, is suggested in the next section.

3.1.2 A modification of the H-M algorithm

Let the model $\mathcal{M}_h$ be defined by the LL scheme. The second proposal density requires solving the ODE (1) at ϕ^c in a bounded neighborhood of $\phi^{(r)}$, for which the LL approximate solution has been computed. Let the second proposal density verify the following property, called property (2): there exists $\eta > 0$ such that, almost surely,

$$\|\phi^{(r)} - \phi^c\|_{\mathbb{R}^k} < \eta. \tag{2}$$

The standard H-M algorithm is slightly modified by including the LL2 scheme. At step $r+1$, given $\phi^{(r)}$, move the chain to a new value $\phi^c = \phi^{(r)} + \delta$. Evaluate the acceptance probability of the move, $\rho^{(2)}(\phi^{(r)}, \phi^c)$ given by

$$\rho^{(2)}(\phi^{(r)}, \phi^c) = \min\left\{1, \frac{p^{(2)}_{h,Y|\phi}(\phi^c)\pi(\phi^c)}{p_{h,Y|\phi}(\phi^{(r)})\pi(\phi^{(r)})}\right\},$$

where $p^{(2)}_{h,Y|\phi}(.)$ is evaluated using the LL2 scheme. If the move is accepted, $\phi^{(r+1)} = \phi^c$, and $f(t, \phi^{(r+1)})$ and the $p_{h,Y|\phi}(\phi^{(r+1)})$ density are evaluated using the LL scheme. If it is not accepted, $\phi^{(r+1)} = \phi^{(r)}$ and the chain does not move.

For cases where the candidate is not accepted, no additional matrix exponential is computed, significantly reducing the computational time.

3.2 Convergence

Both the usual and modified H-M algorithms converge. The rates of convergence of such algorithms have been widely studied (see e.g. Tierney, 1994), hence are not discussed here. We focus on the error induced by the numerical integration scheme by quantifying the distance between the invariant distribution simulated by the H-M algorithm and the original distribution of interest $p_{\phi|Y}$.

Theorem 3 *Let f be the exact solution of ODE (1). Let $p_{\phi|Y}$ be the conditional distribution for the model $\mathcal{M}$. Assume that **H1**, **H2** and **H3** hold.*

(1) Let f_h be the approximation obtained by a numerical integration method of step size h and order p. Let $p_{h,\phi|Y}$ be the conditional distribution of the model $\mathcal{M}_h$. Then on the model $\mathcal{M}_h$, the H-M algorithm converges towards its stationary distribution $p_{h,\phi|Y}$. Furthermore, there exists a constant C_y such that for any small h,

$$D(p_{\phi|Y}, p_{h,\phi|Y}) \leq C_y h^p.$$

where $D(\cdot,\cdot)$ denotes the total variation distance.

(2) The modified H-M algorithm converges towards a stationary distribution $p^{(2)}_{h,\phi|Y}$. Furthermore, let η verify the property (2) and let the chain $(\phi^{(r)})$ remain in K_1. Then, there exists a constant $C^{(2)}_y$ such that

$$D(p_{\phi|Y}, p^{(2)}_{h,\phi|Y}) \leq C^{(2)}_y \eta + C_y h^2.$$

The proof is given in appendix A. This is illustrated by a numerical example in section 5.

4 Estimation of parameters

In the following section, we extend to incomplete data models defined by ODEs, the SAEM algorithm coupled with the modified H-M algorithm for the maximum likelihood estimation, and the Gibbs sampling algorithm for the Bayesian approach.

4.1 Maximum likelihood approach

The EM algorithm proposed by Dempster et al. (1977) maximizes the $Q(\theta|\theta') = E(\log p_{Y,\phi}(\cdot;\theta)|y;\theta')$ function in two steps. At the k^{th} iteration, the E-step is

the evaluation of $Q_k(\theta) = Q(\theta \,|\, \theta_k)$ while the M-step updates θ_k by maximizing $Q_k(\theta)$. For cases where the E-step has no closed form, Delyon et al. (1999) introduce a stochastic version SAEM of the EM algorithm. The $Q_k(\theta)$ integral is evaluated by a stochastic approximation procedure. The E-step is divided into a simulation step (S-step) of the non-observed data ϕ_k with the conditional distribution $p_{\phi|Y}(.;\theta_k)$ and a stochastic approximation step (SA-step):

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k \left(\log(p_{Y,\phi}(.;\theta_k)) - Q_k(\theta)\right),$$

where (γ_k) is a sequence of positive numbers decreasing to 0. They prove the convergence of this algorithm under general conditions in the case where $p_{Y,\phi}$ belongs to a regular curved exponential family. When the non-observed data cannot be directly simulated, Kuhn and Lavielle (2004) suggest using a MCMC scheme by building a Markov chain with $p_{\phi|Y}(\cdot;\theta_k)$ as unique stationary distribution at the k^{th} iteration.

When the regression function is defined by ODE, SAEM is implemented on the model $\mathcal{M}_h$. Let $\pi(\cdot,\beta)$ be such that $p_{h,Y,\phi}$ belongs to the exponential family:

$$p_{h,Y,\phi}(\cdot;\theta) = \exp\left\{-\psi(\theta) + \langle S_h(y,\phi), \Phi(\theta)\rangle\right\},$$

where ψ and Φ are two functions of θ, $\langle\cdot,\cdot\rangle$ is the scalar product and $S_h(y,\phi)$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\mathcal{S}$ of $\mathbb{R}^m$. Let $\pi(\cdot;\beta)$ be of class $\mathcal{C}^m$. At the k^{th} iteration, the SAEM algorithm is:

- S-Step: the non-observed data ϕ_k is simulated by the H-M algorithm developed in section 3 with $p_{h,\phi|Y}(\cdot;\theta_k)$ as unique stationary distribution,
- SA-Step: s_{k+1} is updated by the stochastic approximation:

$$s_{k+1} = s_k + \gamma_k(S_h(y,\phi_k) - s_k),$$

- M-Step: θ_k is updated by

$$\theta_{k+1} = \arg\max_{\theta} \left(-\Psi(\theta) + \langle s_{k+1}, \Phi(\theta)\rangle\right).$$

Kuhn and Lavielle (2004) propose estimates of the Fisher information matrix, using the Louis's missing information principle (Louis (1982)), either by importance sampling or by stochastic approximation. We adapt their estimates when the regression function is not known analytically and, as a consequence, the extended SAEM supplies the standard errors of the estimates.

The convergence of SAEM is proved on $\mathcal{M}_h$ and the distance between the likelihoods of the two models is quantified in the following theorem.

Theorem 4 *Let us consider a numerical scheme of step size h and order p.*

*Let **H1**, **H2** and **H3** hold. Let (γ_k) be a sequence of positive numbers decreasing to 0 such that for any k in $\mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.*

(1) Assuming the sequence $(s_k)_{k\geq 0}$ takes its values in a compact set of $\mathcal{S}$, the sequence $(\theta_k)_{k\geq 0}$ obtained by the SAEM algorithm on $\mathcal{M}_h$, converges almost surely towards a (local) maximum of the likelihood $p_{h,Y}(y)$.
(2) For any $\sigma_0^2 > 0$, there exists a constant θ-independent C such that

$$\sup_{\theta=(\beta,\sigma^2)\ |\sigma^2>\sigma_0^2} |p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq Ch^p.$$

Hence, as a principal consequence of this theorem, and assuming regularity hypotheses on the Hessians of the likelihoods of both models $\mathcal{M}$ and $\mathcal{M}_h$, the bias of the estimates induced by both the numerical approximation and the estimation algorithm, is controlled.

PROOF.

(1) Assumptions of convergence of the SAEM algorithm are checked by the model $\mathcal{M}_h$. See Kuhn and Lavielle (2004) for more details.
(2) In the proof of theorem 3, we obtain the result (1b) for a fixed θ and small enough h that:

$$|p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq \frac{C_y}{(2\pi\sigma^2)^{J/2}} h^p.$$

Consequently, for any $\sigma_0^2 > 0$, for any $\theta \in \Theta$ with $\sigma^2 \geq \sigma_0^2$, there exists a constant C, independent of θ, such that

$$|p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq Ch^p.$$

4.2 Bayesian approach

Let p_θ be a prior distribution on the parameter θ. Following Gilks et al. (1996), σ has a gamma distribution $\mathrm{Ga}(\nu_0/2, \sigma_0\nu_0/2)$, μ is multivariate $\mathrm{N}(c, C)$ and Ω^{-1} is Wishart $\mathrm{W}(\rho, [\rho R]^{-1})$ for known ν_0, σ_0, c, C, R and ρ. The Bayesian approach consists in the evaluation of the posterior distribution $p_{\theta|Y}$. The iterative Gibbs sampling algorithm is outlined as follows (see Huang et al., 2004, for more details).

Step 1. Initialize the iteration counter of the chain $k = 1$ and start with initial values $\Gamma^{(0)} = (\sigma^{-2(0)}, \mu^{(0)}, \Omega^{(0)}, \phi^{(0)})^t$.
Step 2. Obtain a new value $\Gamma^{(k)} = (\sigma^{-2(k)}, \mu^{(k)}, \Omega^{(k)}, \phi^{(k)})^t$ from $\Gamma^{(k-1)}$ through successive generation of values:

(1) $\sigma^{-2(k)} \sim q(\sigma^{-2}|\mu^{(k-1)}, \Omega^{(k-1)}, \phi^{(k-1)}, y)$
$\mu^{(k)} \sim q(\mu|\sigma^{-2(k)}, \Omega^{(k-1)}, \phi^{(k-1)}, y)$
$\Omega^{(k)} \sim q(\Omega|\sigma^{-2(k)}, \mu^{(k)}, \phi^{(k-1)}, y)$
(2) $\phi^{(k)} \sim q(\phi|\sigma^{-2(k)}, \mu^{(k)}, \Omega^{(k)}, y)$

Step 3. Change the counter from k to $k+1$ and return to Step 2 until convergence is reached.

The parameter ϕ is the only parameter which has non-standard full-conditional distribution. To generate a realization of it, Bennet et al. (1996) describe several approaches, such as a Rejection Gibbs, a Ratio Gibbs, independent or Random-walk H-M algorithms. Gilks et al. (1996) recommend the use of initial iterations allowing a "burn-in" phase, followed by a large number of iterations.

For models defined by ODEs, as seen before, the inclusion of such H-M algorithms require, at each iteration, the evaluation of $g(\phi, t_j)$. Hence, the Gibbs algorithm is implemented on the approximate statistical model $\mathcal{M}_h$. The modified version of the H-M algorithm detailed in section 3, can be included at step 2 of the Gibbs algorithm.

In practice, we estimate the posterior distribution of the model $\mathcal{M}_h$ instead of the distribution of interest $p_{\theta|Y}$. The following theorem quantifies the total variation distance between the posterior distribution $p_{h,\theta|Y}$ and this original distribution of interest, $p_{\theta|Y}$.

Theorem 5 *Let us consider a numerical scheme of step size h and order p. Let $p_{\theta|Y}$ and $p_{h,\theta|Y}$ be the posterior distributions respectively of $\mathcal{M}$ and $\mathcal{M}_h$. Assume that **H1**, **H2** and **H3** hold.*

(1) The Gibbs sampling algorithm converges on the model $\mathcal{M}_h$.
(2) There exists a $\mathbf{y}$-dependent constant C_y such that

$$D(p_{h,\theta|Y}, p_{\theta|Y}) \leq C_y h^p.$$

PROOF.

(1) Assumptions of convergence of the Gibbs sampling algorithm are checked on the model $\mathcal{M}_h$, see Carlin and Louis (2000) for more details.
(2) By Bayes theorem, we have

$$p_{\theta|Y} = \frac{p_{Y|\theta} p_\theta}{p_Y}$$

where $p_Y = \int p_{Y|\theta} p_\theta d\theta$, and the same equality for the model $\mathcal{M}_h$. From the result (1b) of the proof of theorem 3, we deduce that there exists a constant C, independent of θ, such that for any $\theta \in \Theta$ with $\sigma^2 > \sigma_0^2 > 0$,

$|p_{Y|\theta}(y) - p_{h,Y|\theta}(y)| \leq Ch^p$ and $|p_Y(y) - p_{h,Y}(y)| \leq Ch^p$. We now bound $|p_{\theta|Y}(\theta) - p_{h,\theta|Y}(\theta)|$:

$$\begin{aligned}|p_{\theta|Y}(\theta) - p_{h,\theta|Y}(\theta)| &\leq \frac{p_\theta(\theta)}{|p_Y(y)|}\left||p_{Y|\theta}(y) - p_{h,Y|\theta}(y)| + \frac{p_Y(y)}{p_{h,Y}(y)}|p_Y(y) - p_{h,Y}(y)|\right| \\ &\leq \frac{Ch^p}{|p_Y(y)|}p_\theta(\theta)\left|1 + \frac{p_{h,Y|\theta}(y)}{p_{h,Y}(y)}\right| = \frac{Ch^p}{|p_Y(y)|}\left(p_\theta(\theta) + p_{h,\theta|Y}(\theta)\right).\end{aligned}$$

Thus

$$D(p_{\theta|Y}, p_{h,\theta|Y}) \leq \frac{Ch^p}{|p_Y(y)|}.$$

5 Application to nonlinear mixed-effects model

Nonlinear mixed-effects models are widely used in pharmacokinetics (PK) and pharmacodynamics (PD) to estimate PK/PD parameters. They are interesting because of their capacity to discriminate the intra- from the inter-subject variabilities and to test covariable effect on the PK/PD parameters. They are modeled by:

$$\begin{aligned}&y_{ij} = C(t_{ij}, \phi_i) + \varepsilon_{ij} \\ &\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ &\phi_i \sim \pi(.; \beta),\end{aligned}$$

where y_{ij} is the observation of the drug concentration C for subject i, $i = 1, \ldots, N$, at time t_{ij}, $j = 1, \ldots, n_i$ and ϕ_i is the vector of individual non-observed PK/PD parameters of subject i.

We consider a case where the drug concentration C is defined through a differential equation without analytical solution. On a simulated dataset, we illustrate applications of the H-M and of the SAEM algorithms. We do not implement the Gibbs algorithm adapted to ODE models, since the PKBugs software already proposes estimation by Gibbs sampling for PK/PD models with ODE.

5.1 Numerical settings

The following one-compartment pharmacokinetic system with a first order absorption and a Michaelis-Menten saturable elimination describes the con-

centration C of a drug:

$$\frac{dC}{dt}(t,\phi) = \frac{k_a D}{V} e^{-k_a t} - \frac{V_m C(t,\phi)}{k_m + C(t,\phi)}, \tag{3}$$

where D is the known administered dose, V the total volume of distribution, k_a the absorption constant, k_m the Michaelis-Menten constant and V_m the maximum rate of metabolism. Hence, the parameter vector is $\phi = [V, k_a, k_m, V_m] \in \mathbb{R}^4$.

We consider the pharmacokinetic parameters of hydroxurea, an anti-cancerous drug, studied by Tracewell et al. (1995): V=12.2 L, k_a=2.72 h^{-1}, k_m=0.37 mmol/L, V_m=0.082 mmol/h/L. One dataset of 20 patients is simulated with a dose D = 13.8 mmol and measurements at time $t = 0, 0.5, 1, 1.5, 2$ hours and then every hour until 12 hours. We choose a Gaussian distribution for ϕ with a diagonal variance-covariance matrix (diagonal components equal to 0.4). The residual variance σ^2 is set to 0.01. The numerical method used to simulate this dataset is the ode45 solving Matlab function, that implements a Runge-Kutta scheme of the fourth order, with a very small maximal step size of resolution h equal to 0.001. Data are plotted on figure B.1.

[Fig. 1 about here.]

5.2 Results

The numerical LL and LL2 schemes are compared to the Runge-Kutta (RK) algorithm as implemented in the ode45 Matlab function. The ODE (3) is solved at a ϕ_0 equal to the simulated parameter values and then at $\phi = 1.05 \times \phi_0$. With a step size $h = 0.1$, the solutions obtained at ϕ_0 are identical for the two numerical methods ode45 and LL. The difference between the solutions obtained at ϕ by ode45, LL, and LL2 is not distinguishable. Furthermore, LL2 is seven times faster than LL or ode45, which is a major advantage, taking into account that it is included in an iterative algorithm.

Several numerical integration methods included in the H-M algorithm are compared on the simulation of the conditional distributions. 1000 independent 200-long Markov chains are generated for each method. The first method is the ode45 function with maximal step size $h = 0.001$ in order to obtain an exact cumulative distribution function; this is the reference as this numerical method is of the fourth order with a very small step size. Then we compare the two cumulative distribution functions obtained by using at first ode45 with the default step size $h = 0.1$ and then the LL and LL2 schemes. As seen on Figure B.2, plotting the empirical cumulative distribution functions, the three numerical methods provide similar simulated conditional distributions.

[Fig. 2 about here.]

Finally, the SAEM algorithm is applied to the simulated dataset. Initial values have been arbitrarily chosen and are presented in Table B.1. The evolutions of each parameter estimate are plotted against iterations on Figure B.3. The estimates converge rapidly to a neighborhood of the simulated values. The parameter estimates and their standard errors obtained by SAEM are presented in Table B.1.

[Fig. 3 about here.]

They are compared with those obtained by the NONMEM software, proposed by Beal and Sheiner (1982) and used by 80% of the pharmacokineticists in drug companies. To date, NONMEM is the only software available which is able to solve the maximum likelihood estimation problem with nonlinear mixed models defined by ODEs. NONMEM is an implementation of the FOCE algorithm (First Order Conditional Estimate), which is based on a first order linearization of the regression function around the conditional estimates of the parameters ϕ. The numerical solving method used is an implementation of the fourth order Runge-Kutta algorithm. The NONMEM software evaluates the standard errors of the estimates by linearization. Results using NONMEM are also presented in Table B.1.

[Table 1 about here.]

In this example, it takes about ten minutes for SAEM to converge using a conventional Intel Pentium IV 3,2 GHz workstation. Using the same computer, the NONMEM software stops after about ten minutes without convergence towards the maximum of the likelihood. The estimate of V_m does not change from its initial value, the estimate of k_m is far from its simulation value, var k_m is estimated near zero while var V_m and var k_a are overestimated. The NONMEM software fails to evaluate the standard errors of all estimates. All the SAEM estimates almost reach the simulated values. The SAEM algorithm achieves the evaluation of the information Fisher matrix, and almost all the standard errors of the estimates are satisfactory.

6 Discussion

This paper extends the statistical approaches used to estimate incomplete data model parameters to the frequent cases where such a model $\mathcal{M}$ is defined by a dynamical system. To that purpose, an approximate model $\mathcal{M}_h$ is introduced, of which regression function is evaluated by a numerical integration method. The standard estimation algorithms are adapted to estimate this approximate

model. The convergence of the H-M, the SAEM and the Gibbs Sampling algorithms on the model $\mathcal{M}_h$ is proved.

This paper quantifies the error induced by the use of a numerical solving method. The errors on the conditional distribution, the likelihood and the posterior distribution between the model $\mathcal{M}$ and the approximate model $\mathcal{M}_h$ are controlled by h^p, where h is the step size and p the order of the numerical integration method. This error is distinct from the error on the estimates induced by the estimation algorithms, which is classically controlled by the standard errors evaluated through the Fisher information matrix of the estimates.

This paper proposes an extended version of the LL scheme, using the dependence to the parameter ϕ of the ODE. As the LL2 scheme does not require any exponential matrix computation, the computational cost of the modified H-M algorithm proposed in this paper is significantly reduced. Biscay et al. (1996), Ramos and García-López (1997) and Ramos (1999) study numerically the LL scheme performances. It is more accurate than a modified explicit second order Euler scheme but less accurate than the explicit fourth-order Runge-Kutta methods. To address this issue, a higher order LL scheme could be implemented, but this may only provide an insignificant improvement when included in stochastic algorithms, as shown by the simulation results of the H-M algorithm numerical properties. Furthermore, Biscay et al. (1996) show that the LL scheme keeps stable properties on stiff dynamical systems. Thus these two LL schemes provide an interesting trade-off between computational cost and numerical stability when included in an iterative stochastic algorithm.

Regarding the maximum likelihood, we study the SAEM algorithm instead of the Monte-Carlo EM proposed by Wei and Tanner (1990) or Wu (2004). With an analytical regression function, the MCEM is computational intensive because of the large sample of non-observed data simulated at each iteration, while the Stochastic Approximation method requires the generation of only one realization of non-observed data at each iteration. Thus, due to computational time considerations, we only extend the SAEM algorithm to the case of regression function implicitly defined.

The SAEM algorithm is applied to a dataset simulated using a pharmacokinetic model defined by ODEs. The SAEM estimates are compared with those obtained by the standard estimation software NONMEM, the only available software providing estimates by maximum likelihood in nonlinear mixed models defined by ODEs. SAEM provides satisfying estimates and standard errors of the parameters, while NONMEM does not converge on this simulated example and fails to evaluate the standard errors. The estimation algorithm implemented in NONMEM is based on the linearization of the regression function. Despite the fact that Vonesh (1996) highlights problems of consis-

tence and convergence of the estimates produced by such estimation methods, NONMEM is used by 80% of the pharmacokineticists in the pharmaceutical industry. The simulation results presented in this paper point out the poor ability of this software to estimate the parameters in nonlinear mixed model defined through ODEs. Thus we recommend using the SAEM algorithm to analyze such incomplete data problems defined by ODE. The SAEM algorithm is implemented by the MONOLIX group in a free Matlab function (http://mahery.math.u-psud.fr/∼lavielle/monolix/index.html). The extension of the monolix function to ODE models will soon be available on the same web-site.

Several methods have been suggested to simulate the non-observed data included in a Bayesian approach. Wakefield et al. (1994) sample the non-observed data using a ratio-of-uniform while Tierney (1994) proposes using a Hastings-Metropolis algorithm. Gilks et al. (1996) summarize these methods in their book. We do not implement the Gibbs algorithm adapted to ODE models, as the PKBugs software already proposes an estimation by Gibbs sampling for PK/PD models with ODE.

We implement SAEM on a non-stiff dynamical system of one dimension with a homoscedastic model. Extended implementation of SAEM should be done on further examples with high-dimension or stiff ODEs and heteroscedastic models. The analysis of a real dataset using SAEM would also extend this work.

Acknowledgements

The authors are grateful to their advisor Professor Marc Lavielle for his constructive advice and help. The authors thank Professor France Mentré for her helpful comments. The authors would like to thank also Rolando Biscay for helpful discussions about Local Linearisation schemes.

A Proof of theorem 3:

(1) (a) Following Tierney (1994), the H-M algorithm provides a uniformly ergodic Markov chain with $p_{h,\phi|Y}$ as invariant distribution, as soon as π is one of the proposal density.

(b) By Bayes theorem, we have $p_{\phi|Y}(\phi) = \frac{p_{Y|\phi}(y|\phi)\pi(\phi)}{p_Y(y)}$, and

$$\left|p_{\phi|Y}(\phi) - p_{h,\phi|Y}(\phi)\right| = \frac{\pi(\phi)}{p_Y(y)}\left[\left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| + \frac{p_{h,Y|\phi}(y)}{p_{h,Y}(y)}\left|p_Y(y) - p_{h,Y}(y)\right|\right].$$

We have $\left|p_{Y|\phi}(y|\phi) - p_{h,Y|\phi}(y|\phi)\right| = |p_{Y|\phi}(y)| \left|1 - \frac{p_{h,Y|\phi}(y)}{p_{Y|\phi}(y)}\right|$, and

$$p_{Y|\phi}(y) = \frac{1}{(2\pi\sigma^2)^{J/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\left(y_j - g(t_j,\phi)\right)^2\right].$$

Furthermore, by **H3**, we have:

$$\begin{aligned}|g_h(t,\phi) - g(t,\phi)| &= |H \circ f_h(t,\phi) - H \circ f(t,\phi)| \\ &\leq L_H \sup_{t,\phi} |f_h(t,\phi) - f(t,\phi)| \leq L_H C h^p.\end{aligned}$$

By **H1** and the assumptions on the proposal densities, (t,ϕ) remains a.s. uniformly in the compact set $[t_0,T] \times K_1$. Thus there exist M and, M_h such that

$$M = \sup_{(t,\phi)\in[t_0,T]\times K_1} |g(t,\phi)|, \qquad M_h = \sup_{(t,\phi)\in[t_0,T]\times K_1} |g_h(t,\phi)|,$$

and, we can prove that $M_h \leq M + Ch^p$. Hence, there exists A_h such that

$$\left|\left(y_j - g_h(t_j,\phi)\right)^2 - \left(y_j - g(t_j,\phi)\right)^2\right| \leq A_h C h^p$$

and, we obtain a ϕ-independent bound:

$$\left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| \leq \frac{1}{(2\pi\sigma^2)^{J/2}}(e^{\frac{1}{2\sigma^2}JA_hCh^p} - 1) \leq \frac{1}{(2\pi\sigma^2)^{J/2}}\left(e^{BA_hh^p} - 1\right),$$

with $B = \frac{JC}{2\sigma^2}$. Consequently, by integrating the previous inequality,

$$|p_Y(y) - p_{h,Y}(y)| \leq \int \left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| \pi(\phi)d\phi \leq \frac{1}{(2\pi\sigma^2)^{J/2}}(e^{BA_hh^p} - 1).$$

Finally, using the previous inequalities, we have:

$$\int \left|p_{\phi|Y}(\phi) - p_{h,\phi|Y}(\phi)\right| d\phi \leq \frac{2}{p_Y(y)(2\pi\sigma^2)^{J/2}}(e^{BA_hh^p} - 1).$$

Let $D(\cdot,\cdot)$ denote the total variation distance. If h is small enough, there exists a h-independent constant C_y such that:

$$D(p_{\phi|Y}, p_{h,\phi|Y}) \leq C_y h^p.$$

(2) (a) Following Tierney (1994), this algorithm provides a uniformly ergodic Markov chain. We quote $p^{(2)}_{h,\phi}$ its stationary distribution.

(b) At step $r+1$, let $\phi^{(r)}$ be the current value and ϕ^c be the candidate of the Markov chain. By definition of the LL and LL2 schemes, we have

$$\left|f_h(t,\phi^c) - f_{h,\phi^{(r)}}(t,\phi^c)\right| = O(\|\phi^c - \phi^{(r)}\|_{\mathbb{R}^k}). \tag{A.1}$$

Moreover, by definition of the acceptance probability,

$$|\rho(\phi^{(r)},\phi^c)-\rho^{(2)}(\phi^{(r)},\phi^c)| \leq \frac{\pi(\phi^c)}{p_{h,Y|\phi}(\phi^{(r)})\pi(\phi^{(r)})}\left|p_{h,Y|\phi}(\phi^c)-p^{(2)}_{h,Y|\phi}(\phi^c)\right|$$

Let quote

$$M_{LL} = \sup_{(t,\phi)\in[t_0,T]\times K_1} |H\circ f_h(t,\phi)| \qquad \text{and} \qquad M_{LL2} = \sup_{(t,\phi)\in[t_0,T]\times K_1} |H\circ f_{h,\phi^{(r)}}(t,\phi)|.$$

These quantities exist as ϕ remains in the compact set K_1 by definition of the proposal density and, because $\phi \to f_0(\phi)$ is $\mathcal{C}^1$. According to (A.1), there exists M such that:

$$\begin{aligned}&\left|(y_j - H\circ f_h(t_j,\phi^c))^2 - \left(y_j - H\circ f_{h,\phi^{(r)}}(t_j,\phi^c)\right)^2\right| \\ &\leq \underbrace{\left(2\max|y_j| + M_{LL} + M_{LL2}\right) L_H M}_{\equiv A} \underbrace{\|\phi^c-\phi^{(r)}\|_{\mathbb{R}^k}}_{\leq\eta \text{ by property (2)}}.\end{aligned}$$

Consequently we have

$$\left|p_{h,Y|\phi}(\phi^c)-p^{(2)}_{h,Y|\phi}(\phi^c)\right| \leq \frac{1}{(2\pi\sigma^2)^{J/2}}\left(e^{\frac{1}{2\sigma^2}JA\eta}-1\right).$$

As $p_{h,Y|\phi}$ is continuous in ϕ, there exists a constant c such that

$$\inf_{\phi\in K_1} p_{h,Y|\phi}(\phi) \geq \frac{c}{(2\pi\sigma^2)^{J/2}}.$$

By **H1**, and combining the previous inequalities, we obtain

$$|\rho(\phi^{(r)},\phi^c)-\rho^{(2)}(\phi^{(r)},\phi^c)| \leq \frac{b}{ca}\left(e^{\frac{1}{2\sigma^2}JA\eta}-1\right) \leq B\eta, \qquad \text{(A.2)}$$

for a η small enough. The transition kernel of the Markov chains simulated by the standard and, modified Hastings-Metropolis algorithms are quoted $\mathcal{K}$ and $\mathcal{K}^{(2)}$ respectively. Using the previous inequality, we have:

$$\begin{aligned}\left|\mathcal{K}(\phi;\{\phi^c\}) - \mathcal{K}^{(2)}(\phi;\{\phi^c\})\right| &= \left|\rho(\phi,\phi^c)-\rho^{(2)}(\phi,\phi^c)\right| q(\phi^c|\phi) \\ &\leq B\eta\ \ q(\phi^c|\phi)\end{aligned}$$

As a consequence, we have

$$\sup_{\phi\in K_1} D\left(\mathcal{K}(\phi;\cdot),\mathcal{K}^{(2)}(\phi,\cdot)\right) \leq B\eta$$

The result follows from the part 1 of this theorem, applied for the Local Linearization scheme, combined with the following lemma.

Lemma 6 *Let $\mathcal{K}$ and $\mathcal{K}^{(2)}$ be the respective transition kernels of two Markov chains defined on a space $\mathcal{E}$ and let p and $p^{(2)}$ be their respective stationary distributions. Assume that $\mathcal{K}$ supplies a uniformly ergodic chain and that there exists a constant C such that*

$$\sup_{\phi\in\mathcal{E}} D\left(\mathcal{K}(\phi;\cdot),\mathcal{K}^{(2)}(\phi,\cdot)\right) \leq C.$$

Then, there exists a constant α such that:

$$D\left(p,p^{(2)}\right) \leq \alpha C.$$

This lemma directly derives from the Poisson equality:

$$p\cdot f - p^{(2)}\cdot f = p^{(2)}(\mathcal{K}^{(2)}-\mathcal{K})\cdot Vf$$

where $Vf(x) = \sum_{n=0}^{\infty}\left(\mathcal{K}^n f(x) - p\cdot f\right)$ and $p\cdot f = \int f(x)p(dx)$ for any measurable function f.

□

B The modified Local Linearization scheme

The LL scheme is based on a local linearization of the second member of ODE (1) with respect to time t and f. The new LL2 scheme is deduced using a Taylor expansion of the right term with respect to t, f and ϕ.

B.1 Principle

Let the equation (1) be solved by the LL scheme at a given ϕ_0. Let ϕ be in a bounded neighborhood of ϕ_0. Let the time interval $[t_0,T]$ be divided in N sub-intervals $[t_n,t_{n+1}[$, $t_n = t_0+nh$, $n=0,\ldots,N-1$, where h is the step size of the method. On each time interval $[t_n,t_{n+1}[$, the linearized equation deriving from the equation (1) at $(t,\,\phi)$ with $t\in[t_n,t_{n+1}[$ is:

$$F(f(t,\phi),t,\phi) \simeq F(f(t_n,\phi_0),t_n,\phi_0) + \frac{dF}{df}\left(f(t_n,\phi_0),t_n,\phi_0\right)\left(f(t,\phi)-f(t_n,\phi_0)\right)$$
$$+\frac{dF}{dt}\left(f(t_n,\phi_0),t_n,\phi_0\right)(t-t_n) + \frac{dF}{d\phi}\left(f(t_n,\phi_0),t_n,\phi_0\right)(\phi-\phi_0). \qquad \text{(B.1)}$$

The following notations are introduced:

$$\begin{cases} f_n(\phi) & = f_h(t_n, \phi) \\ F_n(\phi) & = F(f_n(\phi), t_n, \phi) \\ DF_n(\phi) & = \frac{dF}{df}(f_n(\phi), t_n, \phi) \\ F'_n(\phi) & = \frac{dF}{dt}(f_n(\phi), t_n, \phi) \\ R_k(X, h) & = \int_0^h \exp(uX) u^k du \\ D_\phi F(\phi_0) & = \frac{dF}{d\phi}(f_n(\phi_0), t_n, \phi). \end{cases}$$

The LL2 scheme is:

$$f_{n+1}(\phi) = f_n(\phi_0) + \Lambda_n^m(\phi_0, \phi, h) \tag{B.2}$$

with

$$\begin{aligned} &\Lambda_n^m(\phi_0, \phi, h) = [hR_0(DF_n(\phi_0), h) - R_1(DF_n(\phi_0), h)] F'_n(\phi_0) \\ &+R_0(DF_n(\phi_0), h)(F_n(\phi_0) + D_\phi F(\phi_0)(\phi - \phi_0)) + \exp(hDF_n(\phi_0))(f_n(\phi) - f_n(\phi_0)). \end{aligned}$$

Remark 7 *(1) The previous recursive formula taken at $\phi = \phi_0$ is the same as the one deriving from the LL scheme.*
(2) For autonomous dynamic system, the scheme (B.2) is simplified by:

$$f_{n+1}(\phi) = f_n(\phi_0) + \lambda_n^m(\phi_0, \phi, h),$$

with

$$\begin{aligned} \lambda_n^m(\phi_0, \phi, h) = &R_0(DF_n(\phi_0), h)(F_n(\phi_0) + D_\phi F(\phi_0)(\phi - \phi_0)) \\ &+ \exp(hDF_n(\phi_0))(f_n(\phi) - f_n(\phi_0)). \end{aligned}$$

B.2 Convergence of the method

Biscay et al. (1996) and Ramos and García-López (1997) prove that the LL scheme is of convergence rate h^2. We extend their results to the LL2 scheme:

Lemma 8 (Error estimation) *Let f be the exact solution of ODE (1) and f_{h,ϕ_0} be the appoximate solution obtained by the LL2 scheme with the step size h, given a point ϕ_0. Under assumptions **H1′** and **H2**, there exist two constants C_1 and C_2, ϕ and ϕ_0- independent such that for any $t \in [t_0, T]$ and for any ϕ,*

$$|f(t, \phi) - f_{h,\phi_0}(t, \phi)| \leq \max(C_1 h^2, C_2 \|\phi - \phi_0\|^2_{\mathbb{R}^k}).$$

PROOF. The proof, essentially the same as the Ramos and García-López (1997)'s one, is presented for $d = 1$ but is easily generalizable for any d.

On the interval $[t_n, t_{n+1}[$, $0 \leq n \leq N-1$, f_{h,ϕ_0} is the exact solution of the following linear equation:

$$\begin{cases} \frac{\partial f_{h,\phi_0}(t,\phi)}{\partial t} & = F_{h,\phi_0}(f_{h,\phi_0}(t,\phi), t, \phi) \\ f_{h,\phi_0}(t_0,\phi) & = f_0(\phi), \end{cases}$$

where F_{h,ϕ_0} is the Taylor expansion of F at point (t_n, ϕ_0) defined by the equation (B.1). For any t in $[t_n, t_{n+1}[$, we have:

$$|f(t,\phi) - f_{h,\phi_0}(t,\phi)| = \left| \int_{t_n}^{t} F(f(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi) du + f(t_n,\phi) - f_{h,\phi_0}(t_n,\phi) \right|$$

Assuming **H2**, F is $\mathcal{C}^2$ on its domain of definition. Since F_{h,ϕ_0} is the second order Taylor expansion of F at (t_n, ϕ_0), there exists $\xi \in \mathbb{R}^d \times (t_n, t_{n+1}] \times \mathbb{R}^k$ such that:

$$\begin{aligned} |F(f_{h,\phi_0}(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi)| &\leq \frac{1}{2} \left| \frac{\partial^2 F}{\partial f^2}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))^2 \right| \\ + \frac{1}{2} \left| \frac{\partial^2 F}{\partial t^2}(\xi,\phi_0)(u-t_n)^2 \right| &+ \frac{1}{2} \left| \frac{\partial^2 F}{\partial f \partial t}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))(u-t_n) \right| \\ + \frac{1}{2} \left| \sum_{m=1}^{k} \frac{\partial^2 F}{\partial f \partial \phi_m}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))(\phi_m - \phi_{0_m}) \right| &+ \frac{1}{2} \left| \sum_{m=1}^{k} \frac{\partial^2 F}{\partial t \partial \phi_m}(\xi,\phi_0)(u-t_n)(\phi_m - \phi_{0_m}) \right| \\ + \frac{1}{2} \left| \sum_{m',m=1}^{k} \frac{\partial^2 F}{\partial \phi_m \partial \phi_{m'}}(\xi,\phi_0)(\phi_m - \phi_{0_m})(\phi_{m'} - \phi_{0_{m'}}) \right|. & \end{aligned} \tag{B.3}$$

Assuming **H1′** and **H2**, there exists a constant c independent of t and ϕ, which upper-bounds every second order differential of F. Moreover, a short recursive argument together with **H1′** implies that f_{h,ϕ_0} is $\mathcal{C}^1$ on $[t_0, T] \times K_1$. Thus, there exist constants η and η' such that

$$|f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0)| \leq \max(\eta|u-t_n|, \eta'\|\phi - \phi_0\|_{\mathbb{R}^k})$$

Hence, using (B.3), there exist two constants A and A' such that:

$$|F(f_{h,\phi_0}(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi)| \leq \max(A|u-t_n|, A'\|\phi - \phi_0\|_{\mathbb{R}^k})^2.$$

The F function is $\mathcal{C}^2$. Thus, by quoting L_F its Lipschitz constant, we can write:

$$|F(f(u,\phi),u,\phi) - F(f_{h,\phi_0}(u,\phi),u,\phi)| \leq L_F \|f(u,\phi) - f_{h,\phi_0}(u,\phi)\|$$

Finally, by quoting $E_\phi(u) = \|f(u,\phi) - f_{h,\phi_0}(u,\phi)\|$, and combining the previous inequalities, we have:

$$E_\phi(t) \leq \int_{t_n}^{t} L_F E_\phi(u)du + \int_{t_n}^{t} \max(A|u-t_n|, A'\|\phi-\phi_0\|_{\mathbb{R}^k})^2 du + E_\phi(t_n).$$

As $|u - t_n| \leq |t - t_n| \leq h$, we have:

$$E_\phi(t) \leq \int_{t_n}^{t} L_F E_\phi(u)du + \max(Ah, A'\|\phi-\phi_0\|_{\mathbb{R}^k})^2 (t-t_n) + E_\phi(t_n).$$

The expected result derives from the following lemma and similar arguments as those presented by Ramos and García-López (1997):

Lemma 9 *Let u be a positive function such that*

$$u(t) \leq a + b(t-t_0) + c\int_{t_0}^{t} u(s)ds$$

Then we have

$$u(t) \leq ae^{c(t-t_0)} + \frac{b}{c}(e^{c(t-t_0)} - 1).$$

References

Beal, S., Sheiner, L., 1982. Estimating population kinetics. Crit Rev Biomed Eng 8(3), 195–222.

Bennet, J. E., Racine-Poon, A., Wakefield, J. C., 1996. MCMC for nonlinear hierarchical models, Chapman & Hall, London. 339–358.

Biscay, R., Jimenez, J. C., Riera, J. J., Valdes, P. A., 1996. Local linearization method for the numerical solution of stochastic differential equations. Ann. Inst. Statist. Math. 48(4), 631–644.

Buxton, R. B., Wong, E. C., Franck, L. R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. MRM 39, 855–864.

Carlin, B. P., Louis, T. A., 2000. Bayes and empirical Bayes methods for data analysis, vol. 69 of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Celeux, G., Diebolt, J., 1985. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational. Statistics Quaterly 2, 73–82.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27, 94–128.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39(1), 1–38. With discussion.

Donnet, S., Lavielle, M., Poline, J. B., 2005. Are fMRI event related responses constant across events? In review in Neuroimage .

Gelfand, A. E., Smith, A. F. M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85(410), 398–409.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. Markov chain Monte Carlo in practice. Interdisciplinary Statistics. Chapman & Hall, London.

Hairer, E., Nørsett, S., Wanner, G., 1987. Solving Ordinary Differential Equations I, Nonstiff Problems. Springer, Berlin.

Huang, Y., Liu, D., Wu, H., 2004. Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. Technical Report 04/06 .

Jimenez, J. C., 2002. A simple algebraic expression to evaluate the local linearization schemes for stochastic differential equations. Appl. Math. Lett. 15(6), 775–780.

Jimenez, J. C., Biscay, R., Mora, C., Rodriguez, L. M., 2002. Dynamic properties of the local linearization method for initial-value problems. Appl. Math. Comput. 126(1), 63–81.

Kuhn, E., Lavielle, M., 2004. Coupling a stochastic approximation version of EM with a MCMC procedure. ESAIM P&S , 115–131.

Louis, T. A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44(2), 226–233.

Ramos, J. I., 1999. Linearized methods for ordinary differential equations. Appl. Math. Comput. 104(2-3), 109–129.

Ramos, J. I., García-López, C. M., 1997. Piecewise-linearized methods for initial-value problems. Appl. Math. Comput. 82(2-3), 273–302.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22(4), 1701–1762. With discussion and a rejoinder by the author.

Tracewell, W., Trump, D., Vaughan, W., Smith, D., Gwilt, P., 1995. Population pharmacokinetics of hydroxyurea in cancer patients. Cancer Chemother. Pharmacol. 35(5), 417–22.

Vonesh, E. F., 1996. A note on the use of Laplace's approximation for nonlinear mixed-effects models. Biometrika 83(2), 447–452.

Wakefield, J., Smith, A., Racine-Poon, A., Gelfand, A., 1994. Bayesian analysis of linear and non-linear population models using the gibbs sampler. Appl. Statist 43, 201–221.

Wei, G. C. G., Tanner, M. A., 1990. Calculating the content and boundary of the highest posterior density region via data augmentation. Biometrika 77(3), 649–652.

Wu, L., 2004. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. J. Amer. Statist. Assoc. 99(467), 700–709.

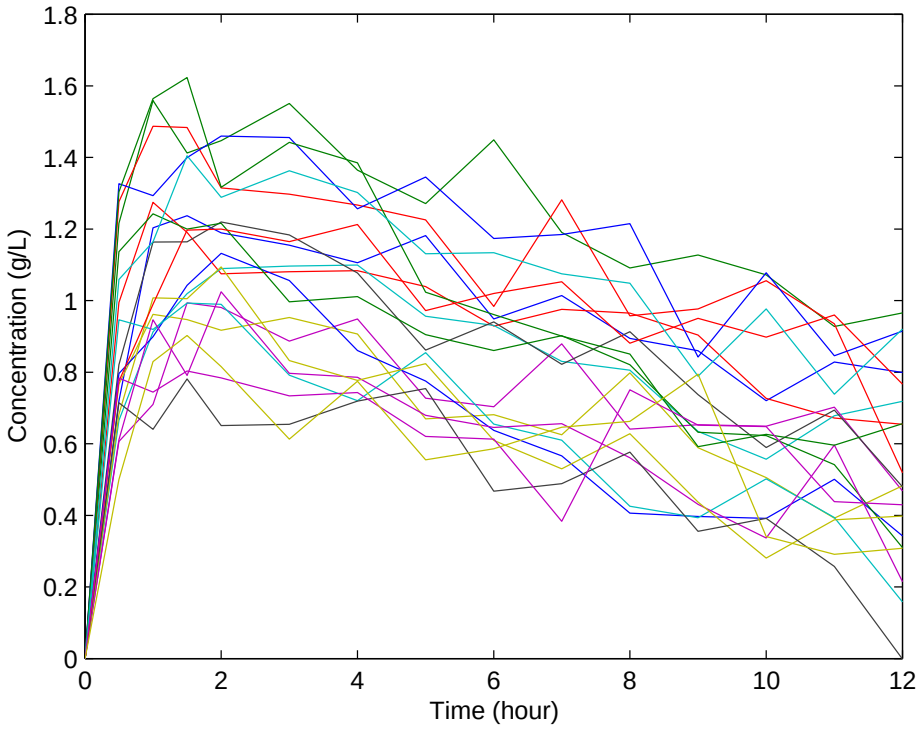


Fig. B.1. Individual concentrations of pharmacokinetic hydroxurea simulated for 20 patients.

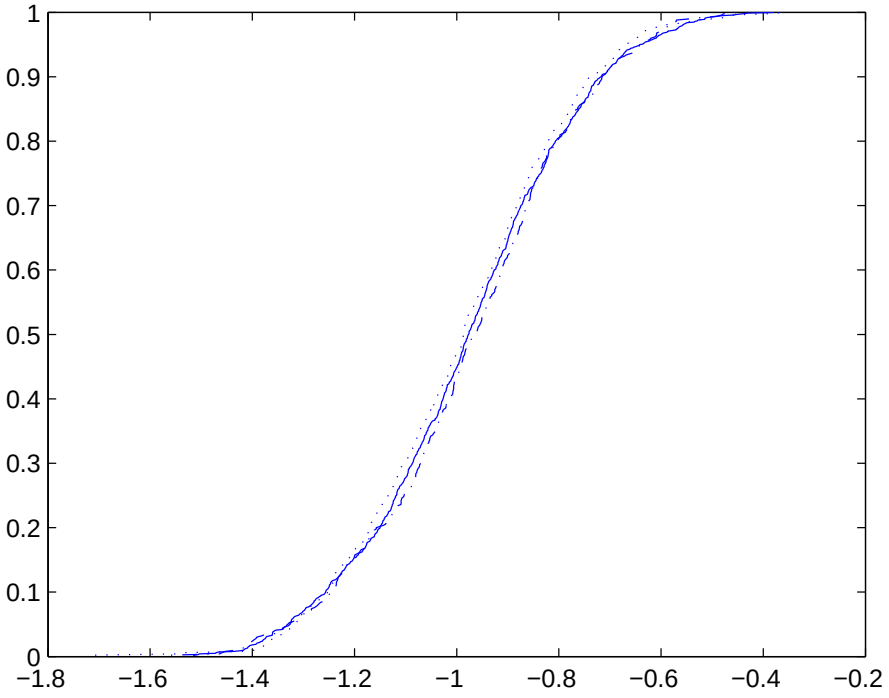


Fig. B.2. Empirical cumulative distribution functions of the k_m conditional distribution simulated by Hastings-Metropolis using a very precise Runge-Kutta solving scheme (plain line), a classical Runge-Kutta scheme (dotted line), or the LL and LL2 schemes (dashed line).

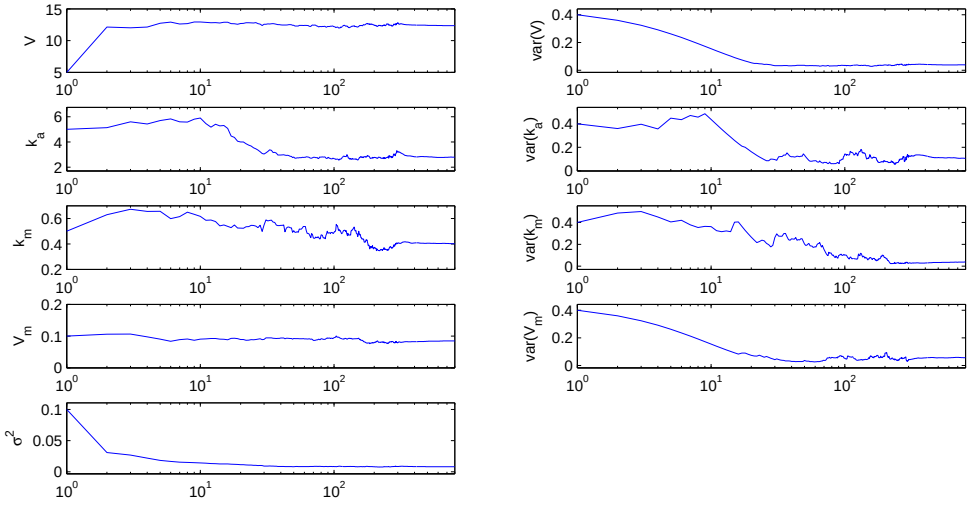


Fig. B.3. Evolution of the estimates in function of the iteration of SAEM algorithm (with a logarithm scale for the abscis axis).

Table B.1
Parameter estimates obtained by the SAEM and NONMEM algorithms on the simulated dataset.

	V	k_a	k_m	V_m	var V	var k_a	var k_m	var V_m	σ^2
initial values	5.0	5.00	0.50	0.100	0.400	0.400	0.400	0.400	0.1000
simulation values	12.2	2.72	0.37	0.082	0.040	0.040	0.040	0.040	0.0100
SAEM									
estimates	12.3	2.79	0.40	0.085	0.038	0.049	0.039	0.038	0.0081
SE	0.5	0.21	0.01	0.004	0.012	0.019	0.013	0.013	6.10^{-4}
NONMEM									
estimates	12.3	2.57	0.60	0.100	0.036	0.068	10^{-8}	0.062	0.0088
SE	-	-	-	-	-	-	-	-	-

6.2 Modèles définis par équations différentielles stochastiques

De nombreux systèmes différentiels ont été proposés pour décrire des processus biologiques. Cependant, dans certains cas, les versions déterministes de ces systèmes sont inadéquates et trop rigides par rapport aux perturbations observées dans la réalité. La modélisation de ces perturbations peut être réalisée en introduisant une variabilité supplémentaire directement dans le système différentiel, aboutissant ainsi à des systèmes différentiels stochastiques. Par exemple Tan et Wu [67] ont proposé une modélisation stochastique de la dynamique simultanée des virus et des $CD4^+$ dans l'infection par le VIH. Ces modèles permettent en particulier de représenter les erreurs résiduelles corrélées dans le temps, dues par exemple à une mauvaise spécification du modèle, à des erreurs sur les temps de prélèvements, les doses, etc.

Dans cet article, nous avons donc considéré des processus de diffusion (solutions d'équations différentielles stochastiques) observés à temps discrets avec bruit de mesure, et dont les paramètres de la fonction de dérive sont aléatoires. Nous avons développé deux méthodes d'estimation, par maximum de vraisemblance (algorithme SAEM) et par approche bayésienne (échantillonneur de Gibbs) adaptées à ces modèles. La méthode d'Euler-Maruyama est utilisée pour approximer le processus de diffusion. Nous avons montré la convergence de ces algorithmes et borné l'erreur due à l'approximation d'Euler-Maruyama en fonction du pas de la discrétisation.

La précision de la méthode d'estimation par maximum de vraisemblance SAEM est illustrée par une étude sur données simulées à partir d'un modèle différentiel à une dimension de pharmacocinétique. Étant donnée la complexité de ces modèles, il n'était pas raisonnable d'espérer évaluer les propriétés de cet algorithme sur un système différentiel de dimension plus élevée. L'analyse des données réelles de la pharmacocinétique de la théophylline (médicament anti-asthmatique bronchodilatateur) illustre la pertinence de l'approche stochastique par rapport à l'approche déterministe, les courbes individuelles étant mieux prédites par le modèle différentiel stochastique.

Ce travail est détaillé dans un article soumis à *Scandinavian Journal of Statistics*.

L'extension de l'algorithme SAEM à l'estimation de systèmes différentiels stochastiques suivait naturellement le travail réalisé dans la section 6.1. Cependant, l'utilisation de tels modèles en pratique reste délicate, leur interprétation clinique étant complexe. Des études supplémentaires sur d'autres jeux de données sont encore nécessaires.

Parametric inference for diffusion processes from discrete-time and noisy observations

Running headline : Estimation for noisy diffusion processes

Sophie Donnet[1] and Adeline Samson[2]

[1] Paris-Sud University, Laboratoire de Mathématiques, Orsay, France
[2] INSERM U738, Paris, France; University Paris 7, Paris, France

Abstract

Noisy discretely observed diffusion processes with random drift function parameters are considered. Maximum likelihood and Bayesian estimation methods are extended to this model, respectively the Stochastic Approximation EM and the Gibbs sampler algorithms. They are based on the Euler-Maruyama approximation of the diffusion, achieved using latent auxiliary data introduced to complete the diffusion process between each pair of measurement instants. A tuned hybrid Gibbs algorithm based on conditional Brownian bridges simulations of the unobserved process paths is included in these two algorithms. Their convergence is proved. Errors induced on the likelihood and the posterior distribution by the Euler-Maruyama approximation are bounded as a function of the step size of the approximation. Results of a pharmacokinetic mixed model simulation study illustrate the accuracy of the maximum likelihood estimation method. The analysis of the Theophyllin real dataset illustrates the relevance of the SDE approach relative to the deterministic approach.

Key words: Bayesian estimation, Brownian bridge, Diffusion process, Euler-Maruyama approximation, Gibbs algorithm, Incomplete data model, Maximum likelihood estimation, Nonlinear mixed effects model, SAEM algorithm

1 Introduction

Time-dependent dynamic processes that follow the laws of finance, physics, physiology or biology are usually described by differential systems. For example, stock price dynamics or short-term interest rates can be described using a wide class of financial differential systems. As another example, in biology, pharmacokinetics consists in the study of the evolution of a drug in an organism. It is described through dynamic systems, the human body being assimilated to a set of compartments within which the drug flows. In these contexts, diffusion models described by stochastic differential equations (SDEs) are natural extensions to the corresponding deterministic models (defined by ordinary differential equations, ODEs) to account for time-dependent or serial correlated residual errors and to handle real life variations in model parameters occurring over time. This variability in the model parameters is most often not predictable, not fully understood or too complex to be modeled deterministically. Thus the SDEs consider errors associated with misspecifications and approximations in the dynamic system.

The parametric estimation of such diffusion processes is a key issue. Estimation of continuously observed diffusion processes is widely studied (see for instance Kutoyants, 1984; Prakasa Rao, 1999). However, for obvious practical purposes, real longitudinal data are always gathered

at discrete points in time (for example stock prices collected once a day, drug concentration measured every hour in patient blood, etc.). Within this framework, statistical inference of discretely observed diffusion processes is a critical question for both maximum likelihood and Bayesian approaches. When the transition probability of the diffusion process is explicitly known, Dacunha-Castelle and Florens-Zmirou (1986) propose a consistent maximum likelihood estimator. Classical Bayesian algorithm such as Gibbs sampling can also be directly applied in this particular case.

However, this transition density has generally no closed form and the estimation methods have to sidestep this difficulty. A short summary of such estimation methods is provided below (see Prakasa Rao, 1999; Sørensen, 2004, for complete reviews). Analytical methods include those of Bibby and Sørensen (1995), Sørensen (2000) – using estimating functions –, Poulsen (1999) – using a numerical solution of the Kolmogorov equation – or Aït-Sahalia (2002) – based on an analytical non-Gaussian approximation of the likelihood function. Other methods approximate the transition density via simulation. They consider the unobserved paths as missing data and introduce a set of auxiliary latent data points between every pair of observations. Along these auxiliary latent data points, the process can be finely sampled using the Gaussian Euler-Maruyama approximation to evaluate the likelihood function via numerical integration as proposed by Pedersen (1995) and Elerian *et al.* (2001), or to evaluate the posterior distribution in a Bayesian analysis again via numerical integration, as discussed by Eraker (2001) and Roberts and Stramer (2001). In this context and for both maximum likelihood and Bayesian estimations, standard Markov Chain Monte-Carlo (MCMC) methods are used to sample the process with the conditional distributions. However, the convergence rate of these estimation methods decreases with the increase in number of latent data points. Different solutions are proposed to overcome this difficulty: Eraker (2001) suggests the sampling of only one element at a time, while Elerian *et al.* (2001) propose to sample block-wise with an importance sampling algorithm. Roberts and Stramer (2001) take a slightly different approach as they sample transformations of the diffusion process. To sidestep the Euler-Maruyama approximation, Beskos *et al.* (2005) develop an exact simulation method of the diffusion process, applicable even without any analytical form of the transition density. This algorithm can be included in a Monte-Carlo procedure to approximate the likelihood function for a classical estimation and in a Gibbs algorithm for a Bayesian inference. However, this exact simulation method is only adapted for time-homogeneous SDEs, which is frequently not the case when studying biological dynamical systems for example. Furthermore, even under the conditions defined by Beskos *et al.* (2005), this exact method requires the inclusion of accept-reject algorithms, which are difficult to implement in the general case of non-linear SDEs and often require a large com-

putational time. Therefore an Euler-Maruyama approximation approach is considered in this paper.

The above-cited papers do not take into account the observation noise on the collected data, which is non-realistic in many cases. For example in the financial context, the daily evolution of an asset price depends on the price fluctuations within each business day. In the biological context, endpoints such as drug concentrations are generally measured with a certain variability due to experimental limits. To reflect this observation noise, we consider the following regression statistical model $\mathcal{M}$: the observed data $y = (y_1, \ldots, y_J)$ are a realization of a random variable Y deduced from a scalar diffusion process Z, as stated by the following equation:

$$Y_j = Z(t_j) + \varepsilon_j, \qquad (\mathcal{M})$$

where $(\varepsilon_j)_{j=1,\ldots,J}$ is a sequence of i.i.d Gaussian random variables of variance σ^2, representing the measurement errors. The diffusion process Z is defined as the solution of the SDE describing the observed dynamic process:

$$dZ(t) = F(Z, t, \phi)dt + \gamma dB(t),$$

driven by a Brownian motion $\{B_t, t_0 \leq t \leq T\}$, a drift function F depending on a parameter ϕ and a volatility coefficient γ. If the volatility coefficient γ is null, the SDE is an ODE, the SDE model parameter ϕ being evidently equivalent to the parameter of the corresponding ODE system, and therefore being interpreted in the same way. In such models, two fundamentally different types of noise have to be distinguished: the dynamic noise γ, reflecting the real random fluctuations around the corresponding theoretical dynamic model, and the measurement noise σ representing the uncorrelated part of the residual variability associated with assay, dosing and sampling errors, for instance, in a biological context. The problem of the parameter estimation of discretely observed diffusion processes with additive measurement noise is evoked in few papers and is not completely solved. In the particular case of linear SDEs, the Kalman filter (Schweppe, 1965) or the EM algorithms (Singer, 1993) can be used. When the observed process is a Gaussian martingale, Jensen and Petersen (1999) and Gloter and Jacod (2001) exhibit estimators and study their theoretical properties. Unfortunately, these explicit forms of maximum likelihood estimates are limited to the linear SDEs case.

In this paper, we assume in addition that the parameter ϕ is a realization of a random variable Φ distributed with a probability π depending on a parameter β. This is the case for example in drug pharmacokinetics studies of which use will be detailed below. Basically, in order to estimate drug pharmacokinetic parameters, the drug concentration is sampled repeatedly among several

individuals, the parameter ϕ being assumed different between the subjects and thus considered as individual non-observed random data.

The main objective of this paper is to develop methods to estimate the parameters vector $\theta = (\beta, \gamma^2, \sigma^2)$ in the general case of non-linear SDEs. Such a method is proposed by Overgaard *et al.* (2005) and Tornøe *et al.* (2005) in the particular case of non-linear mixed effects models. They combine an extended Kalman filter of the diffusion process with an approximated maximum likelihood estimation algorithm based on a linearization of the model. However, the convergence properties of this estimation algorithm based on linearization are not proved. A different point of view can be taken for the parameters estimation, the random quantities Z and Φ being considered as non-observed random data. In that case, the model $\mathcal{M}$ belongs to the framework of incomplete data models, for which several estimation methods are developed for both classical and Bayesian approaches. For classical inference, the Expectation-Maximization (EM) algorithm proposed by Dempster *et al.* (1977) is a broadly applied approach taking advantage of the incomplete data model structure. When the E-step has no closed form, Celeux and Diebolt (1985), Wei and Tanner (1990) and Delyon *et al.* (1999) propose different stochastic versions of this algorithm. These methods require the simulation of the non-observed data using Markov Chain Monte-Carlo (MCMC) algorithms, as proposed by Kuhn and Lavielle (2004). For the Bayesian approach, tuned Gibbs algorithms are developed to estimate the posterior distribution $p_{\theta|Y}(\cdot|y)$ of θ, a specified prior distribution $p_\theta(\cdot)$ for θ being given. When the simulation under the posterior distribution cannot be done in a closed form, hybrid Gibbs sampling algorithms are proposed in the literature, including Metropolis-Hastings procedures (Wakefield *et al.*, 1994; Bennet *et al.*, 1996). To our knowledge, these estimation methods are not yet extended to noisy discretely observed diffusion processes models considered in this paper.

Our objective is thus to propose efficient estimation methods of the vector of parameters θ for the model $\mathcal{M}$, together with theoretical convergence results for both classical and Bayesian inference. We consider an approximate statistical model, of which the regression term is the Euler-Maruyama discretized approximate diffusion process of the SDE. The parameter inference is then performed on this new model, using a stochastic version of the EM algorithm for the classical estimation approach, or using a hybrid version of the Gibbs sampling algorithm for the Bayesian approach.

Section 2 describes the setup of the problem which is considered in this paper, detailing the diffusion process and its Euler-Maruyama approximation. The estimation algorithms for the maximum likelihood and the Bayesian approaches are respectively presented in Sections 3 and 4. These sections detail a tuned MCMC procedure supplying both theoretical and computational

convergence properties to these algorithms. The error on the estimation induced by the Euler-Maruyama scheme is quantified in Section 5. In Section 6, the maximum likelihood algorithm is applied to a non-linear mixed effects model issued from pharmacokinetics. Section 7 concludes with some discussion.

2 Data and Model

2.1 Incomplete data model defined by SDEs

Let $y = (y_j)_{j=0..J}$ denote the vector of the observations measured at times $t_0 \leq t_1 \leq \ldots \leq t_J \leq T$. We consider that y is a realization of the random variable Y defined through the following statistical model $\mathcal{M}$:

$$\left.\begin{array}{rcll} Y_j & = & Z(t_j) + \varepsilon_j, & 0 \leq j \leq J \\ \varepsilon_j & \sim_{i.i.d} & \mathcal{N}(0, \sigma^2), & \\ dZ(t) & = & F(Z, t, \Phi)\, dt + \gamma dB(t)\ ,\ Z(t_0, \Phi) = Z_0(\Phi), & (2.1) \\ \Phi & \sim & \pi(\cdot, \beta) & \end{array}\right\} \quad (\mathcal{M})$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_J)$ represents the measurement error, with a residual variance σ^2. The regression term is a realization of the diffusion process $Z : \mathbb{R} \longrightarrow \mathbb{R}$ defined by the equation (2.1), with B a one-dimensional Brownian motion, γ is the volatility coefficient and the function $F : \mathbb{R} \times [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is the known measurable drift function, non-linearly depending on the non-observed parameter $\Phi \in \mathbb{R}^d$. We assume that Φ is a random variable, distributed with the density π, depending on the parameter $\beta \in \mathbb{R}^p$. The initial condition Z_0 of this process is a deterministic known function of the random parameter Φ (this deterministic function can be a constant).

Our objective is to propose a classical and a Bayesian estimation methods of the parameters vector θ, where $\theta = (\beta, \gamma^2, \sigma^2)$ belongs to some open subset Θ of the Euclidean space $\mathbb{R}^{p+2}$. As the random parameter Φ and the random trajectory Z are not observed, this statistical problem can be viewed as an incomplete data model. The observable vector Y is thus consider as part of a so-called complete vector (Y, Z, Φ).

***Remark* 1** • *This work can be extended to a statistical model with a regression function being equal to $g(Z(t))$, with g a linear or non-linear function, i.e.*

$$Y_j = g(Z(t_j)) + \varepsilon_j, \qquad 0 \leq j \leq J.$$

However, for the simplicity's sake, we only consider the case $g(Z(t)) = Z(t)$ in this paper.

- *The identifiability of this model is a complex problem which is beyond the scope of this paper. However, for simple examples such as linear SDEs the parameters identifiability can be proved.*

2.2 Diffusion model

The diffusion process (2.1) is defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. Statistical inference makes sense only if the existence and uniqueness of a solution of the SDE (2.1) for all $Z(t_0)$, Φ and γ is ensured. Sufficient conditions of existence and uniqueness are the following globally Lipschitz, linear growth and boundedness conditions:

Assumption **(A0)**:

1. For all $\phi \in \mathbb{R}^d$, for all $0 < R < \infty$, there exists $0 < K_R < \infty$ such that for all $t_0 \leq t \leq T$, for all $x, x' \in \mathbb{R}$ with $|x| \leq R$, $|x'| \leq R$

$$|F(x, t, \phi) - F(x', t, \phi)| \leq K_R |x - x'|.$$

2. For all $\phi \in \mathbb{R}^d$, for all $0 < T < \infty$, there exists a constant $0 < C_T < \infty$ such that for all $t_0 \leq t \leq T$, for all $x \in \mathbb{R}$

$$\gamma + |F(x, t, \phi)| \leq C_T(1 + |x|).$$

Under this assumption, for any $t_0 < t < T$, the distribution of $Z(t)$ conditioned by the filtration $\mathcal{F}_{t-}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ ($\mathcal{F}_{t-}$ being the filtration generated by $\{Z(s), s < t\}$). This distribution is denoted $p_{Z|\Phi}(\cdot|\phi; \gamma^2)$ in the following. As a consequence, both Y and (Y, Z, Φ) have density functions, denoted respectively $p_Y(y; \theta)$ and $p_{Y,Z,\Phi}(y, z, \phi; \theta)$ depending on the parameter θ.

2.3 Introduction of an approximate statistical model

For common SDEs, the diffusion density $p_{Z|\Phi}$ has generally no closed form. Consequently neither the likelihood of the observed data $p_Y(y; \theta)$ nor the likelihood of the complete data $p_{Y,Z,\Phi}(y, z, \phi; \theta)$ have analytical forms, which further complicates the parameters estimation. To overcome this difficulty, an approximate statistical model, based on the Euler-Maruyama approximation of the diffusion process is introduced.

2.3.1 Euler-Maruyama approximation of the diffusion process

The Euler-Maruyama scheme is one of the simplest discrete-time approximation of a diffusion process leading to Gaussian approximations of the transition densities. If the time intervals

between the observation instants are too great to obtain a good approximation of the transition density, a natural approach is to introduce a set of auxiliary latent data points between every pair of observations, as first proposed by Pedersen (1995). Let $t_0 = \tau_0 < \tau_1 < \ldots < \tau_n < \ldots < \tau_N = t_J$ denote the deduced discretization of the time interval $[t_0, t_J]$. Let us assume that, for all $j = 0 \ldots J$, there exists an integer n_j verifying $t_j = \tau_{n_j}$, with $n_0 = 0$ by definition. Let $(h_n)_{1 \leq n \leq N}$ be the sequence of the step sizes defined as $h_n = \tau_n - \tau_{n-1}$. Let $h = \max_{1 \leq n \leq N} h_n$ be the maximal step size.

Then the diffusion process denoted W and supplied by the Euler-Maruyama approximation of the SDE is described by the following iterative scheme: for a fixed ϕ, $W_0 = Z_0(\phi)$, and for $n = 1 \ldots N$,

$$\begin{aligned} h_n &= \tau_n - \tau_{n-1}\,, \\ W_n &= W_{n-1} + h_n\, F(W_{n-1}, \tau_{n-1}, \phi) + \gamma\sqrt{h_n}\,\xi_n\,, \\ \xi_n &\sim_{i.i.d} \mathcal{N}(0,1). \end{aligned}$$

Consequently, $(w_{n_0}, \ldots, w_{n_J})$ is an approximation of the original diffusion process at observations instants $(z(t_0), \ldots, z(t_J))$. In the following, let $w = (w_n)_{n=0 \cdots N}$ denote a realization vector of the process W at the discrete times $(\tau_n)_{n=0 \cdots N}$.

2.3.2 Approximate statistical model

Using this approximation of the diffusion process provided by the Euler-Maruyama scheme of step size h, an approximate statistical model denoted model $\mathcal{M}_h$ is defined as:

$$\left.\begin{array}{rcll} Y_j &=& W_{n_j} + \varepsilon_j, & 0 \leq j \leq J\,, \\ \varepsilon_j &\sim_{i.i.d}& \mathcal{N}(0, \sigma^2)\,, & \\ h_n &=& \tau_n - \tau_{n-1}\,, & \\ W_n &=& W_{n-1} + h_n\, F(W_{n-1}, \tau_{n-1}, \Phi) + \gamma\sqrt{h_n}\,\xi_n\,, & 1 \leq n \leq N, \\ \xi_n &\sim_{i.i.d}& \mathcal{N}(0,1)\,, & \\ \Phi &\sim& \pi(\cdot;\beta)\,, & \end{array}\right\} \quad (\mathcal{M}_h)$$

with $W_0 = Z_0(\Phi)$. On this model $\mathcal{M}_h$, Y results from the partial observation of the complete data (Y, W, Φ) where W is the process at the discrete times $(\tau_n)_{n=0 \cdots N}$.

Remark 2 *In this data augmentation framework, the choice of the discretization grid $(\tau_n)_{0 \leq n \leq N}$ is a central issue to guarantee the fast convergence of the estimation algorithms. Indeed, on the one hand, a small step size h ensures a fine Gaussian diffusion approximation. However, on the other hand, it increases the volume of missing data (W, Φ), which can lead to arbitrarily*

poor convergence properties of the algorithms when the missing data volume widely exceeds the volume of actually observed data Y. Furthermore, the time intervals between two observations can be strongly different. Therefore, for practical purposes and to prevent unbalanced volumes of missing data, we propose to adjust the step sizes for each single time interval.

In the following, the distributions referring to the model $\mathcal{M}_h$ are denoted q while those referring to the model $\mathcal{M}$ are denoted p. On $\mathcal{M}_h$, the observation vector y is distributed with density distribution $q_Y(y;\theta)$, which has no closed form because of the SDE non-linearity with respect to ϕ. But by enriching the observed data with the missing data, and by the Markov property of the diffusion process, the complete data likelihood is analytically known:

$$\begin{aligned} q_{Y,W,\Phi}(y,w,\phi;\theta) &= q_{Y|W}(y|w;\sigma^2) \prod_{n=1}^{N} q_{W|\Phi}(w_n|w_{n-1},\phi;\gamma^2)\ \pi(\phi;\beta) \\ &= q_{Y|W}(y|w;\sigma^2) \prod_{n=1}^{N} d(w_n;\ w_{n-1}+h_n\,F(w_{n-1},\tau_{n-1},\phi),\ \gamma^2 h_n)\ \pi(\phi;\beta), \end{aligned}$$

where $d(.;m,v)$ denotes the Gaussian density with mean m and variance v. As a consequence, the estimation of θ can be performed on the model $\mathcal{M}_h$, using a stochastic version of the EM algorithm for a Maximum Likelihood approach or a Gibbs algorithm for a Bayesian approach.

3 Maximum Likelihood Estimation on the model $\mathcal{M}_h$

In this section we propose a maximum likelihood estimation method, the vector of parameters θ being thus estimated as the maximizing value of the likelihood $q_Y(.\,;\theta)$.

3.1 Stochastic versions of the EM algorithm

The Expectation Maximization (EM) algorithm proposed by Dempster *et al.* (1977) takes advantage of the incomplete data model structure. We consider that the observed data Y are the partial observations of the complete data (Y,X) with X the vector of the non-observed data. The EM algorithm is useful in situations where the direct maximization of $\theta \to q_Y(.\,;\theta)$ is more complex than the maximization of $\theta \to Q(\theta|\theta')$, with:

$$Q(\theta|\theta') = E_{X|Y}\left[\log p_{Y,X}(y,x;\theta)|y;\theta'\right].$$

The EM algorithm is an iterative procedure: at the k-th iteration, the E-step is the evaluation of $Q_k(\theta) = Q(\theta\,|\,\theta_{k-1})$ while the M-step updates θ_{k-1} by maximizing $Q_k(\theta)$. For cases where the E-step has no closed form, Delyon *et al.* (1999) propose the Stochastic Approximation EM

algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q_k(\theta)$. The E-step is thus divided into a simulation step (S-step) of the non-observed data $x^{(k)}$ with the conditional distribution $p_{X|Y}(.\,|y;\theta_{k-1})$ and a stochastic approximation step (SA-step):

$$Q_k(\theta) = Q_{k-1}(\theta) + \alpha_k \left[\log\left(p_{Y,X}(y, x^{(k)}; \theta_{k-1})\right) - Q_{k-1}(\theta)\right],$$

where $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of positive numbers decreasing to zero.

The distribution $p_{X|Y}(.|y;\theta_{k-1})$ is likely to be a complex distribution, as for the model $\mathcal{M}_h$, resulting in the impossibility of a direct simulation of the non-observed data x. For such cases, Kuhn and Lavielle (2004) suggest a MCMC scheme by constructing a Markov chain with an unique stationary distribution $p_{X|Y}(\cdot|y;\theta_{k-1})$ at the k-th iteration. They prove the convergence of the estimates sequence provided by this SAEM algorithm towards a maximum of the likelihood under general conditions and in the case where $p_{Y,X}$ belongs to a regular curved exponential family.

3.2 Extension of the SAEM algorithm to the model $\mathcal{M}_h$

In the particular case of the model $\mathcal{M}_h$, the non-observed data vector is equal to $X = (W, \Phi)$. As the simulation under the conditional distribution $q_{W,\Phi|Y}$ can not be performed directly, the SAEM algorithm combined with a MCMC procedure is applied to the model $\mathcal{M}_h$ to estimate the model parameter θ. To ensure the convergence of the SAEM algorithm, the model $\mathcal{M}_h$ is assumed to fulfill some regular conditions:

Assumption (**A1**):

1. $\pi(.\,;\beta)$ is such that $q_{Y,W,\Phi}$ belongs to the exponential family:

$$\log q_{Y,W,\Phi}(y, w, \phi; \theta) = -\psi(\theta) + \langle S(y, w, \phi), \nu(\theta)\rangle,$$

   where ψ and ν are two functions of θ, $S(y, w, \phi)$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\widetilde{\mathcal{S}}$ of $\mathbb{R}^m$ and $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbb{R}^m$.

2. $\beta \longmapsto \pi(\phi;\beta)$ is of class $\mathcal{C}^m$ for all $\phi \in \mathbb{R}^d$.

Under the assumption (**A1**), the SA-step of the SAEM algorithm reduces to the approximation of $E\left[S(y, w, \phi)|y;\theta'\right]$. The k-th iteration of the SAEM algorithm is thus

- S-Step: a realization of the non-observed data $(w^{(k)}, \phi^{(k)})$ is generated through the succession of M iterations of a MCMC procedure providing an uniformly ergodic Markov chain with $q_{W,\Phi|Y}(\cdot|y;\theta_{k-1})$ as unique stationary distribution,

- SA-Step: s_{k-1} is updated using the following stochastic approximation scheme:

$$s_k = s_{k-1} + \alpha_k(S(y, w^{(k)}, \phi^{(k)}) - s_{k-1}),$$

- M-Step: θ_{k-1} is updated to maximize the complete log-likelihood:

$$\widehat{\theta}_k = \arg\max_{\theta} \left(-\psi(\theta) + \langle s_k, \nu(\theta)\rangle\right).$$

For example, the sufficient statistics corresponding to σ^2 and γ^2 are:

$$\begin{aligned} S^{(1)}(y, w, \phi) &= \frac{1}{J+1}\sum_{j=0}^{J}(y_j - w_{n_j})^2, \\ S^{(2)}(y, w, \phi) &= \frac{1}{N}\sum_{n=1}^{N}\frac{(w_n - h_n F(w_{n-1}, \tau_{n-1}, \phi))^2}{h_n}, \end{aligned}$$

and the M-step for σ^2 and γ^2 at iteration k reduces to $\widehat{\sigma^2}_k = s_k^{(1)}$ and $\widehat{\gamma^2}_k = s_k^{(2)}$. The sufficient statistics for β depend on the distribution $\pi(.\,;\beta)$.

3.3 Convergence of the SAEM algorithm on the model $\mathcal{M}_h$

Let denote Π_θ the transition probability of the Markov chain generated by the MCMC algorithm. Following Kuhn and Lavielle (2004), the convergence of the SAEM algorithm combined with MCMC is ensured under the following additional assumption:

Assumption (**A2**):

1. The chain $(w^{(k)}, \phi^{(k)})_{k\geq 0}$ takes its values in a compact set $\mathcal{E}$ of $\mathbb{R}^N \times \mathbb{R}^d$.

2. For any compact subset V of Θ, there exists a real constant L such that for any (θ, θ') in V^2

$$\sup_{\{(w,\phi),(w',\phi')\}\in\mathcal{E}} |\Pi_\theta(w', \phi'|w, \phi) - \Pi_{\theta'}(w', \phi'|w, \phi)| \leq L\|\theta - \theta'\|_{\mathbb{R}^{p+2}}$$

3. The transition probability Π_θ supplies an uniformly ergodic chain of which invariant probability is the conditional distribution $q_{W,\Phi|Y}(\cdot\,;\theta)$, i.e.

$$\exists K_\theta \in \mathbb{R}^+, \quad \exists \rho_\theta \in]0,1[\quad | \quad \forall k \in \mathbb{N} \quad \|\Pi_\theta^k(\cdot|w,\phi) - q_{W,\Phi|Y}(\cdot;\theta)\|_{TV} \leq K_\theta \rho_\theta^k$$

where $\|\cdot\|_{TV}$ is the total variation norm. Furthermore,

$$K = \sup_{\theta\in\Theta} K_\theta < \infty \quad \text{and} \quad \rho = \sup_{\theta\in\Theta} \rho_\theta < 1$$

4. The function S_h is bounded on $\mathcal{E}$.

Theorem 1 *Let assumptions (**A0-A1-A2**) hold. Let $q_{W,\Phi|Y}$ have finite moments of order 1 and 2. Let (α_k) be a sequence of positive numbers decreasing to 0 such that for all k in $\mathbb{N}$, $\alpha_k \in [0,1]$, $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$.*

Assuming the sequence $(s_k)_{k\geq 1}$ takes its values in a compact set, the sequence $(\widehat{\theta}_k)_{k\geq 1}$ obtained by the SAEM algorithm on the model $\mathcal{M}_h$ converges almost surely towards a (local) maximum of the likelihood q_Y.

Proof: Assuming (**A1-A2**) and the existence of finite moments for $q_{W,\Phi|Y}$, the assumptions of Kuhn and Lavielle (2004) are fulfilled and ensure the convergence of the estimates towards a local maximum of the likelihood function.

Remark 3 *If the compactness on $(s_k)_{k\geq 0}$ is not checked or difficult to check, the algorithm can be stabilized using the method of dynamic bounds proposed by Chen* et al. *(1988) and used by Delyon* et al. *(1999).*

A MCMC procedure fulfilling the assumption (**A2**) is proposed in the following part 3.4.

3.4 Simulation of the non-observed data using a MCMC procedure

At the k-th iteration of the SAEM algorithm, given en estimate $\widehat{\theta}_{k-1}$, a realization of the non-observed data $(w^{(k)}, \phi^{(k)})$ is generated through the succession of M iterations of a MCMC procedure. MCMC procedures construct a Markov chain with $q_{W,\Phi|Y}(.|y\,;\widehat{\theta}_{k-1})$ as the invariant distribution, by proposing candidates (ϕ^c, w^c) with any proposal density Q. However, sampling all the missing data at the same time can lead to poor convergence properties. Therefore, a hybrid Gibbs algorithm is implemented and realized successively M times, the m-th iteration being written as:

1. generation of $\phi^{(m)}$, using a Metropolis-Hastings (M-H) procedure with Q_1 as proposal density and such that $q_{\Phi|Y,W}(.\,|y, w^{(m-1)}; \widehat{\theta}_{k-1})$ is the invariant distribution.

2. generation of $w^{(m)}$, using a M-H procedure with Q_2 as proposal distribution and such that $q_{W|Y,\Phi}(.\,|y, \phi^{(m)}; \widehat{\theta}_{k-1})$ is the invariant distribution.

A careful choice of the proposal densities Q_1 and Q_2 will help the algorithm to quickly explore the parameters space. In the following, some proposal densities of which efficiency is proved on numerical examples are detailed. To simplify the notation, the parameter $\widehat{\theta}_{k-1}$ is omitted since this simulation is performed for a fixed $\widehat{\theta}_{k-1}$.

3.4.1 Proposal distributions

1. Simulation of the candidate ϕ^c can be carried out with the prior density π which allows an efficient exploration of the space of parameters. This leads to an independent M-H algorithm. An alternative consists in generating a candidate in a neighborhood of $\phi^{(m-1)}$, $\phi^c = \phi^{(m-1)} + \eta$ with $\eta \sim \mathcal{N}(0, \delta)$ and where δ is a scaling parameter on which the algorithm convergence depends. This results in the so-called random-walk M-H algorithm (see for example Bennet *et al.*, 1996).

2. A trajectory candidate w^c can be generated using the Euler-Maruyama scheme which corresponds to the prior distribution. An alternative to simulate w^c consists in splitting the vector w into two parts $(w_{n_0}, \cdots, w_{n_J})$ and w_{aux}, the former being the process observed at times $(t_j)_{j=0\cdots J}$ and the latter being the process observed at the auxiliary latent times excluding the observation times. The simulation of $(w^c_{n_0}, \cdots, w^c_{n_J})$ can be performed with random walk distributions: $w^c_{n_j} = w^{(m-1)}_{n_j} + \eta'$ where $\eta' \sim \mathcal{N}(0, \delta')$ and δ' is a scaling parameter chosen to ensure good convergence properties. As proposed by Pedersen (1995), the trajectory at the auxiliary times w^c_{aux} can be generated using an unconditioned distribution but it would have poor convergence properties. A more appropriate strategy consists in generating a candidate w^c_{aux} using Brownian bridges, conditioning the proposed bridge on the events $(w^c_{n_j})_{j=0\cdots J}$, as suggested by Eraker (2001) or Roberts and Stramer (2001). More precisely, for $n_{j-1} < n < n_j$, w_{τ_n} is simulated with:

$$w^c_{\tau_n} = w^c_{n_{j-1}} + \frac{w^c_{n_j} - w^c_{n_{j-1}}}{t_j - t_{j-1}}(\tau_n - t_{j-1}) + \overline{B}_{\tau_n}$$

where $\overline{B}$ is a standard Brownian bridge on $[0, 1]$ equal to zero for $t = 0$ and $t = 1$, which can be easily simulated.

3.4.2 Uniform ergodicity of the MCMC procedure

For checking assumption (**A2**-3), it is possible to verify some minoration condition or Doeblin's condition for the transition probability Π_θ (see Chap. 16 of Meyn and Tweedie, 1993). Otherwise, each case has to be considered individually.

For the independent M-H algorithm, its uniform ergodicity is ensured as soon as the proposal distribution verifies:

$$\exists \lambda \in \mathbb{R}^+ \quad | \quad \forall (w, \phi) \in \mathcal{E}, \qquad Q(w, \phi) \geq \lambda q_{W,\Phi|Y}(w, \phi|y),$$

(see Th 2.1 in Tierney (1994) for more details). The proposal distribution equal to the prior

distribution $Q(w,\phi) = q_{W,\Phi}(w,\phi)$ fulfills this condition.

Moreover, in case of a cyclic combination, the uniform ergodicity of the Markov Chain is ensured if one of the proposal distributions satisfies a minoration condition (Prop. 3 and 4 of Tierney (1994)). Thus the introduction of the prior distribution as proposal distribution is sufficient to ensure the uniform ergodicity of the Markov Chain.

4 Bayesian estimation of the model $\mathcal{M}_h$

For a fully Bayesian treatment of the estimation problem, we shall fix prior distributions on all unknown parameters $(\beta, \gamma^2, \sigma^2)$. We assume that β, γ^2 and σ^2 have continuous prior densities $p_\beta(\cdot)$, $p_\gamma(\cdot)$, $p_\sigma(\cdot)$ respectively on $\mathbb{R}^p$, $\mathbb{R}$ and $\mathbb{R}$. The Bayesian approach consists in the evaluation of the posterior distribution $p_{\theta|Y}$. According to the arguments developed in Section 2, the estimation procedure is applied to the model $\mathcal{M}_h$.

A simplistic approach would be to consider a basic Gibbs algorithm which simulates the non-observed data (ϕ, w) and then updates the parameter θ. However, as emphasized and illustrated by Roberts and Stramer (2001), the quadratic variation of the diffusion process satisfies, for almost surely all observation times t_j and t_{j+1}:

$$\lim_{h\to 0} \sum_{n=n_j}^{n_{j+1}-1} (w_{n+1} - w_n)^2 = (t_{j+1} - t_j)\gamma^2. \tag{4.1}$$

Therefore, conditional on any process satisfying (4.1), the posterior distribution of the volatility $q(\gamma^2|y,\phi,w,\sigma^2,\beta) \propto q_{W|\Phi}(w|\phi;\gamma^2)p_\gamma(\gamma^2)$ is just a point mass at γ^2. Consequently this data augmentation scheme is reducible. Roberts and Stramer (2001) propose a reparameterization to avoid this problem and consider the following transformation:

$$\dot{w}_n = \frac{w_n}{\gamma} \quad \text{and} \quad \dot{F}(x,t,\phi) = \frac{F(\gamma x, t, \phi)}{\gamma}.$$

Consequently, Bayesian inference is performed on the approximate model $\dot{\mathcal{M}}_h$ deduced from the model $\mathcal{M}_h$ using the same reparameterization:

$$\left.\begin{array}{rcll} Y_j & = & \gamma\, \dot{W}_{n_j} + \varepsilon_j, & 0 \le j \le J\,, \\ \varepsilon_j & \sim_{i.i.d} & \mathcal{N}(0,\sigma^2)\,, & \\ h_n & = & \tau_n - \tau_{n-1}\,, & \\ \dot{W}_n & = & \dot{W}_{n-1} + h_n\, \dot{F}(\dot{W}_{n-1}, \tau_{n-1}, \Phi) + \sqrt{h_n}\, \xi_n\,, & 1 \le n \le N, \\ \xi_n & \sim_{i.i.d} & \mathcal{N}(0,1)\,, & \\ \Phi & \sim & \pi(\cdot;\beta)\,. & \end{array}\right\} \quad (\dot{\mathcal{M}}_h)$$

A Gibbs algorithm based on this reparameterization is described below. The posterior distributions can be written as:

- $q_{\Phi,\dot{W}|Y,\theta}(\phi,\dot{w}|y,\theta) \propto q_{Y|\dot{W}}(y|\dot{w};\gamma,\sigma^2)q_{\dot{W}|\Phi}(\dot{w}|\phi,\gamma^2)p(\phi;\beta)$,
- $q(\sigma^2|y,\phi,\dot{w},\beta,\gamma^2) \propto q_{Y|\dot{W}}(y|\dot{w};\gamma^2,\sigma^2)p_\sigma(\sigma^2)$,
- $q(\gamma^2|y,\phi,\dot{w},\sigma^2,\beta) \propto q_{Y|\dot{W}}(y|\dot{w};\gamma^2,\sigma^2)q_{\dot{W}|\Phi}(\dot{w}|\phi,\gamma^2)p_\gamma(\gamma^2)$,
- $q(\beta|y,\phi,\dot{w},\sigma^2,\gamma^2) \propto p(\phi;\beta)p_\beta(\beta)$.

These conditional distributions provide the basis for the algorithm, alternating between updating $(\phi,\dot{w})$, β, γ^2 and σ^2 according to their conditional posterior distributions. Updating β, γ^2 and σ^2 can be carried out using standard M-H algorithms and is not discussed in detail here. Updating $(\phi,\dot{w})$ is less straightforward and is detailed in Section 3.4 in the case of the basic data augmentation. This procedure is easily adjustable to the reparameterization case by using the conditional distributions detailed previously. For a practical implementation, we recommend the paper of Roberts and Stramer (2001) which can be adapted to the model $\dot{\mathcal{M}}_h$.

The convergence of this Gibbs algorithm is proved in the following theorem:

Theorem 2 *Let p_θ and $q_{\theta|Y}$ be respectively the prior and the posterior distributions of θ on the model $\mathcal{M}_h$.*

Assuming the proposal distributions specified above, the hybrid Gibbs algorithm detailed previously converges and provides an ergodic Markov Chain generated with the posterior distribution $q_{\theta|Y}$.

Proof: As previously detailed in Section 3.4, the ergodicity of the Markov Chain is ensured if one of the proposal distributions of the cyclic combination fulfills a minoration condition detailed in Tierney (1994), which is the case with the proposal distributions used to generate $(\phi,\dot{w})$.

It is well known that prior distributions must be properly defined and that their choice may have a considerable impact on the posterior distribution evaluation. Classically, standard non-informative prior distributions are assumed. Following Gilks *et al.* (1996), Gamma distributions can be chosen for σ and γ, and a multivariate Gaussian distribution for β.

5 Survey of the error induced by the Euler-Maruyama approximation

Both estimation methods proposed in this paper, respectively the maximum likelihood and the Bayesian schemes, generate two distinct types of errors on the parameter estimates that have to be controlled.

The first type of error is induced by the estimation method itself. For the maximum likelihood approach, the estimation algorithm produces a sequence $(\widehat{\theta}_k)_{k\geq 0}$ of estimates which converges towards θ_h^*, the maximum of the $\mathcal{M}_h$-likelihood $q_Y(y;\cdot)$ function. Delyon *et al.* (1999) prove an asymptotic normal result of convergence of an averaged SAEM procedure. The variance of this estimate $\widehat{\theta}_k$ is classically controlled by the standard error evaluated through the Fisher information matrix of the estimates. Kuhn and Lavielle (2004) propose to estimate this Fisher information matrix by using the stochastic approximation procedure and the Louis' missing information principle (Louis, 1982). For the Bayesian approach, the equivalent of the problem of obtaining the standard errors is to obtain estimated variances for the posterior mean $E(\theta|y)$. Gilks *et al.* (1996) propose MCMC convergence diagnostics tools in their book, the simplest and more generally used one is independent parallel simulations mixed in together. Different variance estimates are proposed such as the effective sample size estimate or the batching approach, that provide 95% confidence interval for $E(\theta|y)$ (see Carlin and Louis, 2000, for a review of such methods). Because this type of error is not specific to the situation exposed in this paper, it is not further discussed here.

A second type of error is induced on the estimates by the Euler-Maruyama scheme. Indeed, for the reasons evoked in Section 2, the estimation algorithms are applied to the model $\mathcal{M}_h$ instead of to the model $\mathcal{M}$. In the maximum likelihood approach, the algorithm maximizes the $\mathcal{M}_h$-likelihood function q_Y instead of the $\mathcal{M}$-likelihood function p_Y; in the Bayesian framework, the parameters are generated under the posterior distribution $q_{\theta|Y}$ instead of the posterior distribution $p_{\theta|Y}$.

The aim of this section is to study this second type of error induced by the Euler-Maruyama scheme on the conditional distribution $q_{W,\Phi|Y}$, on the likelihood function q_Y and on the posterior distribution $q_{\theta|Y}$. In Theorem 3 we propose bounds of these three errors as functions of the maximal step size of the Euler-Maruyama scheme h. In the following, some additional assumptions hold:

Assumption (**A3**):

The function $F : \mathbb{R} \times [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is infinitely differentiable in the variable space and its partial derivatives of any order are uniformly bounded with respect to x and ϕ.

Assumption (**A4**):

The assumption **(UH)** of Bally and Talay (1996) is satisfied. More precisely, let A_0 and A_1 denote the vector fields defined respectively by $A_0 = F(\cdot)\partial_z$ and $A_1 = \gamma\partial_z$. For multiindices $a = (a_1, \ldots, a_\ell) \in \{0,1\}^\ell$, let the vector fields A_1^a be defined by induction: $A_1^\varnothing := A_1$ and for $j = 0$ or 1 $A_1^{(a,j)} := [A_1^j, A_1^a]$ where $[\cdot,\cdot]$ denotes the Lie bracket. For $L \leq 1$, let define the quadratic form $V_L(\xi,\eta) := \sum_{|a|\leq L-1} \langle A_1^a(\xi), \eta\rangle^2$. We assume that

$$V_L(\xi) := 1 \wedge \inf_{\|\eta\|=1} V_L(\xi,\eta) \geq 0$$

Theorem 3 *Let the assumptions (**A0-A4**) hold.*

1. *Let Z and W be the diffusion processes of the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively, at the observation times: $Z = (Z(t_0), \cdots, Z(t_J))$ and $W = (W(t_0), \cdots, W(t_J))$.*

   Let $p_{Z,\Phi|Y}$ and $q_{W,\Phi|Y}$ be the conditional distributions on the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. There exists a constant $C(y)$ such that, for any $0 < h < H_0$,

   $$\left\|p_{Z,\Phi|Y} - q_{W,\Phi|Y}\right\|_{TV} \leq C(y)h,$$

   where $\|\cdot\|_{TV}$ denotes the total variation distance.

2. *Let p_Y and q_Y be the likelihoods of the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. There exists a constant $C_2(y)$ such that for all $0 < h < H_0$*

   $$\sup_{\{\theta=(\beta,\gamma^2,\sigma^2),\sigma^2>\sigma_0^2,\gamma^2>\gamma_0^2\}} |p_Y(y;\theta) - q_Y(y;\theta)| \leq C_2(y)h.$$

3. *In the Bayesian approach, let p_θ denote the prior distribution. Let $p_{\theta|Y}$ and $q_{\theta|Y}$ be the posterior distributions of the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. There exists a constant $C_3(y)$ such that for all $0 < h < H_0$*

   $$\left\|q_{\theta|Y} - p_{\theta|Y}\right\|_{TV} \leq C_3(y)h.$$

Theorem 3 is proved in Appendix A. These results are based on the convergence rate of the transition densities proposed by Bally and Talay (1996).

***Remark* 4** *Assumption (**A3**) requires only the derivatives of the function F to be bounded and not F itself. Assumption (**A4**) is easily satisfied for linear drift functions F: $F(x,t,\phi) = A(\phi,t) + xB(\phi,t)$.*

6 Theophyllin pharmacokinetic example

The maximum likelihood estimation method developed in Section 3 is applied below to a pharmacokinetics example.

6.1 Pharmacokinetics and Non-linear mixed effects models

Pharmacokinetics (PK) studies the time course of drug substances in the organism. This can be described through dynamic systems, the human body being assimilated to a set of compartments within which the drug evolves with time. In general, these systems are considered in their deterministic version. However, in a recent book on PK modeling, Krishna (2004) claims that the fluctuations around the theoretical pharmacokinetic dynamic model may be appropriately modeled by using SDEs rather than ODEs. Overgaard *et al.* (2005) suggest the introduction of SDEs to consider serial correlated residual errors due for example to erroneous dosing, sampling history or structural model misspecification. This new variability is distinct from the standard measurement noise representing the experimental uncertainty such as assay error.

Generally, several patients are followed up in a clinical trial, their drug concentration being measured along time repeatedly. Longitudinal data are thus gathered at discrete times and classically analyzed using non-linear mixed-effects models. Indeed, the mixed-effects models are a means to discriminate the intra-subject variability from the inter-subject variability, the parameter ϕ being a random parameter proper to each subject. The non-linear mixed-effects model can be written as follows:

$$\left.\begin{array}{rcl} y_{ij} & = & Z(t_{ij},\phi_i)+\varepsilon_{ij} \\ \varepsilon_{ij} & \sim & \mathcal{N}(0,\sigma^2) \\ \phi_i & \sim & \mathcal{N}(\mu,\Omega) \end{array}\right\} \qquad (\mathcal{M}_{mix})$$

where y_{ij} is the observation for subject i, $i=1,\ldots,I$ at time t_{ij}, $j=1,\ldots,J_i$ and ϕ_i is the vector of individual and non-observed parameters of subject i.

In a deterministic approach, the regression function Z is defined as the solution of a PK ordinary differential system: $dZ(t)/dt = F(Z(t),t,\phi)$ with $Z(t_0)=Z_0$, each component of the vector ϕ having a PK meaning. For example, a classic one compartment PK model with first order absorption and first order elimination is described by the following dynamic equation: $Z_0 = Dose$ and

$$\frac{dZ(t,\phi)}{dt} = \frac{Dose\cdot K_a K_e}{Cl} e^{-K_a t} - K_e Z(t,\phi), \tag{6.1}$$

where Z is the drug concentration, $Dose$ is the known drug oral dose received by the subject,

K_e is the elimination rate constant, K_a is the absorption rate constant and Cl is the clearance of the drug. A stochastic differential system can be deduced from the ODE:

$$dZ(t,\phi) = \left(\frac{Dose \cdot K_a K_e}{Cl} e^{-K_a t} - K_e Z(t,\phi) \right) dt + \gamma dB_t \tag{6.2}$$

where B_t is a Brownian motion and γ is the volatility coefficient of the SDE.

In its SDE version, the non-linear mixed-effects model ($\mathcal{M}_{mix}$) is a particular case of the model $\mathcal{M}$ previously presented i.e. a diffusion process is observed at discrete times with noise measurement and its drift function parameters are random.

6.2 Simulation study

The aim of this simulation study is to illustrate the accuracy (bias and root mean square errors) of the extended SAEM algorithm developed in Section 3.2 on a PK application.

We use the previous PK model to mimic the Theophyllin drug pharmacokinetic. To prevent the parameters from taking unrealistic negative values, the vector $\phi \in \mathbb{R}^3$ is classically composed of the log parameters $\phi = (\log(K_e), \log(K_a), \log(Cl))$. The individual parameters ϕ are thus simulated with Gaussian distributions $\mathcal{N}(\mu, \Omega)$, with μ equal to $(-2.52, 0.40, -3.22)$ as proposed by Pinheiro and Bates (1995). A diagonal variance-covariance matrix Ω is assumed for the Gaussian distribution of ϕ. Let $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2)$ denote the vector of these variances. The inter-subject variability is set equal for the three parameters: $\omega_1^2 = \omega_2^2 = \omega_3^2 = 0.01$, corresponding to a variation coefficient of 10%. We set a volatility coefficient equal to $\gamma^2 = 0.2$ and an additive Gaussian measurement error $\sigma^2 = 0.1$. We generate 100 datasets with $I = 36$ subjects and with nine blood samples per patient ($J = 8$), taken at 15 minutes, 30 minutes, 1, 2, 3.5, 5, 7, 9, 12 hours after dosing. The drug oral dose ($Dose$) received by the subject is chosen arbitrarily between 3 and 6 mg.

To evaluate the accuracy of the estimates of $\theta = (\mu, \omega^2, \gamma^2, \sigma^2)$ produced by the SAEM algorithm, the estimation of the parameters is performed on the 100 simulated datasets using the extension of the SAEM algorithm presented in Section 3.2.

The Euler-Maruyama scheme included in the SAEM algorithm is implemented on a grid with auxiliary latent data points introduced between each pair of observation instants as detailed in Section 2.3.1. The number of auxiliary points has to be chosen carefully because a volume of missing data too large can induce arbitrarily poor convergence properties of the Gibbs algorithm. In this example, we divide each time interval $[t_{i,j}, t_{i,j+1}]$ into 20 sub-intervals of equal length. This choice supplies a reasonable volume of missing data, avoids unbalance between the observation-time intervals and proves its numerical efficiency in accurately approximating

the solution of the SDE.

The implementation of the Gibbs procedure included in the SAEM algorithm requires subtle tuning in practice. In particular, the simulation of the diffusion process w on the auxiliary grid is highly critical. An unconditioned trajectory simulation with $q(w_{n_j}|w_{n_{j-1}};\theta)$ as proposed by Pedersen (1995) provides poor numerical results in the case of this example. Indeed, a great number of these simulated trajectories produce large jumps $(w_{\tau_{n_j}} - w_{\tau_{n_j}-1})$. The probability of such trajectories being close to zero, it induces too low an acceptance rate. As suggested by Eraker (2001) or Roberts and Stramer (2001) and detailed in Section 3.4, a conditioned trajectory simulation through Brownian bridge distributions is preferred. Moreover, we update the missing trajectories at once for each subject, as recommended by Elerian *et al.* (2001) to avoid a high level of rejection. In this example, we obtain acceptance rates in the neighborhood of 25%.

The implementation of the SAEM algorithm requires initial values and the choice of the stochastic approximation sequence $(\alpha_k)_{k\geq 0}$. The initial values of the parameters are chosen arbitrarily and set to $\theta_0 = (-3, 1, -3, 0.1, 0.1, 0.1, 2, 1)$. The step of the stochastic approximation scheme is chosen as recommended by Kuhn and Lavielle (2005): $\alpha_k = 1$ during the first iterations $1 \leq k \leq K_1$, and $\alpha_k = (k - K_1)^{-1}$ during the subsequent iterations. Indeed, the initial guess θ_0 might be far from the maximum likelihood value and the first iterations with $\alpha_k = 1$ allow the sequence of estimates to converge to a neighborhood of the maximum likelihood estimate. Subsequently, smaller step sizes during $K - K_1$ additional iterations ensure the almost sure convergence of the algorithm to the maximum likelihood estimate. We implement the extended SAEM algorithm with $K_1 = 200$ and $K = 500$ iterations. Figure 1 illustrates the convergence of the parameter estimates provided by the extended SAEM algorithm as a function of the iteration number in a logarithmic scale. During the first iterations of SAEM, the parameter estimates fluctuate, reflecting the Markov chain construction. After 200 iterations, the curves smooth out but still continue to converge towards a neighborhood of the likelihood maximum. Convergence is obtained after 500 iterations.

[Figure 1 about here.]

The relative bias and relative root mean square error (RMSE) for each component of θ are computed and presented in Table 1.

[Table 1 about here.]

The estimates of the mean parameter μ have very low bias (<5%). The variance parameters have small bias (<9%) except γ^2, this variance parameter being slightly over-estimated (13%).

The RMSE are very satisfactory for the mean parameter (<9%). The RMSE for the variance parameters are greater but still satisfactory (≤40%) in comparison to the small number of subjects ($I = 36$). The RMSE of σ^2 is particularly satisfactory (<20%) considering the complexity of the variability model.

In conclusion, even if this simulation study is performed on a complex model, the convergence of the extended SAEM algorithm towards the maximum likelihood neighborhood is computationally efficient. In addition despite the fact that the number of subjects is small, the extended SAEM algorithm all in all supplies accurate estimations of the parameters. Furthermore, the accuracy is comparable to that obtained with the classic SAEM algorithm for an ODE version of a mixed model ($\mathcal{M}_{mix}$) i.e. for a model with one less variability level.

6.3 A real data example

The extended SAEM algorithm is used to estimate the PK parameters of the Theophyllin drug PK real dataset. This new analysis of the Theophyllin dataset aims at illustrating the advantage of the SDE approach over the ODE approach.

In this clinical trial, twelve subjects received a single oral dose of 3 to 6 mg of Theophyllin. Ten blood samples were taken 15 minutes, 30 minutes, 1, 2, 3.5, 5, 7, 9, 12 and 24 hours after dosing. The individual data are displayed in Figure 2.

[Figure 2 about here.]

The Theophyllin PK is classically described by the one compartment model with first order absorption and first order elimination presented previously. We fit the Theophyllin data with the regression term successively defined as the solution (6.1) and then as that of the SDE (6.2).

In the ODE approach, the differential equation (6.1) has an explicit solution. Thus, the parameters estimates are obtained using the SAEM algorithm combined with a MCMC procedure proposed by Kuhn and Lavielle (2004). The individual concentration profiles are predicted by $\widehat{Z}_{ij} = Z(t_{ij}, \hat{\phi}_i)$ for all i and j where Z is the solution of (6.1) and $\hat{\phi}_i$ is the posterior mean evaluated during the last iterations of the SAEM algorithm. In the SDE approach, the same implementation of the extended SAEM algorithm as the one detailed for the simulation study is used. The individual concentration predictions $E(Z(t_{ij}, \phi_i)|y_i; \hat{\theta})$ for all i and j are evaluated by $\widehat{Z}_{ij} = 1/100 \sum_{k=K-99}^{K} Z^{(k)}(t_{ij}, \phi_i^{(k)})$ where $Z^{(k)}(t_{ij}, \phi_i^{(k)})$ is simulated under the conditional distribution $q_{W,\Phi|Y}(\ .\ |y_i; \widehat{\theta})$ during the 100 last iterations of the extended SAEM algorithm.

The ODE and SDE predictions are overlaid on the data in Figure 3 for six typical subjects. Both ODE and SDE predicted curves for the other six subjects are satisfactory and thus not presented here.

[Figure 3 about here.]

For these six subjects, the ODE predicted curves miss some of the observed data, particularly the last one or the initial concentration peak. The SDE predicted curves improve almost all of these individual profiles.

In conclusion, in this real dataset case study, the individual predictions supplied by the SDE model fit the data better than those obtained by the ODE model. Consequently, in this case, the SDE approach has to be preferred to the ODE approach.

7 Discussion

This paper proposes estimation methods for models defined by a discretely observed diffusion process including additive measurement noise and with random drift function parameters. To that end, an approximate model $\mathcal{M}_h$ is introduced, of which the regression term is evaluated using a Gaussian Euler-Maruyama approximation of maximal step size h. A Gibbs sampler based on the reparameterization of the model suggested by Roberts and Stramer (2001) in the Bayesian framework and the SAEM algorithm in the Maximum Likelihood approach are extended to this model. These two estimation algorithms require the simulation of the missing data (w, ϕ) with the conditional distribution $q_{W,\Phi|Y}$. The choice of the proposal distributions governs the convergence properties of the algorithm and thus is a key issue. A tuned MCMC procedure to perform this simulation is thus proposed, combining a hybrid Gibbs algorithm with independent or random walk Metropolis-Hastings schemes.

Moreover, the error induced by the Euler-Maruyama Gaussian approximation of the diffusion process on the conditional distribution, the likelihood and the posterior distribution of the model $\mathcal{M}_h$ are controlled by the step size h of the numerical scheme. This error is distinct from the error on the estimates induced by the estimation algorithms.

In the maximum likelihood approach, the stochastic version of the EM algorithm SAEM proposed by Kuhn and Lavielle (2004) is preferred to the Monte-Carlo EM (MCEM) developed by Wei and Tanner (1990) or Wu (2004) because of its computational properties. Indeed, SAEM requires the generation of only one realization of the non-observed data at each iteration. In a context where the missing data have to be simulated by a MCMC method, decreasing the size of these missing data is a key issue to ensure acceptable computational times.

The accuracy of the extended SAEM algorithm is illustrated on a pharmacokinetic simulation study using a non-linear mixed effect model defined by SDEs. The relevance of the SDEs approach with respect to the deterministic one is exemplified on a real dataset.

The estimation of such models is mentioned in few papers and is not completely solved. In

the general case of non-linear SDE, the only method proposed is exclusively adapted to the particular case of mixed models and for a maximum likelihood approach (Overgaard *et al.*, 2005). Furthermore, this method relies on the linearization of the model and its convergence is not established. In this paper, we propose estimation methods not only for classic but also Bayesian inference that are adapted to more general missing data models. In addition, the convergence of the proposed algorithms is demonstrated.

References

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica* **70**, 223–262.

Bally, V. and Talay, D. (1995). The law of the Euler Scheme for Stochastic Differential Equations: I. Convergence Rate of the Density. Tech. Rep. 2675, INRIA.

Bally, V. and Talay, D. (1996). The law of the Euler scheme for stochastic differential equations (II): convergence rate of the density. *Monte Carlo Methods Appl.* **2**, 93–128.

Bennet, J. E., Racine-Poon, A. and Wakefield, J. C. (1996). *MCMC for nonlinear hierarchical models*, pp. 339–358. Chapman & Hall, London.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. (2005). Retrospective exact simulation of diffusion sample paths with applications. *submitted* .

Bibby, B. M. and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1**, 17–39.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*, vol. 69 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational. Statistics Quaterly* **2**, 73–82.

Chen, H. F., Lei, G. and Gao, A. J. (1988). Convergence and robustness of the robbins-monroe algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.* **27**, 217–231.

Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19**, 263–284.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27**, 94–128.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38. With discussion.

Elerian, O., Chib, S. and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**, 959–993.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *J. Bus. Econ. Statist.* **19**, 177–191.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall, London.

Gloter, A. and Jacod, J. (2001). Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Probab. Stat.* **5**, 225–242.

Jensen, J. and Petersen, N. (1999). Asympotical normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27**, 514–535.

Krishna, R. (2004). *Applications of Pharmacokinetic principles in drug development*. Kluwer Academic/Plenum Publishers, New York.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM Probab. Stat.* **8**, 115–131.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.* **49**, 1020–1038.

Kusuoka, S. and Stroock, D. (1985). Applications of the malliavin calculus, part II. *J. Fac. Sci. Univ. Tokyo. Sect. IA, Math.* **32**, 1–76.

Kutoyants, T. (1984). *Parameter estimation for stochastic processes*. Helderman Verlag Berlin.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226–233.

Meyn, S. and Tweedie, R. (1993). *Markov chains and stochastic stability*. Comm. Control. Engrg. Ser. Springer-Verlag London Ltd., London.

Overgaard, R., Jonsson, N., Tornøe, C. and Madsen, H. (2005). Non-linear mixed-effects models with stochastic differential equations: Implementation of an estimation algorithm. *J Pharmacokinet. Pharmacodyn.* **32**, 85–107.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.* **22**, 55–71.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the non-linear mixed-effect models. *J. Comput. Graph. Statist.* **4**, 12–35.

Poulsen, R. (1999). Approximate maximum likelihood estimation of discretely observed diffusion process. *Center for Analytical Finance* **Working paper 29**.

Prakasa Rao, B. (1999). *Statistical Inference for Diffusion Type Processes.* Arnold Publisher.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika* **88**, 603–621.

Schweppe, F. (1965). Evaluation of likelihood function for gaussian signals. *IEEE Trans. Inf. Theory* **11**, 61–70.

Singer, H. (1993). Continuous-time dynamical systems with sampled data, error of measurement and unobserved components. *J. Time Series Anal.* **14**, 527–545.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *Int. Stat. Rev* **72**, 337–354.

Sørensen, M. (2000). Prediction-based estimating functions. *Econom. J.* **3**, 123–147.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1762.

Tornøe, C., Overgaard, R., Agersø, H., Nielsen, H., Madsen, H. and Jonsson, E. (2005). Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharm. Res.* **22**, 1247–58.

Wakefield, J., Smith, A., Racine-Poon, A. and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models using the gibbs sampler. *Appl. Statist* **43**, 201–221.

Wei, G. C. G. and Tanner, M. A. (1990). Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika* **77**, 649–652.

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *J. Amer. Statist. Assoc.* **99**, 700–709.

A Proof of theorem 3

1. The aim is to bound

$$\left\|p_{Z,\Phi|Y} - q_{W,\Phi|Y}\right\|_{TV} = \int \left|p_{Z,\Phi|Y}(x,\phi|y;\theta) - q_{W,\Phi|Y}(x,\phi|y;\theta)\right| dxd\phi,$$

Using the fact that the conditional distributions $p_{Y|Z}(x;\sigma^2)$ and $q_{Y|W}(x;\sigma^2)$ are equal, the Bayes theorem application provides:

$$\begin{aligned}&\left|p_{Z,\Phi|Y}(x,\phi|y;\theta) - q_{W,\Phi|Y}(x,\phi|y;\theta)\right| \leq \\ &\frac{p_{Y|Z}(x;\sigma^2)\pi(\phi;\beta)}{p_Y(y;\theta)}\left[\left|p_{Z|\Phi}(x|\phi;\gamma^2) - q_{W|\Phi}(x|\phi;\gamma^2)\right| + \frac{q_{W|\Phi}(x|\phi;\gamma^2)}{q_Y(y;\theta)}\left|p_Y(y;\theta) - q_Y(y;\theta)\right|\right]\end{aligned}$$

As a consequence, the total variation distance is bounded by:

$$\begin{aligned}\left\|p_{Z,\Phi|Y} - q_{W,\Phi|Y}\right\|_{TV} \quad &\leq \frac{\int p_{Y|Z}(x;\sigma^2)\pi(\phi;\beta)dxd\phi}{p_Y(y;\theta)}\left[\sup_{x,\phi}\left|p_{Z|\Phi}(x|\phi;\gamma^2) - q_{W|\Phi}(x|\phi;\gamma^2)\right|\right. \\ &\left. + \frac{|p_Y(y;\theta) - q_Y(y;\theta)|}{q_Y(y;\theta)}\sup_{x,\phi} q_{W|\Phi}(x|\phi;\gamma^2)\right] \qquad \text{(A.1)}\end{aligned}$$

(a) The quantity $\sup_{x,\phi}\left|p_{Z|\Phi}(x|\phi;\gamma^2) - q_{W|\Phi}(x|\phi;\gamma^2)\right|$ is bounded using a result demonstrated by Bally and Talay (1995). This result, based on the Malliavin Calculus, controls the density convergence rate in the case of the Euler-Maruyama scheme.

- By the assumption (**A3**) and because the volatility function is constant, the Hörmander's condition detailed in Bally and Talay (1995) is verified. Thus, there exists a constant $C(\phi,\gamma^2,t_j - t_{j-1})$ independent of h, x_j and x_{j-1} such that

$$\left|p_{Z|\Phi}\left(x_j|x_{j-1},\phi;\gamma^2\right) - q_{W|\Phi}\left(x_j|x_{j-1},\phi;\gamma^2\right)\right| \leq C(\phi,\gamma^2,t_j - t_{j-1})h.$$

The constant depends on the bounds of the derivatives of the drift function, independent of ϕ under assumption (**A3**). Besides, if γ^2 is contained in $[\gamma_0,\Gamma_0]$, there exists C_1 independent of γ^2 such that, for all $j = 1\cdots J$,

$$\left|p_{Z|\Phi}\left(x_j|x_{j-1},\phi;\gamma^2\right) - q_{W|\Phi}\left(x_j|x_{j-1},\phi;\gamma^2\right)\right| \leq C_1 h \qquad \text{(A.2)}$$

- Under the assumption (**A4**) and using a result of Kusuoka and Stroock (1985) based on the Malliavin calculus (corollary 3.25), for all $j = 1 \cdots J$, there exists a constant $C_2(\phi, \gamma^2, t_j - t_{j-1})$ such that

$$p_{Z|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right) \leq C_2(\phi, \gamma^2, t_j - t_{j-1})$$

  By the same arguments as before, this constant is bounded independently of ϕ and γ^2. Hence, there exists C_2 such that, for all $j = 1 \cdots J$,

$$p_{Z|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right) \leq C_2 \tag{A.3}$$

- In addition, we can write:

$$\begin{aligned}
&|q_{W|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right)| \\
&\leq |q_{W|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right) - p_{Z|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right)| + |p_{Z|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right)| \\
&\leq hC_1 + C_2.
\end{aligned} \tag{A.4}$$

- Finally, the Markov property provides:

$$\left|p_{Z|\Phi}(x|\phi; \gamma^2) - q_{W|\Phi}(x|\phi; \gamma^2)\right| = \left|\prod_{j=1}^{J} p_{Z|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right) - \prod_{j=1}^{J} q_{W|\Phi}\left(x_j|x_{j-1}, \phi; \gamma^2\right)\right| \tag{A.5}$$

  for any $j = 1 \cdots J$. So, by combining the (A.5), (A.2), (A.3) and (A.4), there exists a bound C_3 independent of h, j and γ^2 such that

$$\sup_{x,\phi} \left|p_{Z|\Phi}(x|\phi; \gamma^2) - q_{W|\Phi}(x|\phi; \gamma^2)\right| \leq C_3 h \tag{A.6}$$

(b) By the Markov decomposition of the probability $q_{W|\Phi}(x|\phi; \gamma^2)$, and using the inequality (A.4), there exists C_4 such that

$$\sup_{x,\phi} q_{W|\Phi}(x|\phi; \gamma^2) \leq C_4 \tag{A.7}$$

By integration and using the inequality (A.6), we have:

$$|p_Y(y; \theta) - q_Y(y; \theta)| \leq C_3 h \int p_{Y|Z}(x; \sigma^2)\pi(\phi; \beta)dxd\phi = C_3 h \tag{A.8}$$

(c) The quantity $q_{W|\Phi}(x|\phi;\gamma^2)$ can be down-bounded. Indeed,

$$\begin{aligned} q_Y(y;\theta) &\geq p_Y(y;\theta) - |p_Y(y;\theta) - q_Y(y;\theta)| \\ &\geq p_Y(y;\theta) - C_3 h \qquad \text{following the inequality (A.8)} \\ &\geq p_Y(y;\theta) - C_3 H_0 \qquad \text{for } h < H_0 \text{ and } H_0 \text{ small enough.} \end{aligned}$$

Hence there exists $C_5(y)$ such that

$$q_Y(y;\theta) \geq C_5(y) \tag{A.9}$$

(d) Finally, the inequalities (A.1) , (A.7), (A.8) and (A.9) provide the final result:

$$\left\| p_{Z,\Phi|Y} - q_{W,\Phi|Y} \right\|_{TV} \leq \frac{1}{p_Y(y;\theta)} \left[C_3 h + \frac{C_3 h}{C_5(y)} C_4 \right]$$

□

2. The proof of the part 2 of Theorem 3 directly derives from (A.8).

3. By Bayes theorem, we have:

$$p_{\theta|Y}(\theta|y) = \frac{p_{y|\theta}(y|\theta)p(\theta)}{p_Y(y)}$$

where $p_Y(y) = \int p_{y|\theta}(y|\theta)p(\theta)d\theta$. From (A.8), there exists a constant C_3, independent of θ such that $|p_{y|\theta}(y|\theta) - q_{Y|\theta}(y|\theta)| \leq hC_3$. Consequently $|p_Y(y) - q_Y(y)| \leq C_3\ h$ and

$$\begin{aligned} |p_{\theta|Y}(\theta|y) - q_{\theta|Y}(\theta|y)| &\leq \frac{p(\theta)}{p_Y(y)} \left| |p_{y|\theta}(y|\theta) - q_{Y|\theta}(y|\theta)| + \frac{q_{Y|\theta}(y)}{q_Y(y)} |q_Y(y) - p_Y(y)| \right| \\ &\leq \frac{C_3 h}{p_Y(y)} p(\theta) \left| 1 + \frac{q_{Y|\theta}(y|\theta)}{p_Y(y)} \right| = C_6(y)\ h \left[p(\theta) + q_{\theta|Y}(\theta|y) \right]. \end{aligned}$$

The final result can be directly deduced:

$$\begin{aligned} \left\| q_{\theta|Y} - p_{\theta|Y} \right\|_{TV} &= \int |p_{\theta|Y}(\theta|y) - q_{\theta|Y}(\theta|y)| d\theta \\ &\leq C_6(y)\ h \int (p(\theta) + q_{\theta|Y}(\theta|y)) d\theta \leq 2\ C_6(y)\ h \end{aligned}$$

□

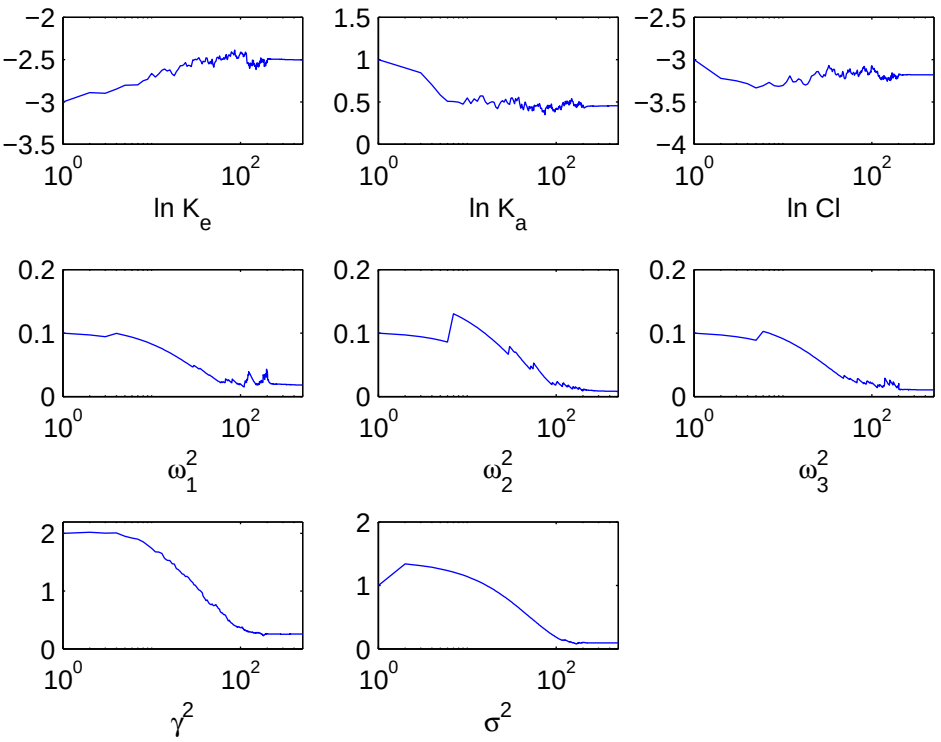


Figure 1: Evolution of the SAEM parameter estimates function of the iteration number in a logarithmic scale

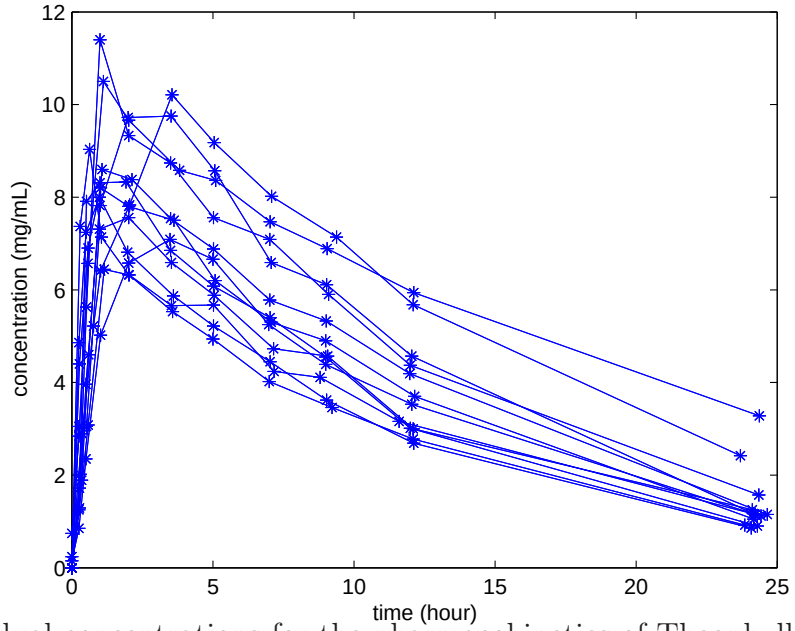


Figure 2: Individual concentrations for the pharmacokinetics of Theophyllin for 12 subjects.

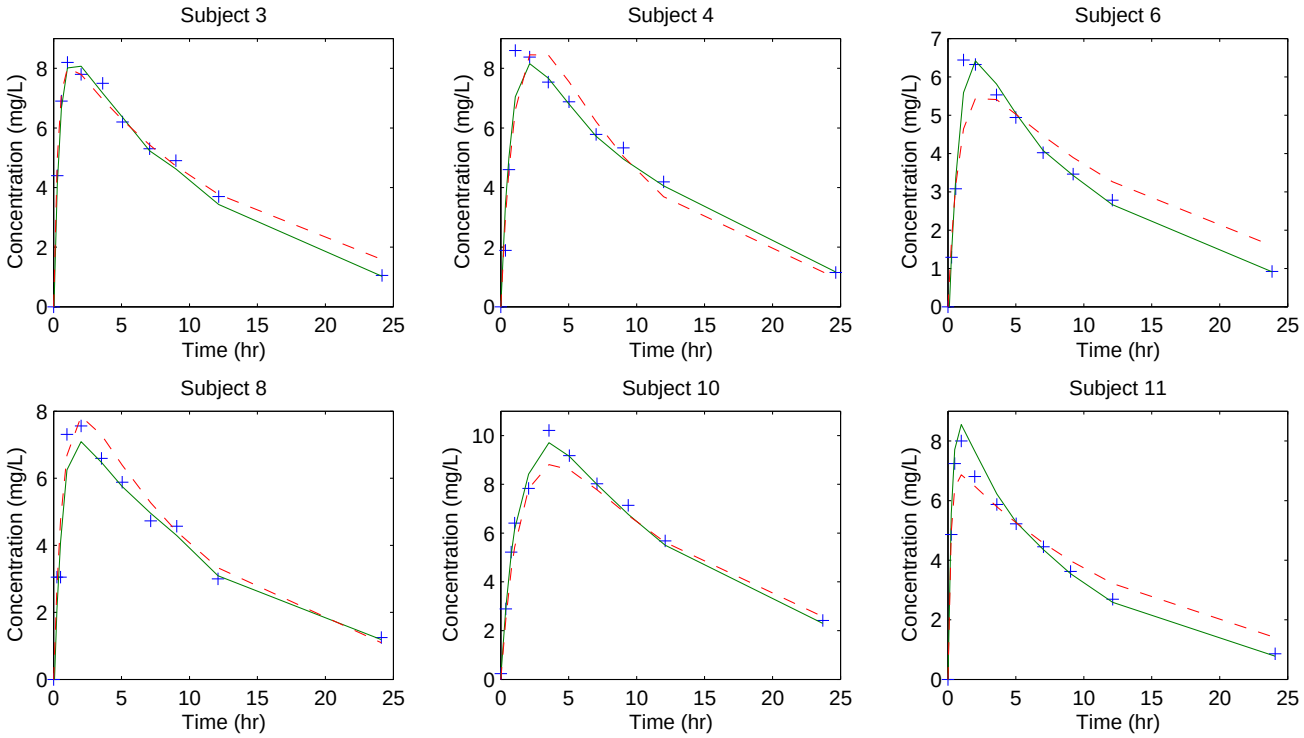


Figure 3: Six individual concentration cruves predictec by SAEM with the ODE approach (dotted line) and the SDE approach (plain line) overlaid on the data points for the pharmacokinetic of Theophyllin

Table 1: Relative bias (%) and relative root mean square error (RMSE) (%) of the estimated parameters evaluated by the SAEM algorithm from 100 simulated trials with $I = 36$ subjects.

Parameters	Bias (%)	RMSE (%)
log K_e	0.42	-3.19
log K_a	4.14	8.95
log Cl	-0.23	-2.27
ω_1^2	3.83	40.03
ω_2^2	8.49	36.76
ω_3^2	-8.81	37.52
γ^2	13.02	21.31
σ^2	-4.44	18.79

Chapitre 7

Analyse de la dynamique du VIH dans l'essai COPHAR II-ANRS 111

Muni d'un outil adapté à l'estimation des paramètres d'un modèle mixte défini par un système différentiel, nous nous sommes ensuite intéressés à la modélisation conjointe de la décroissance de la charge virale du VIH et de la croissance des lymphocytes $CD4^+$ sous traitement anti-rétroviral, modélisation s'appuyant sur un système différentiel décrivant le processus d'infection par le virus. De nombreux systèmes différentiels ont été présentés depuis une dizaine d'années, l'un des premiers modèles ayant été proposé par Perelson et al. [2].

Ainsi avant toute initiation d'un traitement anti-rétroviral, chaque virus peut pénétrer dans une cellule immunitaire lymphocyte $CD4^+$ avec un taux γ. Le lymphocyte devient alors infecté et capable de produire de nouveaux virus à une fréquence II. Soient T_{NI} et T_I les concentrations de lymphocytes $CD4^+$ non-infectés et infectés respectivement et soit V la concentration du virus dans le plasma. Sous ces hypothèses, le système différentiel décrivant la dynamique virale avant traitement s'écrit

$$\begin{aligned} \frac{dT_{NI}}{dt} &= \lambda - \gamma T_{NI} V - \mu_{TNI} T_{NI} \\ \frac{dT_I}{dt} &= \gamma T_{NI} V - \mu_{TI} T_I \\ \frac{dV}{dt} &= \Pi T_I - \mu_V V \end{aligned}$$

où λ est le taux de production de $CD4^+$, et μ_{TNI}, μ_{TI} et μ_V sont respectivement les taux de mort des $CD4^+$ non infectés, infectés, et des virus.

La prise en charge thérapeutique de l'infection par le VIH implique actuellement la combinaison d'au moins trois médicaments, en général un inhibiteur de protéase et deux inhibiteurs de la transcriptase inverse. L'inhibiteur de protéase (*protease inhibitor* PI) permet d'éviter la production de nouveaux virus, ou plus exactement, les nouveaux virus créés ne sont pas infectieux. On introduit alors une nouvelle

catégorie de virus dans le système dynamique, dont la concentration est notée V_{NI}, la concentration des virus infectants étant notée V_I. Le système différentiel s'écrit alors :

$$\begin{aligned}
\frac{dT_{NI}}{dt} &= \lambda - \gamma T_{NI} V_I - \mu_{TNI} T_{NI} \\
\frac{dT_I}{dt} &= \gamma T_{NI} V_I - \mu_{TI} T_I \\
\frac{dV_I}{dt} &= (1 - \eta_{PI}) \Pi T_I - \mu_V V_I \\
\frac{dV_{NI}}{dt} &= \eta_{PI} \Pi T_I - \mu_V V_{NI}
\end{aligned}$$

où η_{PI} est un nombre compris entre 0 et 1 représentant l'efficacité de l'inhibiteur de protéase. Perelson et al. [2] font l'hypothèse que l'IP est 100% efficace, c'est-à-dire que $\eta_{PI} = 1$. En supposant que la concentration de $CD4^+$ non-infectés reste constante (c'est-à-dire que le taux de production λ évolue en fonction de la concentration instantanée), ils montrent que la concentration de virus $V = V_I + V_{NI}$ a une solution analytique du type bi-exponentielle. Cependant, leurs hypothèses simplificatrices ne sont pas du tout satisfaisantes ni réalistes.

L'introduction d'un inhibiteur de la transcriptase inverse (*reverse transcriptase inhibitor* RTI) a pour action de bloquer la production d'ADN pro-viral par un virus au sein d'un lymphocyte $CD4^+$. On introduit alors un paramètre compris entre 0 et 1 représentant l'efficacité de ce médicament η_{RTI} :

$$\begin{aligned}
\frac{dT_{NI}}{dt} &= \lambda - (1 - \eta_{RTI}) \gamma T_{NI} V_I - \mu_{TNI} T_{NI} \\
\frac{dT_I}{dt} &= (1 - \eta_{RTI}) \gamma T_{NI} V_I - \mu_{TI} T_I \\
\frac{dV_I}{dt} &= (1 - \eta_{PI}) \Pi T_I - \mu_V V_I \\
\frac{dV_{NI}}{dt} &= \eta_{PI} \Pi T_I - \mu_V V_{NI}
\end{aligned}$$

Différentes extensions de ce modèle ont ensuite été proposées. Certains systèmes intègrent une nouvelle catégorie de lymphocytes $CD4^+$ dit *quiescents* ou *latents*. Ces $CD4^+$ quiescents, dont la concentration est notée T_Q, peuvent redevenir actifs avec un taux α et se désactiver avec un taux r. Deux hypothèses s'affrontent. L'hypothèse la plus communément acceptée est celle d'un réservoir de lymphocytes non-infectés, et qui ne peuvent pas être infectés ni produire de virus tant qu'ils n'ont pas été

réactivés [68]. Cette hypothèse aboutit au système différentiel suivant :

$$\begin{aligned}
\frac{dT_Q}{dt} &= \lambda + rT_{NI} - \alpha T_Q - \mu_Q T_Q \\
\frac{dT_{NI}}{dt} &= \alpha T_Q - (1-\eta_{RTI})\gamma T_{NI} V_I - rT_{NI} - \mu_{TNI} T_{NI} \\
\frac{dT_I}{dt} &= (1-\eta_{RTI})\gamma T_{NI} V_I - \mu_{TI} T_I \\
\frac{dV_I}{dt} &= (1-\eta_{PI})\Pi T_I - \mu_V V_I \\
\frac{dV_{NI}}{dt} &= \eta_{PI}\Pi T_I - \mu_V V_{NI}.
\end{aligned} \tag{7.1}$$

La seconde hypothèse, très controversée, repose sur un travail de Finzi et al. [69] qui montrent que dans le réservoir de $CD4^+$ latents, une partie des $CD4^+$ sont des cellules infectées par le virus, mais ne produisant pas de nouveaux virus, l'ADN obtenu par rétrotranscription de l'ARN viral n'étant ni lu ni transcrit par le $CD4^+$. Lorsque ces $CD4^+$ sont ré-activés, l'ADN obtenu par rétrotranscription de l'ARN viral est lu, conduisant à la production de nouveaux virus. Cette hypothèse aboutit alors au système différentiel suivant proposé par Di Mascio et al [70] :

$$\begin{aligned}
\frac{dT_{QNI}}{dt} &= \lambda + r_{NI} T_{NI} - \alpha_{NI} T_{QNI} - \mu_{QNI} T_{QNI} \\
\frac{dT_{NI}}{dt} &= -(1-\eta_{RTI})\gamma T_{NI} V_I - r_{NI} T_{NI} + \alpha_{NI} T_{QNI} - \mu_{TNI} T_{NI} \\
\frac{dT_{QI}}{dt} &= r_I T_I - \alpha_I T_{QI} - \mu_{QI} T_{QI} \\
\frac{dT_I}{dt} &= (1-\eta_{RTI})\gamma T_{NI} V_I + \alpha_I T_{QI} - r_I T_I - \mu_{TI} T_I \\
\frac{dV_I}{dt} &= (1-\eta_{PI})\Pi T_I - \mu_V V_I \\
\frac{dV_{NI}}{dt} &= \eta_{PI}\Pi T_I - \mu_V V_{NI}.
\end{aligned}$$

où r_{NI}, r_I, α_{NI} et α_I, T_{QNI} et T_{QI} sont respectivement les taux d'activation et de désactivation et les concentrations de $CD4^+$ quiescents non-infectés et infectés. Ce système est extrêmement complexe, et très difficile à résoudre numériquement. Des hypothèses simplificatrices comme celles de Perelson et al. [2] sont alors proposées par les auteurs, permettant à nouveau d'approcher la solution de la concentration de virus $V = V_I + V_{NI}$ par des sommes de trois ou quatre exponentielles. L'intérêt d'introduire cette nouvelle catégorie de lymphocytes $CD4^+$ est alors discutable.

Dans la suite de ce travail, nous nous sommes intéressés au système (7.1), qui est résolvable numériquement, mais nécessite des méthodes de résolution numériques performantes, puisqu'il est complexe et *stiff* (une équation différentielle est appelée *stiff* si toute méthode de résolution numérique explicite échoue à la résoudre). Il faut alors avoir recours à des méthodes de résolution implicite et à plusieurs étapes.

D'autres extensions de ces modèles ont été proposées, en particulier prenant en compte des effets des médicaments dépendant du temps. Huang et al. [71], Wu et al. [72] ont proposé d'intégrer la pharmacocinétique de l'inhibiteur de protéase et de l'inhibiteur de la transcriptase inverse dans ce modèle différentiel. Les paramètres d'efficacité sont alors des fonctions dépendantes du temps décrites par un modèle pharmacodynamique du type E_{max} :

$$\begin{aligned} \eta_{IP}(t) &= \frac{C_{PI}(t)}{C_{50,PI} + C_{PI}(t)} \\ \eta_{RTI}(t) &= \frac{C_{RTI}(t)}{C_{50,RTI} + C_{RTI}(t)} \end{aligned}$$

où C_{PI} et C_{RTI} sont les concentrations de l'inhibiteur de protéase et de l'inhibiteur de la transcriptase inverse respectivement. Ceci requiert la modélisation préalable de la pharmacocinétique des médicaments, et leur intégration dans le système différentiel.

Nous avons analysé les données des patients infectés par le VIH et recevant une multi-thérapie contenant l'inhibiteur de protéase lopinavir, et suivis dans le cadre de l'essai COPHAR II-ANRS 111. Le but de ce travail était d'estimer les paramètres dynamiques du système (7.1) à partir de la modélisation de l'évolution conjointe de la charge virale et du nombre de $CD4^+$. En particulier, l'objectif était d'évaluer l'efficacité du lopinavir (paramètre η_{PI} du système différentiel) et sa variabilité dans cette population de patients répondant au traitement.

Nous avons utilisé l'algorithme SAEM et les différentes extensions de cet algorithme réalisées au cours de cette thèse pour résoudre de façon satisfaisante ce problème, à savoir une estimation par maximum de vraisemblance de l'ensemble des paramètres d'un système d'équations différentielles décrivant l'évolution conjointe de la charge virale et du nombre de $CD4^+$, et prenant en compte la censure des mesures de charges virales. Le schéma de linéarisation locale de résolution numérique développé au chapitre 6.1 a été nécessaire pour parvenir à la convergence numérique de l'algorithme d'estimation SAEM.

Ce travail fait l'objet d'un article en préparation pour *Antiviral Therapy*.

L'effet des différentes covariables (age, sexe, concentration résiduelle de l'inhibiteur de protéase, échelle d'adhérence au traitement, etc) sur l'efficacité des traitements pourra être testée ultérieurement à partir des estimations a posteriori des paramètres dynamiques individuels du modèle. Cette étude sur les covariables n'a pas encore été réalisée.

Estimation of parameters of a simultaneous long-term model of HIV and CD4+ dynamics in patients initiating a lopinavir containing HAART.

Adeline Samson[1,2], Xavière Panhard[1,2], Marc Lavielle[3], France Mentré[1,2]

March 29, 2006

[1] INSERM, U738, Paris, F-75018 France;

[2] Université Paris 7, Faculté Xavier Bichat, Paris, F-75018 France

[3] Université Paris-Sud, Bat. 425, Orsay, F-91000 France

Abstract

A long-term HIV dynamic model was proposed to estimate dynamic parameters from patients treated by a HAART containing the protease inhibitor lopinavir in the COPHAR II-ANRS 111 trial. These patients were protease inhibitor-naive and followed up during 48 weeks. HIV dynamic parameters were estimated by fitting the model to both the viral load and the $CD4^+$ concentration data. A large inter-patient variability was observed in estimated dynamic parameters. We estimated the efficacy of the protease inhibitor for each patient. The simultaneous modeling of the viral load decrease and the $CD4^+$ increase enabled to differentiate individual dynamic trends that could not be distinguished when modeling only the viral load decrease. These results suggested that viral dynamic parameters may play an important role in determining and predicting treatment long term success. The proposed mathematical model and statistical techniques could be used to simulate and predict antiviral response for individual patients.

Introduction

One of the main consequences of the infection by the human immunodeficiency virus (HIV) is the depletion in $CD4^+$ T cells. Because of their central role in the immune regulation, their depletion can have widespread deleterious effects on the functioning of the immune system as a whole. Over the past decade, a number of mathematical models have been developed to describe the immune system, the population dynamics of the virus and its interaction with the human cells [Perelson et al., 1996, Perelson et al., 1997, Nowak and Bangham, 1996]. Recent reviews of these models can be found in articles by [Perelson and Nelson, 1997, Nowak and May, 2000, Callaway and Perelson, 2002, Perelson, 2002].

Two types of antiviral drugs are commonly used in standard anti-HIV treatment. The first are protease inhibitors (PIs). They inhibit the cleavage of long precursor proteins into smaller functional units. They do not inhibit the formation of viral particles, but they dramatically decrease the infectious capacity of a large fraction of viral particles, and therefore disable the virus ability to produce any offspring. The second class of antiviral drugs considered in anti-retroviral therapy is reverse transcriptase inhibitors (RTIs). They inhibit the transcription of the viral RNA into double-stranded DNA, that can then be integrated into the DNA of the host-cell. The specific effects of these two classes of anti-retroviral drugs have also been incorporated into these models.

The estimation of the parameters of these differential systems from HIV data is crucial to better understand the viral dynamic on a long term. Particularly, antiviral response varying extensively among patients, the estimation of individual dynamic parameters enables to individualize anti-retroviral treatments. The estimation of these individual dynamic parameters is therefore pivotal to improve the disease management. The objective of this paper is to estimate such parameters for patients treated by a multi-therapy containing the lopinavir protease inhibitor.

In this context, mixed models are able to distinguish two sources of variabilities: the variability between the individuals called *inter*-patient and the *residual* variability. Furthermore these models are adapted to sparse design, i.e when few observations are available per subject, which is frequent in HIV studies. The principle of mixed effects models is to evaluate the distribution of the dynamic model parameters among the whole population by considering the statistical model individual parameters as random variables (called random effects) centered around a mean value (fixed effects). Thus a mean population curve can be deduced, reflecting the average evolution of the dynamic process observed among patients. Individual curves are also predicted for each patient reflecting its own mechanism according to its covariables.

Several HIV dynamic studies used these mixed models to analyze the HIV viral dynamic [Wu et al., 1998, Fitzgerald et al., 2002, Putter et al., 2002, Wu and Zhang, 2002, Wu, 2004]. How-

ever, most of them have only considered the viral load decrease under unrealistic assumptions: constant concentration of $CD4^+$ and/or total drug efficiency [Wu et al., 1998, Ding and Wu, 2001, Fitzgerald et al., 2002]. A first assumption proposed by [Perelson et al., 1996, Wu et al., 1998] is to consider a short time period and thus assume that the concentration of non-infected $CD4^+$ cells is a constant. In this case, the differential system becomes linear with an analytical solution. Aiming at studying both the viral load decrease and the $CD4^+$ increase, this hypothesis of a constant $CD4^+$ concentration is unrealistic. In such cases this differential system remains non-linear, and has no longer any analytical solution. A second unrealistic assumption often made is that the initiated therapy inhibits any new infection. Under that assumption, the system can be solved explicitly as proposed by [Wu et al., 1998, Putter et al., 2002]. [Putter et al., 2002] proposed a first simultaneous estimation of the viral load and $CD4^+$ dynamic based on a differential system under this unrealistic assumption. Therefore, they only focused on the first two weeks of the dynamic after initiation of an anti-retroviral treatment. But if a perfect effect of protease inhibitor is not presupposed, the system can no longer be solved analytically. Recently, [Wu et al., 2005] estimated only the viral load (i.e. the $CD4^+$ were not taken into account) with a dynamic system describing the long-term HIV dynamics and considering drug potency, drug exposure/adherence and drug resistance during chronic treatment of HIV-1 infected patients. However, they did also consider a simplified model. They did neither consider separately the compartments of HIV-producing infected cells nor latent cells and not decompose the virus compartment into infectious and noninfectious virions as proposed by [Perelson et al., 1996, Perelson et al., 1997].

An additional difficulty in such statistical analysis is that viral load measurements are subject to left censoring, as the experimental devices are not able to measure low-level of viral load with sufficient accuracy. Therefore, when viral load are below a given limit, namely the limit of quantification (LOQ), the exact value of the viral load is unknown. This limit is generally between 20 and 400 copies/ml. Besides, the proportion of subjects with viral load below LOQ has increased since highly active anti-retroviral treatments have been introduced. Working with such left-censored data further complicates the study of longitudinal viral load data. [Thiebaut et al., 2005] jointly modeled the $CD4^+$ and the viral load dynamic handling the censored viral load data. They also included the potential informative dropout with a survival model. However, they used a bivariate linear model, which is an analytical solution of a simplified differential system. [Wu et al., 2005] did not take into account the left-censored viral load data, whereas it is well known that this censoring may induce biased parameter estimates. At the contrary, [Putter et al., 2002] took into account the left-censored viral load data in the statistical analysis.

The objective of this analysis was to estimate HIV dynamic parameters from PI-naive HIV-infected patients beginning a highly active anti-retroviral treatment (HAART) containing the lopinavir

protease inhibitor. The aim was also to estimate the HIV dynamic parameters variabilities among these patients, particularly the variability of the lopinavir effect, in order to better understand the variety of dynamic viral responses and to be able to predict viral dynamic trend for each patient. These patients were followed-up during 48 weeks in the AIDS clinical trial COPHAR II-ANRS 111 developed by the French Agence Nationale de Recherche contre le Sida (ANRS). We proposed to use a long-term dynamic system including latent T cells and noninfectious virus. Indeed this model is more appropriate and closer to the reality than those used by [Putter et al., 2002] or [Wu et al., 2005]. We also proposed to analyze simultaneously the HIV viral load decrease and the $CD4^+$ increase. We used in this analysis a new Stochastic Approximation Expectation Maximization algorithm, proposed by [Delyon et al., 1999] to estimate the physiological dynamic parameters by maximizing the likelihood of this mixed model.

Materials and methods

Study patients

COPHAR II-ANRS 111 was a prospective, multicenter, 48-week open study which was conducted in HIV-infected patients naive from any protease inhibitor (PI) treatment. They received one PI among three (indinavir, lopinavir or nelfinavir) and a double nucleoside reverse transcriptase inhibitor combination with open choice. The main aim of the COPHAR II-ANRS 111 study was to evaluate the feasibility and the efficacy of Therapeutic Drug Monitoring (TDM) of the three PI to improve antiviral efficacy and tolerance of PI-containing Highly Active Antiretroviral Treatment (HAART) and to warrant virological success and safety of HAART.

All study sites were approved to perform the study by institutional review boards, and all subjects gave their written informed consent before their participation. Eligible subjects were HIV-1 infected individuals older than 18 years with a plasma HIV-1 RNA concentration above 1000 copies/ml within the 4-6 weeks before their entry, no prior use of any PI-containing therapy. Pregnant women, patients with HIV acute infection, and those with chronic diarrhoea, diabetes mellitus, renal or liver or cardiac diseases or a history of nephrolithiasis were excluded. Eligible patients were assigned to receive a PI-containing antiviral therapy among the following combination: a double nucleoside reverse transcriptase inhibitor combination with open choice and a PI chosen among one of the following: lopinavir, indinavir or nelfinavir. We only considered in this paper patients receiving lopinavir administered with ritonavir PI as follows: 400 mg of lopinavir and 100 mg of ritonavir twice daily. The strategy of TDM proceeded as follows: if the trough plasma concentration was out of the adequate predefined range (2500-7000 ng/mL for lopinavir), the PI doses were adjusted by increments of one pill bid: 133/33 mg for lopinavir/ritonavir.

Study visits occurred at screening, at inclusion (W0) and at weeks 2, 4, 8, 16, 24 and 48

following initiation of treatment. Plasma HIV RNA dosing was performed at visits W0, W2, W8, W16, W24 and W48. Standard biological exams, including $CD4^+$ cell counts, were performed at all but the second week visit.

HIV dynamics modeling

We distinguished quiescent $CD4^+$ T cells, T_Q, target or activated non-infected T cells, T_{NI}, productively infected T cells, T_I, and virus particles, V. In our model, only activated $CD4^+$ T cells are assumed to may become infected by HIV, and quiescent cells are assumed to be resistant to infection [Stilianakis et al., 1997, De Boer and Perelson, 1998]. Quiescent $CD4^+$ T-cells T_Q are generated through the hematopoietic differentiation process at a constant rate λ. In adults, as the $CD4^+$ T cell compartment is largely maintained by self-renewal [Stilianakis et al., 1997], the dynamic model allows the quiescent $CD4^+$ T cells to become activated at a low constant rate α. Quiescent $CD4^+$ T cells are assumed to die at a rate μ_Q, and to appear by the deactivation of activated T cells at a rate r. The activated $CD4^+$ T cells appear by activation of quiescent cells at a rate α as stated before, they revert to the quiescent stage at a rate r and they are infected by the virus at a rate γ per susceptible cell and virus. They die at a rate μ_{TNI}. The productively infected $CD4^+$ T cells appear by the infection of target cells by the virus at a rate γ and die at a rate μ_{TI} either due to the action of the virus or the immune system. The total number of $CD4^+$ T cells is therefore defined by $T_{tot} = T_Q + T_{NI} + T_I$. Infected $CD4^+$ T-cells produce virus at a rate of Π per infected cell. The virus are cleared at a rate of μ_V.

The effects of the anti-retroviral treatments are also incorporated into the model. Two types of antiviral drugs are commonly used in standard anti-HIV treatment, protease inhibitors (PIs) and reverse transcriptase inhibitors (RTIs).

The protease inhibitors decrease the infectious capacity of a large fraction of viral particles, and therefore disable the virus ability to produce any offspring. This process is incorporated in the model by adding a compartment for the non-infectious virus denoted V_{NI} and a parameter η_{PI}, which denotes the fraction of produced virus being non-infectious. Therefore, this parameter η_{PI} is a fraction between 0 and 1. A value of $\eta_{PI} = 1$ corresponds to a completely effective drug that results only in the production of non-infectious virus. The lifetime of infectious and non-infectious virus is assumed to be identical.

The reverse transcriptase inhibitors (RTIs) prevent susceptible cells from becoming infected. This process is incorporated in the model by adding a parameter η_{RTI} which denotes the fraction of susceptible cells prevented to be infected. Therefore, this parameter η_{RTI} is a fraction between 0 and 1. A value of $\eta_{RTI} = 1$ corresponds to a completely effective drug that results in preventing all new infections of T cells.

Hence, the system of differential equations describing the viral dynamic after initiation of an anti-retroviral treatment and summarizing the effects of drug therapy can be written as

$$\begin{aligned}
\frac{dT_Q}{dt} &= \lambda + rT_{NI} - \alpha T_Q - \mu_Q T_Q \\
\frac{dT_{NI}}{dt} &= -(1-\eta_{RTI})\gamma T_{NI} V_I + \alpha T_Q - rT_{NI} - \mu_{TNI} T_{NI} \\
\frac{dT_I}{dt} &= (1-\eta_{RTI})\gamma T_{NI} V_I - \mu_{TI} T_I \qquad (S) \\
\frac{dV_I}{dt} &= (1-\eta_{PI})\Pi T_I - \mu_V V_I \\
\frac{dV_{NI}}{dt} &= \eta_{PI}\Pi T_I - \mu_V V_{NI}.
\end{aligned}$$

The total number of virions is therefore defined by $V_{tot} = V_{NI} + V_I$.

The parameters γ, η_{RTI} and η_{PI} are mathematically not all identifiable. Furthermore it is rather optimistic to hope to be able to estimate both η_{RTI} and η_{PI} from the data. To simplify this issue [Putter et al., 2002] proposed to assume $\eta_{RTI} = \eta_{PI}$ and to estimate a single compound treatment effect η. However, RTIs and PIs act in different parts of the HIV replication cycle and no evidence of any theoretical results linking η_{RTI} and η_{PI} can be found in the literature. We decided to focus on the effect of PIs, and therefore we proposed to estimate the two following parameters $\tilde{\gamma} = (1-\eta_{RTI})\gamma$ and η_{PI}.

As proposed by [Perelson et al., 1996], we assumed that all the newly produced viruses are fully infectious before the introduction of a PI treatment. We also assumed that before the treatment initiation, the system has reached an equilibrium state. Thus the initial condition of the differential system (S) can be written as:

$$\begin{aligned}
T_Q(t=0) &= (1/(\alpha+\mu_Q))(\lambda + (r\mu_V\mu_{TNI})/(\Pi\tilde{\gamma})) \\
T_{NI}(t=0) &= \frac{\delta c}{\beta p} \\
T_I(t=0) &= V_I(t=0)\mu_v/\Pi \qquad (0.1) \\
V_I(t=0) &= (\alpha\frac{T_Q(t=0)}{T_{NI}(t=0)} - r - \mu_{TI})/\tilde{\gamma} \\
V_{NI}(t=0) &= 0.
\end{aligned}$$

Nonlinear mixed effects model

As viral dynamic response to treatments is highly variable among patients, the use of a non-linear mixed statistical model allows to analyse the whole data and to estimate both the dynamic parameters and their *inter*-patient variabilities. Let N be the number of subjects and $n_i^{(1)}$, $n_i^{(2)}$ the number of measurements on the i^{th} subject of respectively the viral load and the $CD4^+$ concentration.

The observed log viral load (cp/mL) $y_{ij}^{(1)}$ and the CD4$^+$ concentration (cells/mm^3) $y_{ij}^{(2)}$ of patient i, measured respectively at the sampling time $t_{i,j}^{(1)}$ end $t_{i,j}^{(2)}$, are related as follows

$$\begin{aligned} y_{ij}^{(1)} &= log_{10}(V_{tot}(t_{ij}^{(1)}, \phi_i) \cdot 1000) + \epsilon_{ij}^{(1)} \\ y_{ij}^{(2)} &= T_{tot}(t_{ij}^{(2)}, \phi_i) + T_{tot}(t_{ij}^{(2)}, \phi_i)\, \epsilon_{ij}^{(2)} \end{aligned}$$

where $V_{tot} = V_{NI} + V_I$ (cp/mm^3) and $T_{tot} = T_Q + T_{NI} + T_I$ (cells/mm^3) are the solution functions fo the system (S).

All the parameters are log-transformed in order to ensure positive values, except the parameter η_{PI} which takes its values between $[0, 1]$, and which is thus transformed with a logistic function. Therefore the vector of the transformed physiological parameters of patient i is

$$\phi_i = (\log(\alpha)_i, \log(r)_i, \log(\mu_q)_i, \log(\mu_{ti})_i, \log(\lambda)_i, \log(\tilde{\gamma})_i, \mathrm{logit}(\eta_{PI})_i, \log(\mu_{tni})_i, \log(\mu_v)_i, \log(\Pi)_i).$$

The $\epsilon_{ij}^{(l)}$ are the residual errors for the log viral load ($l = 1$) or the CD4$^+$ concentration ($l = 2$). We assume that the errors $\epsilon_{ij}^{(l)}$ are independent and normally distributed with a null mean and a variance $\sigma_{ij}^{(l)}$. We assume that the individual parameters ϕ_i are random vectors, independent of $\epsilon_{ij}^{(l)}$ and decomposed as $\phi_i = \beta + b_i$. where β is the population mean vector and b_i is the random effect of subject i, which is assumed to be normally distributed with zero mean and diagonal covariance matrix Ω.

On a practical ground, when viral load data $y_{ij}^{(1)}$ are inferior to the limit of quantification (LOQ), the exact value $y_{ij}^{(1)}$ is not known, instead, the observation is $y_{ij}^{(1)} \leq LOQ$. These data are classically named left-censored data. In this study, the limit of quantification was 50 cp/mL. This censoring of observed response presents therefore an additional challenge in the analysis of longitudinal HIV-1 data.

The random effects structure of non-linear mixed-effects models leads to intractable likelihoods, and therefore the maximum likelihood estimate is not available in closed form. We proposed to use the SAEM algorithm, a stochastic version developed by [Delyon et al., 1999, Kuhn and Lavielle, 2005] of the Expectation-Maximization algorithm introduced by [Dempster et al., 1977]. SAEM algorithm is a true maximum likelihood estimation method, for which convergence results are proved. [Donnet and Samson, 2006] proposed a version of this algorithm adapted to differential mixed models. The SAEM algorithm also enables to take into account the left-censored viral load data accurately [Samson et al., 2005]. Furthermore, this estimation method evaluates the expectation of the viral load values below the limit of quantification, conditionally of the observed values.

Viral load data and CD4$^+$ concentrations were modeled jointly in all patients. Models with random effects on all physiological parameters were built. Goodness-of-fit plots (population and individual predicted concentrations versus observed concentrations, weighted residuals versus predicted concentrations and versus time) were examined.

Results

Thirty-height HIV-1 infected patients were enrolled in this study and received lopinavir. Note that of the 38 subjects, 32 were included in this analysis; the remaining 6 subjects were not assessable at week 16. Seventy-one % of subjects were male. The mean age of subjects was 39 years. Fifteen were taking a 400/100 mg bid of lopinavir/ritonavir at week 16, that is they did not change their posology since visit W0, eleven patients were taking 266/66 mg bid at week 16 and six received 533/133 mg bid at week 16. Together with lopinavir, they received respectively 100 mg bid, 66 mg bid and 133 mg bid of ritonavir. Seventy-two % of patients receiving lopinavir were considered highly adherent according to their answers to the adherence questionnaire.

One hundred and ninety nine and one hundred and seventy seven concentrations of respectively the viruses and the $CD4^+$ (out of an anticipated 224 and 192 respectively) were obtained from the 32 patients. Individual viral load and $CD4^+$ evolutions are displayed in Figure 1. Eighty-five (42.7% of the total) viral load concentrations are below the limit of quantification of 50 cp/mL.

[Figure 1 about here.]

The dynamic model (S) was fitted to the viral load and the $CD4^+$ concentration data from patients in the lopinavir arm using the SAEM algorithm. The individual predictions are plotted against the individual observations for the logarithm of the viral load and the $CD4^+$ concentration on Figure 2.

[Figure 2 about here.]

The individual weighted residuals are plotted against the individual predictions for the log viral load and the $CD4^+$ concentration on Figure 3. These graphs show the adequacy of the final model with the data.

[Figure 3 about here.]

We reported the dynamic parameter estimates in Table 1 for the first model.

[Table 1 about here.]

The residual variances were estimated at a level of 0.51 for the log viral load and 0.09 for the $CD4^+$ concentrations.

We observed a large inter-patient variability in the estimates of all individual dynamic parameters (the coefficient of variation ranges from 1% to 114% for the different dynamic parameters), but especially for four parameters, α the activation rate of quiescent T cells, r the desactivation rate of quiescent T cells, μ_v the death rate of the HIV viruses and Π the rate of producted virus. For

instance, the smallest virus clearance rate (μ_v) was estimated as 23 day^{-1} with a corresponding half-life of 0.02 days or 0.48 hour, and the largest corresponded to a half-life of 0.0001 days or 0.14 minute. The smallest rate of virions produced per infected cell was 50.62 and the largest rate of virions produced per infected cell was 11801.

As an example, Figure 4 presents the model-fitting results for 4 subjects for both the viral load and the CD4$^+$ concentration.

[Figure 4 about here.]

For example, for patient 1, the PI efficacy was estimated at 0.97 corresponding to a high level of efficacy. Thus this patient had successful viral load and CD4$^+$ concentration trajectories.

The estimated dynamic parameters for the patients 8 and 17 are reported in Table 2.

[Table 2 about here.]

These two patients have initial similar viral load values (respectively 4.87 and 4.98 at W0, 2.41 and 2.52 at W2, and below LOQ for all the following measurements), therefore similar individual dynamic parameters should be expected. However, their CD4$^+$ concentrations are different. Consequently, by modeling conjointly the CD4$^+$ concentrations and the viral load, the estimation of some individual dynamic parameters differs between these two patients, particularly the virus clearance rate μ_V, the rate of activation of the quiescent T cells α or the production rate of CD4+ T cells λ. These differences explain the two different profiles predicted for the log viral load of the two subjects which are plotted on Figure 4.

Discussion

Most studies of HIV dynamics modeled only viral load data in a short period (2-6 weeks) after the initiation of an anti-retroviral treatment. In this article, we established the relation of virologic responses (viral load decrease and CD4$^+$ increase) to drug effect using a complex viral dynamic model. To estimate viral dynamic parameters, we used all the data (both viral load and CD4$^+$ measurements) obtained during 48 weeks in PI-naive patients starting a treatment with lopinavir. We showed the matter of estimate dynamic parameters from both biological markers. Indeed, for some patients with similar viral load trend, the simultaneous modeling of the CD4$^+$ increase reached to different individual dynamic parameters.

The estimation of the parameters of such a model is a very difficult statistical and computational challenge, never carried through up to now. [Putter et al., 2002] and [Wu et al., 2005] were compelled to assume simplifying hypothesis in order to afford the statistical estimation, leading to unsatisfactory models of the HIV viral dynamic. For example, [Putter et al., 2002] and

[Wu et al., 2005] adopted a Bayesian estimation approach because they failed to implement an efficient maximum likelihood estimation method. Several algorithms proposed to approximate the likelihood by linearization or using Laplace's approximation, such as the Lindstrom and Bates' algorithm implemented in the nlme function of Splus software (Insightful, Seattle, Washington) [Lindström and Bates, 1990, Pinheiro and Bates, 1995]. However, there is no published proof of any relevant statistical property for these algorithms, furthermore, they frequently fail to converge. To overcome this issue, we used the SAEM algorithm which is computationally efficient on this model. The analysis of such data is also complicated by the left-censoring of the viral load data due to the lower limit of detection of experimental devices. It has been proved that omitting to correctly handle this censored data provides biased estimates of dynamic parameters [Samson et al., 2005]. While [Putter et al., 2002] included their analysis in their algorithm, [Wu et al., 2005] did not realize it. We incorporated in the SAEM algorithm a new way to manage the left-censored data, and evaluated the expectation of the censored values conditioning on the observed values. Therefore the results obtained here in the field of virology and anti-retroviral therapy are important by providing reliable estimation of dynamic parameters.

We chose to distinguish three types of $CD4^+$ T cells (quiescent, activated infected or not) because the influence of latent T cells becomes important after a rather long time ranging from several weeks to months [Perelson et al., 1997, Perelson and Nelson, 1997, Callaway and Perelson, 2002, Mocroft et al., 2000]. As we fit data collected up to 48 weeks after the therapy start, this latent $CD4^+$ T cells reservoir is crucial to obtain a good fit. We tried a more simplified model with only four equations but failed to fit the data. We also distinguished infectious virions from non-infectious virions as proposed by [Perelson et al., 1996]. Thus the different mechanisms of action of NRTI and PI drug effects were modeled. The mechanism-based dynamic model is powerful and efficient to characterize relations of the anti-retroviral response to drugs. Long-term HIV dynamics can be reasonably modeled with this model. Dynamic parameters for individual subjects can be estimated using the SAEM algorithm.

Transformations are frequently used on the $CD4^+$ concentrations to allow the statistical model to give a more satisfactory fit to the observations. A variety of transformations including log, square root or fourth-root have been used by others authors [Taylor and Sy, 1994, Taylor and Law, 1998, Thiebaut et al., 2005] to achieve approximate Gaussian distributions and homogeneity of variance of the measurement error component. However, the interpretation of such transformations being complex, we chose not to present such models in this paper.

The estimated parameters showed the efficacy of the HAART including lopinavir as a PI. From the estimated viral dynamic parameters, the clearance rate of free HIV virus ($\mu_v = 228$ day^{-1}) was greater that those previously estimated in literature [Ho et al., 1995, Wei et al., 1995,

Perelson et al., 1996, Perelson et al., 1997, Wu et al., 1999, Wu et al., 2003, Perelson and Nelson, 1997], including the value estimated recently by [Wu et al., 2005] which was yet greater than the previous ones. This might be explained by the difference between the two populations of patients. While the patients in the COPHAR II trial were PI-naive, the data analyzed by [Wu et al., 2005] came from patients receiving the PI indinavir instead of lopinavir and who failed their first PI-containing regimen. The rate of quiescent T cells production was lower ($\lambda = 7.17$) than the one estimated by [Wu et al., 2005], but they used a simplified model of the T cells compartments. The death rate of productively infected cells (μ_{TI}) was smaller than those previously estimated in literature. Moreover, the inter-patient variabilities were large in all viral dynamic, especially for α the activation rate of quiescent T cells, r the desactivation rate of quiescent T cells, μ_v the death rate of respectively the HIV viruses and Π the rate of producted virus. This may indicate a significant heterogeneity in host characteristics and viral species in different hosts, suggesting the importance of individualized treatments.

It is important to point out some of the limitations of the model that we are using. This model does not recognize the fact that HIV undergoes rapid mutation in the presence of antiretroviral therapy. Of course, considering such phenomenons in the model may introduce many more parameters. We have attempted to keep the model itself as simple as possible, in order to keep its parameters estimation as clear as possible. We considered a constant treatment effect, however, the effect of antiviral treatment appears to change over time, probably due to pharmacokinetic processes variation, fluctuating patient adherence, the emergence of drug resistance mutations and/or other factors. [Huang et al., 2003, Wu et al., 2005] proposed viral dynamic models to evaluate antiviral response as a function of time-varying concentrations of drug in plasma, etc. A more elaborate model would thus promisingly include these additional extensions. Nevertheless, these limitations would not offset the major findings from our modeling approach, although further improvement may be brought.

Acknowledgements

The authors thank the scientific committee of the COPHAR II-ANRS111 trial for giving us access to the patients's viral load measurements, especially the coordinators Pr Dominique Salmon-Céron from Infectious Diseases Unit at the Cochin Hospital, Paris, Dr Xavier Duval from Infectious Diseases Unit at the Bichat-Claude Bernard Hospital, Paris, and the pharmacologist Pr Treluyer from Pharmacology Unit at the St Vincent de Paul Hospital, Paris. The authors would like to thank Jérémie Guedj, INSERM E03-38, for his insightful comments and suggestions for the choice of the HIV dynamic model.

References

[Callaway and Perelson, 2002] Callaway, D. and Perelson, A. (2002). HIV-1 infection and low steady state viral loads. *Bull. Math. Biol.*, 64:29–64.

[De Boer and Perelson, 1998] De Boer, R. and Perelson, A. (1998). Target cell limited and immune control models of HIV infection: a comparison. *J. Theor. Biol.*, 190:201–214.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27:94–128.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39:1–38.

[Ding and Wu, 2001] Ding, A. and Wu, H. (2001). Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics*, 2:13–29.

[Donnet and Samson, 2006] Donnet, S. and Samson, A. (2006). Estimation of parameters in mising data models defined by differential equations. *submitted*.

[Fitzgerald et al., 2002] Fitzgerald, A., DeGruttola, V., and Vaida, F. (2002). Modelling HIV viral rebound using non-linear mixed effects models. *Stat. Med.*, 21:2093–2108.

[Ho et al., 1995] Ho, D., Neumann, A., Perelson, A., Chen, W., Leonard, J., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–126.

[Huang et al., 2003] Huang, Y., Rosenkranz, S., and Wu, H. (2003). Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. *Math. Biosci.*, 184(2):165–186.

[Kuhn and Lavielle, 2005] Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49:1020–1038.

[Lindström and Bates, 1990] Lindström, M. and Bates, D. (1990). Nonlinear mixed-effects models for repeated measures data. *Biometrics*, 46:673–687.

[Mocroft et al., 2000] Mocroft, A., Miller, V., Chiesi, A., Blaxhult, A., Katlama, C., Clotet, B., Barton, S., and Lundgren, J. (2000). Virological failure among patients on HAART from accross europe: results from the eurosida study. *Antivir. Ther.*, 5:107–112.

[Nowak and Bangham, 1996] Nowak, M. and Bangham, C. (1996). Population dynamics of immune responses to persistent viruses. *Science*, 272(5258):74–79.

[Nowak and May, 2000] Nowak, M. and May, R. (2000). *Virus dynamics: mathematical principles of immunology and virology*. Oxford University Press.

[Perelson, 2002] Perelson, A. (2002). Modelling viral and immune system dynamics. *Nat. Rev. Immunol.*, 2:28–36.

[Perelson et al., 1997] Perelson, A., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D. (1997). Decay characteristics of HIV-1 infected compartments during combination therapy. *Nature*, 387:188–191.

[Perelson and Nelson, 1997] Perelson, A. and Nelson, P. (1997). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review*, 41:3–44.

[Perelson et al., 1996] Perelson, A., Neumann, A., Markowitz, M., Leonard, J., and Ho, D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271:1582–6.

[Pinheiro and Bates, 1995] Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the non-linear mixed-effect models. *J. Comput. Graph. Stat.*, 4:12–35.

[Putter et al., 2002] Putter, H., Heisterkamp, S., Lange, J., and de Wolf, F. (2002). A bayesian approach to parameter estimation in HIV dynamical models. *Stat. Med.*, 21:2199–214.

[Samson et al., 2005] Samson, A., Lavielle, M., and Mentré, F. (2005). The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *submitted*.

[Stilianakis et al., 1997] Stilianakis, N., Boucher, C., De Jong, M., Van Leeuwen, R., Schuurman, R., and De Boer, R. (1997). Clinical data sets of human immunodeficiency virus type 1 reverse transcriptase-resistant mutants explained by a mathematical model. *J. Virol.*, 71:161–168.

[Taylor and Law, 1998] Taylor, J. and Law, N. (1998). Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Stat. Med.*, 17:2381–2394.

[Taylor and Sy, 1994] Taylor, J.M.G., W. C. and Sy, J. (1994). A stochastic model for analysis of longitudinal aids data. *J. Amer. Stat. Assoc.*, 89:727–736.

[Thiebaut et al., 2005] Thiebaut, R., Jacqmin-Gadda, H., Babiker, A., and Commenges, D. C. C. (2005). Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat. Med.*, 24(1):65–82.

[Wei et al., 1995] Wei, X., Ghosh, S., Taylor, M., Johnson, V., Emini, E., Deutsch, P., Lifson, J., Bonhoeffer, S., Nowak, M., and Hahn, B. e. a. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510):117–122.

[Wu et al., 1998] Wu, H., Ding, A., and De Gruttola, V. (1998). Estimation of HIV dynamic parameters. *Stat. Med.*, 17:2463–85.

[Wu et al., 2005] Wu, H., Huang, Y., Acosta, E., Rosenkranz, S., Kuritzkes, D., Eron, J., Perelson, A., and Gerber, J. (2005). Modeling long-term hiv dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance. *J. Acquir. Immune. Defic. Syndr.*, 39:272–283.

[Wu et al., 1999] Wu, H., Kuritzkes, D., McClernon, D., Kessler, H., Connick, E., Landay, A., Spear, G., Heath-Chiozzi, M., Rousseau, F., Fox, L., Spritzler, J., Leonard, J., and Lederma, M. (1999). Characterization of viral dynamics in human immunodeficiency virus type 1-infected patients treated with combination antiretroviral therapy: relationships to host factors, cellular restoration, and virologic end points. *J. Infect. Dis.*, 179(4):799–807.

[Wu et al., 2003] Wu, H., Mellors, J., Ruan, P., McMahon, D., Kelleher, D., and Lederman, M. (2003). Viral dynamics and their relations to baseline factors and longer term virologic responses in treatment-naive HIV-1-infected patients receiving abacavir in combination with HIV-1 protease inhibitors. *J. Acquir. Immune. Defic. Syndr.*, 33(5):557–563.

[Wu and Zhang, 2002] Wu, H. and Zhang, J.-T. (2002). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. *Stat. Med.*, 21(23):3655–75.

[Wu, 2004] Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *J. Am. Stat. Assoc.*, 99:700–709.

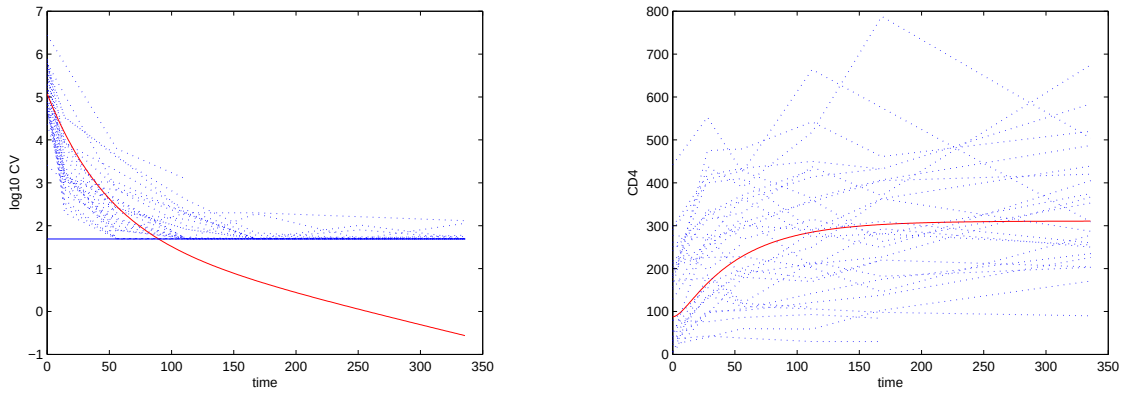


Figure 1: Population predicted curves (plain line) overlaid on the observed individual (doted lines) log viral load decreases (left) and $CD4^+$ increases (right) in the lopinavir group of the COPHAR II-ANRS111 trial. The left-censored viral load data are plotted at the LOQ value 1.69.

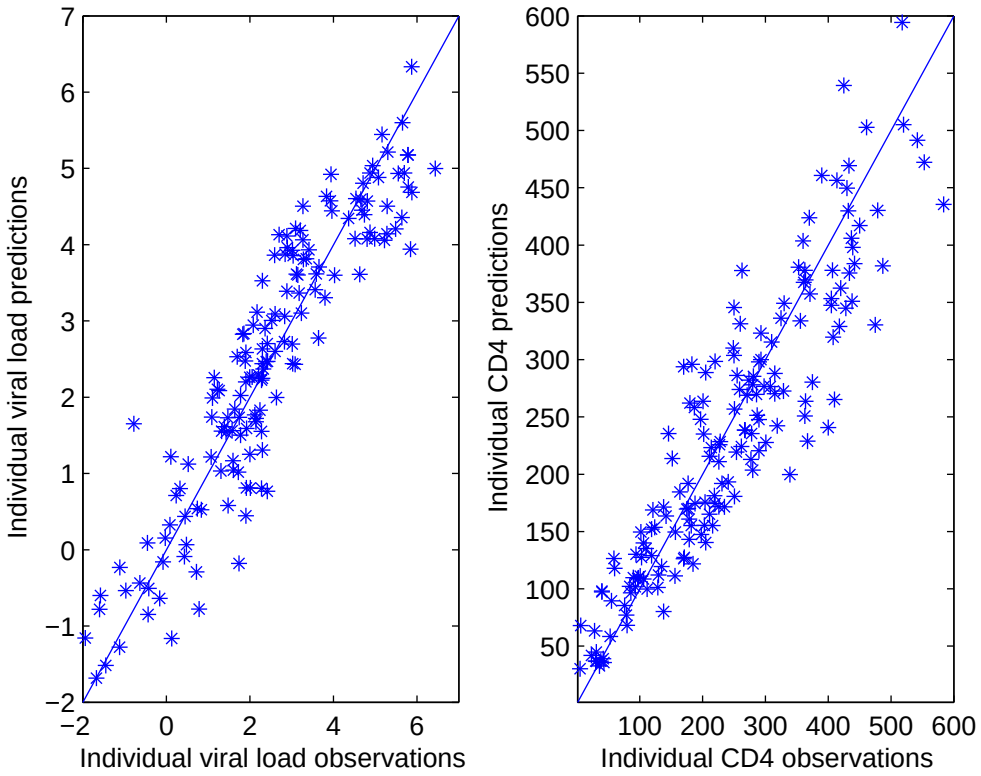


Figure 2: Individual predictions versus individual observations of the log viral load data (left) and the $CD4^+$ concentration (right).

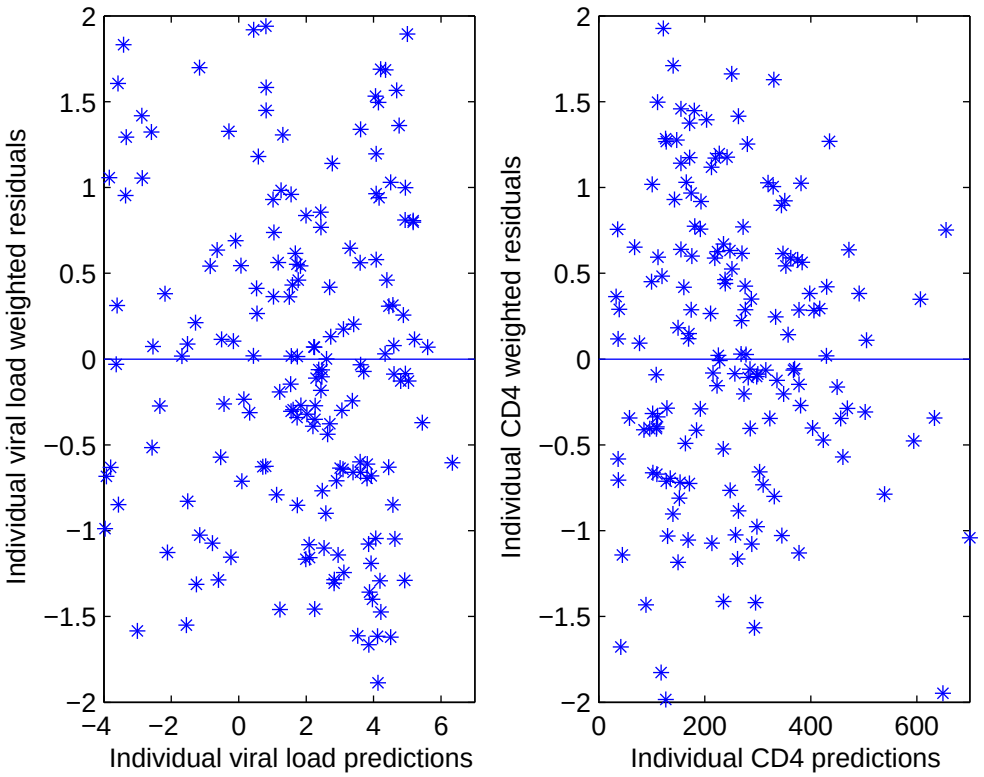


Figure 3: Individual weigthed residuals plotted against individual predictions of the log viral load data (left) and the $CD4^+$ concentration (right)

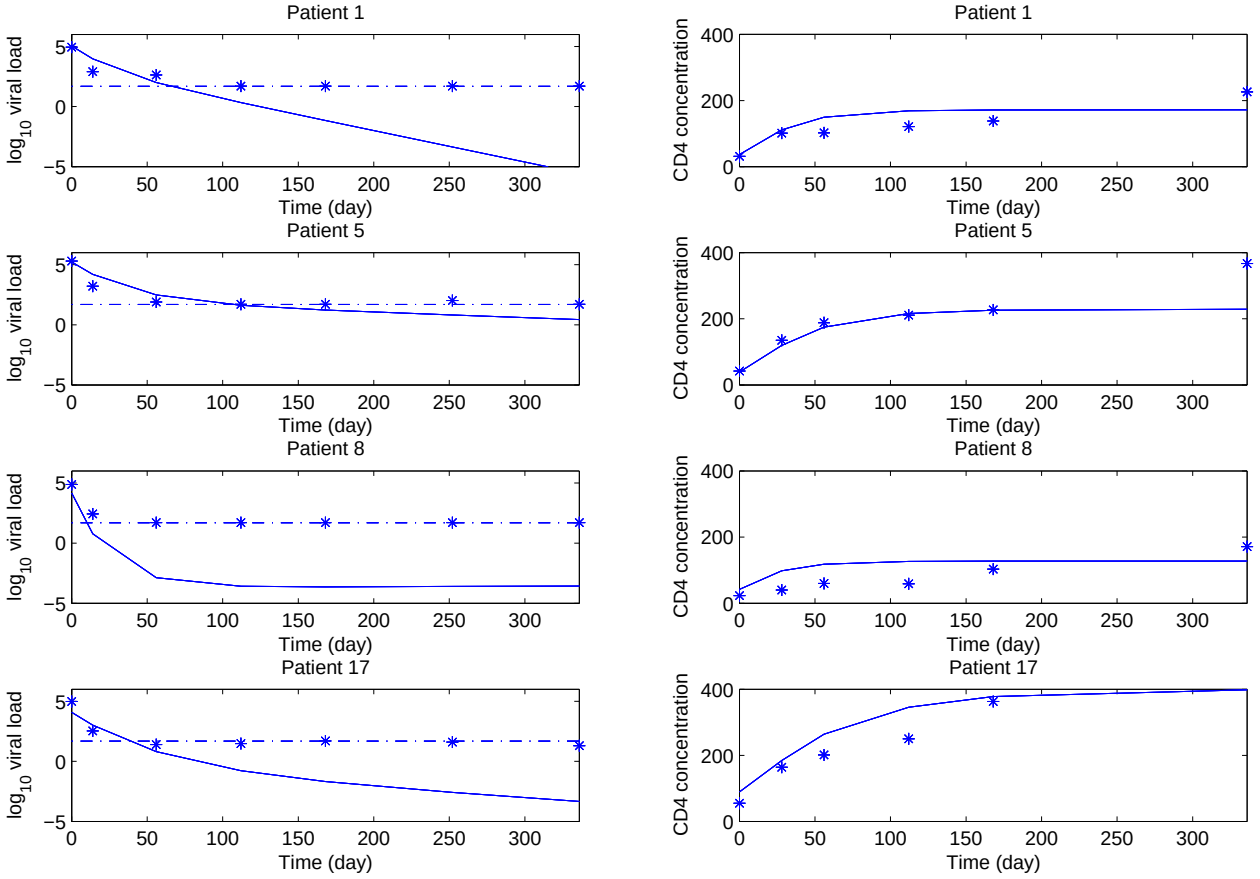


Figure 4: Observed and predicted individual log viral load decrease (left) and $CD4^+$ increase (right) for four subjects in the lopinavir group of the COPHAR II-ANRS111 trial.

Table 1: Estimates of the dynamic parameters and their inter-patient variability expressed as coefficient of variation (CV %) with the corresponding Standard Errors (SE) of estimation.

		fixed effect	(SE)	CV (%)	(SE)
α	(day^{-1} mm^{-3})	0.63	(0.11)	109	(0.37)
r	(day^{-1} mm^{-3})	2.43	(0.44)	114	(0.26)
μ_Q	(day^{-1})	0.01	($<10^{-3}$)	47	(0.06)
μ_{ti}	(day^{-1})	0.04	(0.01)	76	(0.12)
λ	(day^{-1} mm^{-3})	7.17	(0.57)	45	(0.05)
γ	(day^{-1} $cell^{-1}$ mm^{-3})	0.01	($<10^{-3}$)	60	(0.07)
η		0.86	(0.14)	18	(0.25)
μ_{tni}	(day^{-1})	0.15	(0.02)	56	(0.11)
μ_v	(day^{-1})	228.15	(39.6)	107	(0.25)
Π	(day^{-1} $cell^{-1}$ mm^{-3})	428.37	(83.6)	99	(0.28)

Table 2: Individual estimates of the dynamic parameters for patients 8 and 17.

		Patient 8	Patient 17
α	(day^{-1} mm^{-3})	0.41	8.07
r	(day^{-1} mm^{-3})	0.60	0.34
μ_Q	(day^{-1})	0.02	0.03
μ_{ti}	(day^{-1})	0.08	0.02
λ	(day^{-1} mm^{-3})	4.44	7.42
γ	(day^{-1} $cell^{-1}$ mm^{-3})	0.02	0.002
η		0.74	0.76
μ_{tni}	(day^{-1})	1.15	0.23
μ_v	(day^{-1})	35.16	774.24
Π	(day^{-1} $cell^{-1}$ mm^{-3})	179.45	506.04

Résultats complémentaires

Plusieurs auteurs recommandent d'analyser des transformations des concentrations de $CD4^+$, qui permettraient d'obtenir des résidus gaussiens plus satisfaisants qu'à partir d'une analyse des concentrations de $CD4^+$ non transformées. Par exemple Taylor et al. et Thiebaut et al. [73, 74, 63] proposent de considérer les logarithmes, les racines carrées, ou quatrièmes des concentrations de $CD4^+$. Sous ces hypothèses, un modèle d'erreur additif homoscédastique est utilisé sur les $CD4^+$ transformés.

Nous avons donc également réalisé l'analyse conjointe de la dynamique du virus et des $CD4^+$ en utilisant un modèle d'erreur homoscédastique sur la racine quatrième des $CD4^+$. Les évolutions prédites avec ce modèle pour la charge virale et les concentrations de $CD4^+$ sont présentées sur la figure 7.1.

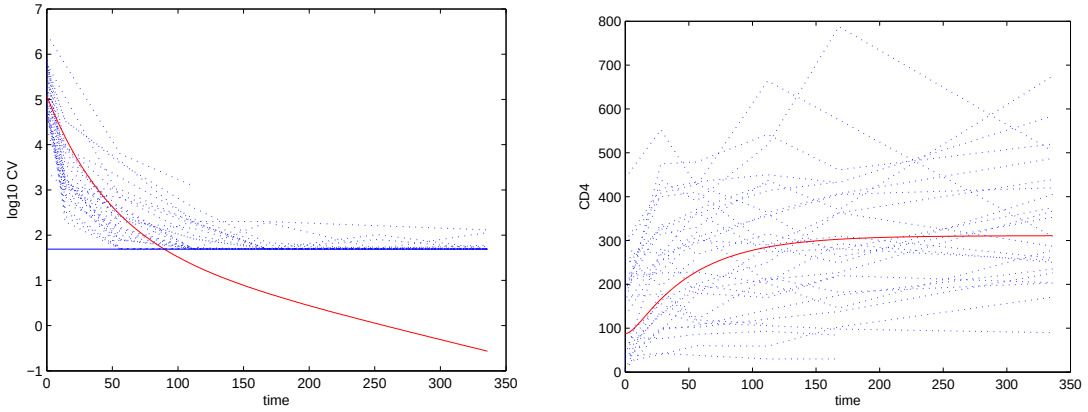


Figure 7.1: Courbes individuelles observées et courbes moyennes prédites pour les charges virales (gauche) et les concentrations de $CD4^+$ (droite) à partir d'un modèle d'erreur homoscédastique sur les racines quatrièmes des concentrations de $CD4^+$.

Les prédictions en fonction des observations des charges virales et des concentrations de $CD4^+$ sont tracées sur la figure 7.2. Certaines concentrations de $CD4^+$ sont sous-prédites, en particulier les grandes valeurs de concentrations. De même, les charges virales sont prédites de façon moins précise avec ce modèle qu'avec le modèle d'erreur hétéroscédastique présenté dans l'article. Ce modèle d'erreur semble donc moins satisfaisant que le précédent. Ces diagnostiques graphiques ne sont cependant pas suffisants pour discriminer ces deux modèles. Une comparaison par critère d'Akaike ou BIC pourrait être réalisée.

Néanmoins, la transformation des concentrations de $CD4^+$ étant difficile à interpréter biologiquement, nous nous sommes limités, dans l'article, à la présentation des résultats obtenus avec le modèle d'erreur hétéroscédastique sur les concentrations de $CD4^+$ non transformées.

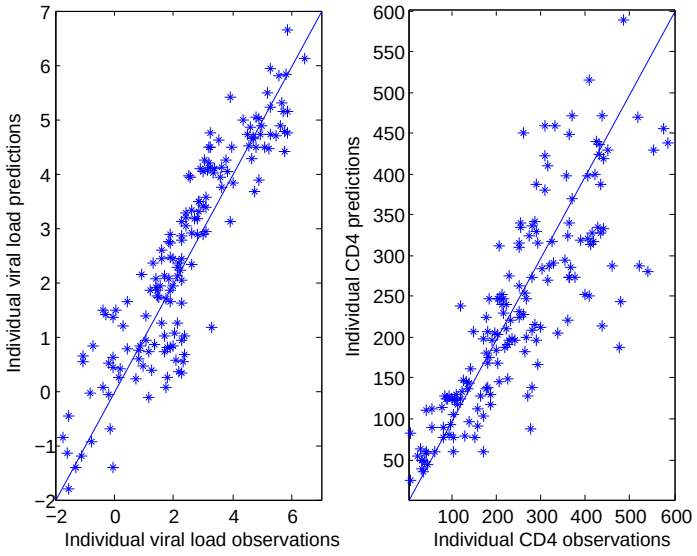


Figure 7.2: Prédictions individuelles en fonction des observations individuelles pour les charges virales (gauche) et les concentrations de $CD4^+$ (droite) à partir d'un modèle d'erreur homoscédastique sur les racines quatrièmes des concentrations de $CD4^+$.

Chapitre 8

Étude des interactions pharmacocinétiques

8.1 Contexte

Dans l'essai COPHAR II-ANRS 111 présenté dans le chapitre 7, les patients recevaient une combinaison de médicaments. Par exemple les patients sous inhibiteur de protéase lopinavir (combiné au ritonavir) recevaient également des analogues nucléosidiques comme la zidovudine et/ou la lamivudine. Il est donc important d'étudier les interactions pharmacocinétiques entre ces médicaments. D'autre part, il essentielle d'étudier et de modéliser la variabilité intra-patient de ces médicaments car certains auteurs ont montré qu'elle est en général particulièrement grande [64]. Ces deux types d'études reposent sur une méthodologie similaire, l'analyse de données de concentration recueillies lors de plusieurs périodes d'observations. Dans ce chapitre, nous nous sommes principalement intéressés à l'étude de l'interaction pharmacocinétique de deux molécules.

L'interaction pharmacocinétique entre deux médicaments se caractérise par la modification d'un paramètre pharmacocinétique de l'un d'entre eux, cette modification ayant une pertinence clinique. Par exemple, dans la classe des inhibiteurs de protéase, il a été montré que le ritonavir est moins efficace que l'indinavir. Cependant, différents auteurs [75, 76] ont montré que lorsque le ritonavir est administré avec l'indinavir, le ritonavir a pour effet de ralentir l'élimination de l'indinavir, c'est-à-dire qu'il augmente son exposition et donc son efficacité. Cette augmentation de l'efficacité de l'indinavir ne peut pas être obtenue en augmentant les doses de ce médicament sans risquer des problèmes de toxicités rénales, les seuils d'efficacité et de toxicité en terme de doses de cette molécule étant proches. Ainsi, la combinaison de l'indinavir et du ritonavir permet, avec des doses non toxiques, d'obtenir une efficacité augmentée de l'indinavir.

L'administration conjointe de deux molécules peut également avoir des effets

négatifs. C'est par exemple le cas pour deux inhibiteurs nucléosidiques de la transcriptase inverse, la zidovudine et la stavudine. En effet, il a été montré que ces deux médicaments utilisent la même voie d'activation. Quand ils sont administrés ensemble, il y a donc un risque de diminution de l'efficacité de chacun de ces médicaments par antagonisme compétitif dans leur voie d'activation, et l'administration conjointe de ces deux médicaments est proscrite [77].

Les interactions pharmacocinétiques sont étudiées en plusieurs étapes. Une première étape *in vitro* a pour but d'étudier biochimiquement les processus d'interaction. Ensuite, l'interaction est étudiée chez des patients au cours d'essais cliniques spécifiques, à plusieurs périodes observations. Les patients reçoivent lors de la première période un seul des deux médicaments, que nous appelons A dans la suite. Les patients sont ensuite suivis lors d'une deuxième période, pendant laquelle ils reçoivent les deux médicaments A et B simultanément. Dans ces essais, chaque patient est pris comme son propre témoin, ce qui permet de différencier la variabilité intra-patient de la variabilité résiduelle.

A partir des données recueillies au cours de ces essais spécifiques, une première analyse non-compartimentale permet de déterminer les paramètres pharmacocinétiques du médicament A seul ou de A co-administré avec B, et de les comparer. Les agences réglementaires américaine et européenne du médicament préconisent pour ces essais d'évaluer les interactions sur le logarithme de deux paramètres pharmacocinétiques : l'aire sous la courbe (AUC) et la concentration maximale (C_{max}). Dans un deuxième temps, une approche par modélisation permet de différentier les variabilités inter et intra patients de la variabilité résiduelle, d'identifier les facteurs cliniquement pertinents pouvant expliquer cette interaction (démographique, fonction rénale, ...) et enfin de montrer sur quels paramètres pharmacocinétiques se répercute l'interaction médicamenteuse.

Le but de ce chapitre est d'adapter l'algorithme SAEM à l'analyse de ces essais à plusieurs périodes d'observation et à la modélisation de ce niveau supplémentaire de variabilité. L'extension de l'algorithme SAEM à ce cadre est présentée dans la section 8.3. Nous l'avons ensuite évalué par simulation sur un essai pharmacocinétique à deux périodes d'observations, détaillée dans la section 8.4. Enfin, dans la section 8.5, nous avons utilisé cette méthode pour étudier l'interaction du ténofovir, un inhibiteur nucléotidique de la transcriptase inverse, sur la pharmacocinétique de l'atazanavir, un inhibiteur de protéase, en analysant des données de concentration recueillies lors de l'essai Puzzle 2-ANRS 107.

8.2 Modèles et notations

Dans un souci de clarté, nous nous limitons dans la suite de ce chapitre à un essai comprenant 2 périodes d'observations. La $j$ème concentration ($j = 1, \ldots, n_i$) du sujet i ($i = 1, \ldots, N$) pour la période k ($k = 1, 2$) prélevée au temps t_{ijk} est notée y_{ijk}. Notons $y_{ik} = (y_{ijk})_{1 \leq j \leq n_i}$, le vecteur d'observations du sujet i pour la k-ème période et $y = (y_{ijk})_{i,j,k}$ le vecteur de toutes les observations. Le modèle mixte s'écrit alors

$$\begin{aligned} y_{ijk} &= f(t_{ijk}, \phi_{ik}) + \varepsilon_{ijk} g(t_{ijk}, \phi_{ik}) \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, 1) \\ \phi_{ik} &= X_i \mu + b_i + c_{ik} \\ b_i &\sim \mathcal{N}(0, \Omega) \\ c_{ik} &\sim \mathcal{N}(\beta \mathbb{1}_{k=2}, \Gamma) \end{aligned} \tag{8.1}$$

où ϕ_{ik} est le vecteur de taille p des paramètres du sujet i pour la période k, ε_{ijk} est l'erreur de mesure, supposée indépendante de ϕ_{ik}, normalement distribuée, de moyenne nulle et de variance 1. Le modèle d'erreur peut être homoscédastique

$$g(t_{ijk}, \phi_{ik})^2 = \sigma^2,$$

ou hétéroscédastique

$$g(t_{ijk}, \phi_{ik})^2 = (a + \sigma f(t_{ijk}, \phi_{ik}))^2,$$

avec a une constante fixée et σ^2 la variance résiduelle.

On note μ la matrice des effets fixes, X_i le vecteur de covariables du sujet i, b_i est un effet aléatoire dépendant du sujet i, normalement distribué, de moyenne nulle et de matrice de variance-covariance Ω. Le vecteur c_{ik} est un effet aléatoire dépendant du sujet i *et* de la période k, qui est également normalement distribué et de matrice de variance-covariance Γ. Pour assurer l'identifiabilité des paramètres, on suppose que les effets aléatoires c_{i1} de la période 1 sont centrés, alors que les vecteurs c_{i2} de la période 2 sont de moyenne β, qui représente l'effet de l'interaction du traitement B sur le traitement A. La variance Ω représente la variabilité *inter-sujet* alors que Γ représente la variabilité *intra-sujet* ou *inter-occasion*.

Le vecteur des paramètres à estimer est $\theta = (\mu, \beta, \Omega, \Gamma, \sigma^2)$. On note $\psi_i = \mu + b_i$, $\psi = (\psi_1, \ldots, \psi_N)$, $c = ((c_{i1}, c_{i2})_{1 \leq i \leq N})$.

8.3 L'algorithme SAEM pour la modélisation de la variabilité intra-sujet

Le modèle (8.1) est un modèle à données non observées, le vecteur des données non observées étant $x = (\psi, c)$. La vraisemblance des données observées y s'écrit alors

$$p(y;\theta) = \int p(y,\psi,c;\theta)d\psi dc,$$

où $p(y,\psi,c;\theta)$ est la vraisemblance des données complètes (y,ψ,c). La vraisemblance des données complètes s'écrit

$$p(y,\psi,c;\theta) = p(y|\psi,c;\sigma^2)p(\psi;\mu,\Omega)p(c;\beta,\Gamma),$$

et sous les hypothèses de normalité des effets aléatoires, appartient à la famille exponentielle, c'est à-dire qu'elle s'écrit

$$p(y,\psi,c;\theta) = \exp\left\{-h(\theta) + \langle S(y,\psi,c), H(\theta)\rangle\right\}, \tag{8.2}$$

où h et H sont deux fonctions du paramètre inconnu θ, $\langle\cdot,\cdot\rangle$ est le produit scalaire, et $S(y,\psi,c)$ sont les statistiques suffisantes minimales du modèle complet. Ces statistiques comprennent, outre les statistiques déjà présentées dans le chapitre 4, les statistiques liées au deuxième niveau d'effet aléatoire : $\sum_{ik} c_{ik}$, $\sum_{ik} c_{ik}^2$, $\sum_i c_{i2}$.

Nous avons proposé une version de l'algorithme d'estimation SAEM adaptée à ce modèle. Les étapes d'Approximation Stochastique et de Maximisation sont adaptées aux nouvelles statistiques suffisantes du modèle. L'étape de Simulation est plus complexe. Nous avons proposé de simuler la chaîne de Markov (ψ, c), ayant pour unique loi stationnaire la loi $p(\psi, c|y;\theta_{(m-1)})$, par algorithme de Gibbs hybride. A l'itération m, la simulation de $(\psi^{(m)}, c^{(m)})$ est donc réalisée en combinant deux procédures de Metropolis-Hastings :

1. simulation de $c^{(m)}$ selon $p(.|y,\psi^{(m-1)};\theta_{(m-1)})$ par Metropolis-Hastings en utilisant les trois lois instrumentales suivantes :
   (a) la distribution a priori de c, c'est-à-dire une loi normale $\mathcal{N}(\widehat{\beta}_{(m-1)}\mathbb{1}_{k=2}, \widehat{\Gamma}_{(m-1)})$,
   (b) une loi normale multidimensionnelle générant une marche aléatoire $\mathcal{N}(c^{(m-1)}, \rho_c\widehat{\Gamma}_{(m-1)})$,
   (c) la succession de p lois normales unidimensionnelles générant une marche aléatoire, chaque composante de c étant actualisée successivement.
2. simulation de $\psi^{(m)}$ selon $p(.|y,c^{(m)};\theta_{(m-1)})$ par Metropolis-Hastings en utilisant les trois lois instrumentales suivantes :

(a) la distribution a priori de ψ, c'est-à-dire la loi normale $\mathcal{N}(\widehat{\mu}^{(m-1)}, \widehat{\Omega}_{(m-1)})$,

(b) une loi normale multidimensionnelle générant une marche aléatoire $\mathcal{N}(\psi^{(m-1)}, \rho_\psi \widehat{\Omega}_{(m-1)})$,

(c) la succession de p lois normales unidimensionnelles générant une marche aléatoire, chaque composante de ψ étant actualisée successivement.

Les paramètres ρ_ψ et ρ_c utilisés dans les marches aléatoires multidimensionnelles sont des paramètres d'échelle choisis de manière à obtenir un taux d'acceptation satisfaisant, c'est-à-dire proche de 30%.

8.4 Évaluation par simulation des propriétés de l'algorithme

Nous avons évalué les propriétés statistiques (biais et erreurs quadratiques moyennes) de cette extension de SAEM par simulation à partir de l'exemple de la pharmacocinétique de la théophylline, un anti-asthmatique, que nous avons déja utilisé dans le chapitre 3.3 pour illustrer les propriétés de l'algorithme SAEM dans le cadre des modèles dynamiques stochastiques.

8.4.1 Méthodes

Une dose orale de théophylline de 4 mg est administrée aux patients. Dix prélèvements sont réalisés 15 minutes, 30 minutes, 1, 2, 3.5, 5, 7, 9, 12 et 24 heures après la prise.

Un modèle à un compartiment avec absorption et élimination du premier ordre est utilisé pour décrire ce processus

$$f(\phi, t) = \frac{Dose \cdot k_a}{V k_a - Cl} \left(e^{-\frac{Cl}{V} t} - e^{-k_a t} \right) \tag{8.3}$$

où V est le volume de distribution, k_a le taux d'absorption et Cl la clairance d'élimination du médicament. Le vecteur ϕ des paramètres est composé de $\log(V)$, $\log(k_a)$ et $\log(AUC)$, où $AUC = Dose/Cl$.

Nous avons simulé 1000 jeux de données avec $N = 12$ sujets dans chaque groupe de traitement. Les valeurs des effets fixes étaient celles estimées sur le jeu de données réelles : $\mu = (-0.73, 0.39, 4.61)$. Les variabilités inter-patient ont été choisies comme suit : 10% pour le volume $\log(V)$ et 20% pour $\log(k_a)$ et $\log(AUC)$. Les variabilités intra-patient ont été choisies comme étant égales à la moitié des variabilités inter-patient pour chacun des trois paramètres. Nous avons choisi un modèle d'erreur

	Biais		RMSE	
	SAEM	`nlme`	SAEM	`nlme`
Effet fixe	-0.04	-0.16	0.04	2.10
Effet traitement	-0.09	0.03	0.07	1.09
Variance inter-patient	-14.55	-12.00	15.11	47.71
Variance intra-patient	-1.86	14.01	7.33	68.86
Variance résiduelle	-1.71	55.44	1.88	61.57

Table 8.1: Biais et RMSE relatifs (%) sur les paramètres estimés pour $\log(AUC)$ avec SAEM et `nlme`

hétéroscédastique combiné

$$g(\phi_{ik}, t_{ij})^2 = (1 + \sigma f(\phi_{ik}, t_{ij}))^2,$$

simulé avec une variance $\sigma^2 = 0.01$. Nous n'avons pas simulé d'effet interaction, ce qui correspond à un vecteur β nul.

Nous avons analysé les 1000 bases simulées avec l'algorithme SAEM et la fonction `nlme` de R, implémentant un algorithme FOCE. Pour chaque jeu de données, nous avons estimé les effets fixes $\log(V/F)$, $\log(k_a)$ et $\log(AUC)$, ainsi que les variances inter- et intra- patient sur ces trois paramètres. Nous avons également estimé un effet traitement sur chaque effet fixe. La variance résiduelle σ^2 est estimée. Nous avons calculé les biais et les erreurs moyennes quadratiques moyens obtenus par les deux algorithmes d'estimation pour l'ensemble des paramètres.

8.4.2 Résultats

Les résultats obtenus par l'algorithme SAEM et la fonction `nlme` en termes de biais relatifs et de RMSE relatives se rapportant au paramètre $\log(AUC)$ sont présentés dans le tableau 8.1.

Les biais de la variance intra-patient, et de la variance résiduelle sont nettement plus faibles avec l'algorithme SAEM qu'avec la fonction `nlme`. Les biais des autres paramètres sont du même ordre pour les deux méthodes. Les RMSE sur les variances sont nettement améliorés avec SAEM. Ces résultats illustrent les bonnes propriétés statistiques de l'algorithme SAEM étendu à ce cadre.

8.5 Application à l'étude de l'interaction du ténofovir sur la cinétique de l'atazanavir

8.5.1 Essai Puzzle 2 - ANRS 107

L'essai Puzzle 2 - ANRS 107 est un essai prospectif, ouvert, randomisé, réalisé chez des patients infectés par le VIH-1 en échec thérapeutique. Une sous-étude pharmacocinétique a été effectuée au début de l'essai chez 11 d'entre eux. Tous les patients avaient signé un consentement éclairé avant de participer à l'essai et à la sous-étude pharmacocinétique, qui avaient préalablement été approuvés par le Comité Consultatif de Protection des Personnes se prêtant à la Recherche Biomédicale de l'Hôpital Saint-Antoine (Paris). Les patients éligibles devaient répondre aux critères suivants : traitement anti-rétroviral stable depuis un mois, charge virale HIV-1 supérieure à 10 000 copies/ml, échec documenté du précédent traitement anti-rétroviral contenant aux moins deux inhibiteurs de protéase et un analogue non-nucléosidique de la transcriptase inverse, ainsi que l'absence de cardiomyopathie ou de maladie du système de conduction cardiaque.

Les patients ont été randomisés dans deux groupes, le groupe 1 continuant à prendre le même traitement qu'avant l'inclusion pendant deux semaines, et le groupe 2 recevant de l'atazanavir (300 mg en une prise quotidienne) combiné au ritonavir (100 mg en une prise quotidienne) à la place des inhibiteurs de protéase reçus avant l'inclusion.

A partir de la troisième semaine, tous les patients recevaient de l'atazanavir (300 mg en une prise quotidienne) combiné avec du ritonavir (100 mg en une prise quotidienne). Ils recevaient également un inhibiteur nucléotidique de la transcriptase inverse, le tenofovir disoproxil fumarate (DF) (300 mg en une prise quotidienne) et d'autres analogues nucléosidiques choisis individuellement en fonction du profil de résistance du virus isolé chez chaque patient.

La sous-étude pharmacocinétique de cet essai s'est déroulée chez 11 patients du groupe 2. L'objectif de cette étude était d'estimer les paramètres pharmacocinétiques de l'atazanavir (administré avec le ritonavir) avant ou après l'initiation du traitement par le ténofovir DF, afin d'évaluer s'il existe une interaction du ténofovir sur la cinétique de l'atazanavir. Le but de cette étude était d'obtenir une modélisation complète et satisfaisante de ces données avec l'algorithme SAEM.

Nous nous sommes intéressés aux données de concentration obtenues dans cette sous-étude.

8.5.2 Prélèvements pharmacocinétiques et mesure des concentrations

Des prélèvements sanguins destinés à la détermination des concentrations d'atazanavir ont été réalisés à la semaine 2, c'est-à-dire avant l'initiation du traitement par le ténofovir DF, et à la semaine 6, c'est-à-dire après l'initiation du traitement. Ces prélèvements ont eu lieu avant la première dose du matin puis 1, 2, 3, 5, 8 et 24 heures après l'administration du traitement. Le délai exact entre la prise du traitement et le prélèvement sanguin a été recueilli pour ces prélèvements prévus après la dose du matin. Pour le prélèvement effectué avant la dose du matin, le délai avec la dose du soir précédent était évalué à partir de l'heure de prise rapportée par le patient et l'horaire exact du prélèvement. Ce délai étant sujet à une importante incertitude, nous avons décidé d'exclure la concentration correspondante de l'analyse. L'un des onze patients a été exclu de l'étude après le deuxième semaine en raison de la survenue d'extrasystoles. Il a par conséquent été exclu de l'analyse. Les données aux deux visites sont présentées sur la Figure 8.1.

8.5.3 Modèle de pharmacocinétique de population

Nous avons utilisé un modèle à un compartiment avec une absorption d'ordre zéro et une élimination d'ordre un, se plaçant à l'équilibre après l'administration de plusieurs doses de médicament. Les paramètres de ce modèle sont la biodisponibilité F, la durée d'absorption T_a, le volume de distribution V et la constante d'élimination k_e. En utilisant un délai τ de 24 heures entre chaque prise de médicament, l'équation du modèle est la suivante

$$\frac{FDose}{T_a V k_e}\left((1-e^{-k_e t})\mathbb{1}_{t<T_a} + \frac{e^{-k_e\tau\mathbb{1}_{t<T_a}}(1-e^{-k_e T_a})e^{-k_e(t-T_a)}}{(1-e^{-k_e\tau})}\right)$$

Afin de tester la présence d'une interaction pharmacocinétique du ténofovir sur l'exposition des patients à l'atazanavir, le modèle a été reparamétré en T_a, V et AUC en utilisant $AUC = FDose/k_eV$. L'atazanavir étant administré par voie orale, les paramètres identifiables sont T_a, V/F et AUC. Afin de garantir la positivité des paramètres estimés, nous avons considéré les logarithmes de T_a, V/F et AUC, le vecteur de paramètres de ce modèle est donc $\phi = (\ln T_a, \ln V/F, \ln AUC)$.

Nous avons analysé avec l'algorithme SAEM les données de concentrations obtenues aux deux visites pharmacocinétiques. Nous avons estimé la variabilité inter- et intra-patient sur les trois paramètres pharmacocinétiques. Nous avons également testé par test de Wald l'influence de l'association avec le ténofovir sur les trois effets fixes du modèle. Nous avons utilisé un modèle d'erreur homoscédastique.

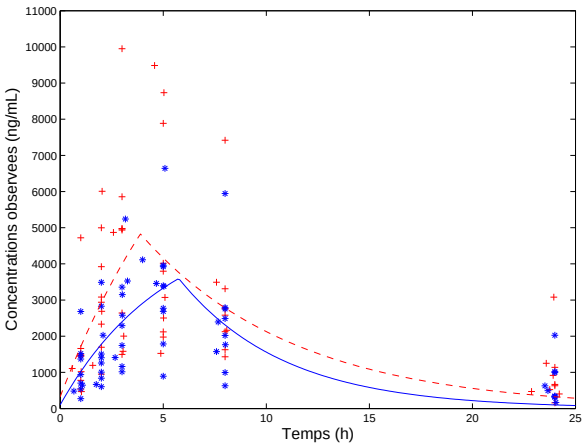


Figure 8.1: Courbes prédites et concentrations observées de l'atazanavir en l'absence de ténofovir (courbe $--$, observations $+$) et en présence du ténofovir (courbe —, observations $*$) chez 10 patients de l'essai Puzzle 2 - ANRS 107.

	V/F	AUC	T_a
log effet fixe	3.92	10.70	1.36
effet interaction	0.02	-0.38	0.39
CV (%) inter-patient	0.00	48.68	0.00
CV (%) intra-patient	53.30	0.41	14.05

Table 8.2: Paramètres estimés à partir des données de concentration d'Atazanavir

8.5.4 Résultats

Les estimations des paramètres et des variances figurent dans le tableau 8.2. La variance résiduelle a été estimée à $\sigma = 737$ ng/mL. La prise de tenofovir a un effet significatif sur les paramètres pharmacocinétiques $\log(AUC)$ ($p < 10^{-4}$) et $\log(T_a)$ ($p < 0.0025$) de l'atazanavir, engendrant une baisse de 1.46 de l'AUC, et une augmentation de 1.47 du temps d'absorption. Cet effet du ténofovir sur la

pharmacocinétique de l'atazanavir est visible sur les courbes de population prédites, tracées sur la figure 8.1.

Les graphiques d'adéquation du modèle, présentés dans la figure 8.2 sont satisfaisants.

8.5.5 Conclusion

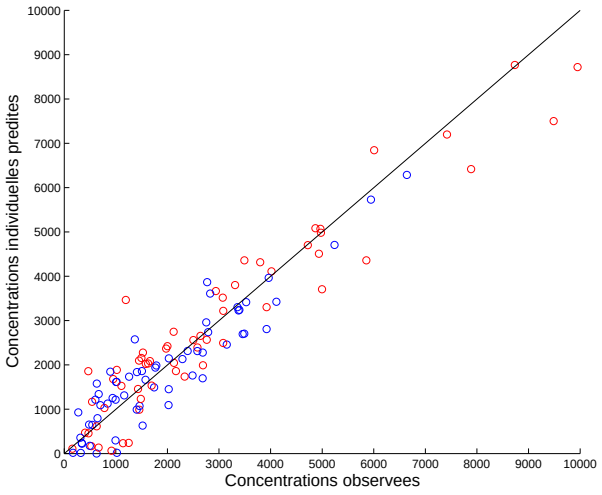
Le développement de l'algorithme SAEM nous a permis d'analyser les concentrations d'atazanavir obtenues aux deux visites pharmacocinétiques de l'essai Puzzle 2 - ANRS 107 en évaluant l'influence du ténofovir DF sur les trois paramètres pharmacocinétiques du modèle. Nos résultats confirment les conclusions de l'analyse non compartimentale des données de cet essai, où une diminution significative de l'AUC de l'atazanavir avait été observée lorsque les patients prenaient également du ténofovir. L'agence du médicament américaine FDA recommande désormais de modifier les doses d'atazanavir si ce dernier doit être administré avec le ténofovir. Nous avons d'autre part pu estimer les variabilités inter- et intra-patient des trois paramètres pharmacocinétiques.

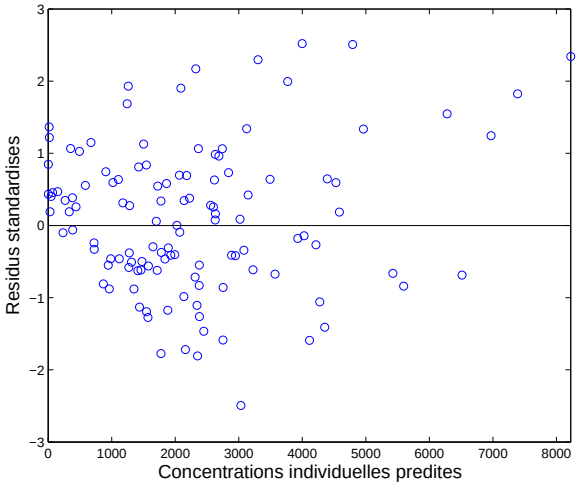
8.6 Discussion

Nous avons proposé dans ce chapitre une extension de l'algorithme SAEM adaptée à l'estimation de la variabilité intra-patient dans les modèles non-linéaires à effets mixtes. Nous avons montré par simulation les propriétés satisfaisantes de cet algorithme, en particulier par rapport à la fonction `nlme`, classiquement utilisée dans ce genre d'analyses.

De plus, l'application à l'étude de l'interaction du ténofovir sur la cinétique de l'atazanavir a montré que cette méthode permettait d'estimer les variabilités inter- et intra-patient sur l'ensemble des paramètres pharmacocinétiques du modèle, ce qui n'était pas possible en analysant les mêmes données avec le logiciel `nlme`.

La méthode d'estimation SAEM est donc particulièrement adaptée à l'analyse de données de concentrations obtenues chez des patients pour l'étude d'interaction pharmacocinétique. Cet algorithme peut également être utilisé pour l'estimation de la variabilité intra-patient des paramètres pharmacocinétiques de médicament étudiés au cours de plusieurs observations, y compris en l'absence d'études d'interaction. Il en est de même pour l'analyse de données recueillies au cours d'essais de bioéquivalence à plusieurs périodes d'observations, dont le but est de montrer l'équivalence pharmacocinétique entre deux formulations différentes (comprimé, soluble, ...) d'une même molécule.


Concentrations individuelles predites
Concentrations observees
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


Residus standardises
Concentrations individuelles predites
−3
−2
−1
0
1
2
3
0
1000
2000
3000
4000
5000
6000
7000
8000

Chapitre 9

Conclusion générale et perspectives

Ce travail de thèse est consacré aux développements de méthodes d'estimation permettant l'évaluation de la dynamique virale sous traitement dans l'infection par le VIH. Cette évaluation repose sur l'analyse de données longitudinales par modèles non-linéaires à effets mixtes. Les méthodes d'estimation que nous avons développées reposent sur l'algorithme SAEM, dont la convergence a été démontrée par Delyon et al. [43] et Kuhn et Lavielle [30]. Ces extensions rendent notamment possible l'estimation des paramètres d'un modèle non-linéaire mixte défini par équations différentielles (ordinaires ou stochastiques), la prise en compte des données censurées par une limite de quantification, ou encore la modélisation d'un niveau supplémentaire de variabilité (par exemple la variabilité inter-occasion). Ces extensions sont appuyées par des développements mathématiques, et les propriétés de convergence de l'algorithme SAEM ont été étendues à ces différents cas.

Nous avons illustré au cours de cette thèse les capacités de l'algorithme SAEM à estimer les paramètres de modèles complexes à partir de jeux de données réelles. Cet algorithme s'affranchit des problèmes de convergence extrêmement lente rencontrés avec les autres algorithmes d'estimation exacte tels que MC-EM, SA-NR, etc. La version « approximation stochastique »de l'algorithme EM semble donc la plus performante numériquement. D'autre part, l'utilisation des logiciels NONMEM et `nlme` engendre des problèmes contraignants de choix des points initiaux des paramètres, les résultats de convergence de ces algorithmes étant fortement dépendant de ces choix. Le caractère stochastique de SAEM permet à partir de tout point initial d'explorer l'espace des paramètres et de converger rapidement vers un voisinage du maximum de vraisemblance, évitant ainsi les problèmes de convergence. Cette méthode d'estimation est donc extrêmement prometteuse, tant au niveau des résultats théoriques de convergence qu'au niveau algorithmique. Son extension à l'analyse de données longitudinales binaires ou catégorielles serait d'un grand intérêt pour les applications en recherche biomédicale. Cependant ces modèles sortent du cadre de la famille exponentielle, en dehors duquel les résultats de convergence ne sont pas

établis.

L'algorithme SAEM dans sa forme initiale est implémenté dans le logiciel MONOLIX[1] qui a été développé au sein du groupe de travail Monolix par Marc Lavielle. L'intérêt de ce logiciel a été démontré lors d'une comparaison avec les différents logiciels disponibles (NONMEM, `nlme`, SAS, ...) présentée au congrès « Population Approach Group in Europe »de 2005. Le logiciel MONOLIX a obtenu les meilleurs résultats en terme de qualité et de précision d'estimation et les temps de calcul pour ce logiciel sont également parmi les plus bas parmi les méthodes d'estimation exacte[2]. Ces résultats ont confirmé ceux que nous avions obtenu et qui sont présentés dans le chapitre 4. Ce logiciel a déja été utilisé pour l'analyse de plusieurs études, en pharmacocinétique [78, 79] comme en agronomie [80]. Trois des développements de cette thèse (modèles définis par équations différentielles ordinaires, modèles avec données censurées, modèles avec une variabilité inter-occasion) vont être intégrés à la prochaine version du logiciel MONOLIX, facilitant ainsi leur diffusion et leur utilisation pratique.

Actuellement, le logiciel MONOLIX nécessite de la part de l'utilisateur le choix (délicat) du nombre d'itérations de l'algorithme. Pour des algorithmes déterministes tels que celui de Newton-Raphson, la convergence est obtenue dès lors qu'entre deux itérations successives, la différence entre les estimateurs des paramètres est plus petite qu'un seuil prédéfini. Cependant, pour un algorithme stochastique, cette condition peut être remplie purement par hasard et ne peut donc pas être utilisée comme critère d'arrêt ou de convergence. Dans le cas de l'algorithme EM, il serait par exemple possible d'utiliser une règle d'arrêt reposant sur sa propriété fondamentale d'accroissement de la vraisemblance, que nous avons rappelé en introduction de cette thèse. Cependant, dans le cas des modèles non-linéaires mixtes, la vraisemblance n'étant pas évaluable directement, le critère d'arrêt pourrait reposer sur une différence inférieure à un seuil prédéfini, entre deux valeurs successives de la fonction Q évaluée par exemple par approximation de Monte Carlo. Cependant, en raison du caractère stochastique de l'algorithme SAEM, cette différence devrait être calculée sur plusieurs itérations successives, afin d'éviter un arrêt prématuré dû à l'aléatoire.

Nous avons montré que l'efficacité des traitements anti-rétroviraux VIH peut être évaluée à l'aide de tests statistiques fondés sur l'analyse des données longitudinales de décroissance de charge virale. En effet, les tests que nous avons proposés ont des propriétés statistiques satisfaisantes. C'est en particulier le cas en présence de données de charge virale censurées, comme nous l'avons montré dans le chapitre 5. L'utilisation de l'extension de l'algorithme SAEM dans l'analyse de l'essai clinique randomisé TRIANON-ANRS 81 a mis en évidence une meilleure réponse des patients à l'un des deux traitements comparés sur la première pente de la décroissance, ce qui

[1] Téléchargeable sur le site http ://math.u-psud.fr/∼lavielle/monolix
[2] http ://www.page-meeting.org/page/page2005/PAGE2005008.pdf

n'a pas pu être montré par une approche de modélisation prenant classiquement en compte les données censurées. Nous avons ainsi retrouvé par modélisation la supériorité du traitement comprenant la lamivudine sur celui comprenant la nevirapine, qui avait été montré par Launay et al [66]. Cependant, nous n'avons analysé que les données disponibles, en particulier, nous n'avons pas pris en compte les sorties d'étude alors que 22% et 42% des patients des bras lamivudine et nevirapine respectivement, avaient arrêté leur traitement. Une analyse prenant en compte ces sorties d'étude prolongerait cette étude.

Nous nous sommes ensuite intéressés à l'étude plus globale de la dynamique virale, comprenant l'analyse simultanée de la décroissance de la charge virale et de la croissance de la concentration de lymphocytes $CD4^+$. Pour ce faire, nous avons développé une extension de l'algorithme SAEM adapté à l'estimation des paramètres d'un modèle mixte défini par un système différentiel. Dans un souci d'optimisation du temps de calcul, argument important pour son application pratique, nous avons proposé un nouveau schéma de résolution numérique de systèmes différentiels ordinaires. Ce schéma a été indispensable pour l'application à l'analyse des données de l'essai COPHAR II-ANRS 102. Une étude approfondie des données de cet essai est encore à réaliser. L'effet de différentes covariables sur l'efficacité des traitements sera testée à partir des estimations a posteriori des paramètres individuels du modèle. Les différentes covariables pourront être intégrées directement dans le modèle mixte, ce qui permettra, en particulier, une comparaison simultanée des données des trois bras de traitement de l'essai clinique COPHAR II-ANRS 102.

Enfin, nous nous sommes intéressés à l'étude de la pharmacocinétique des traitements anti-rétroviraux, en particulier à leur variabilité intra-patient. Cette variabilité peut être estimée à partir d'essais réalisés en plusieurs périodes d'observations. L'interaction d'un médicament sur la pharmacocinétique d'un autre médicament est également étudiée à travers le suivi pharmacocinétique de patients au cours de plusieurs périodes d'observations. Nous avons donc étendu l'algorithme SAEM à ce cadre. L'utilisation de cette méthode a permis de montrer une interaction du ténofovir sur la pharmacocinétique de l'atazanavir à partir de données de l'essai PUZZLE II-ANRS 107. L'utilisation de cet algorithme pour l'étude de la bioéquivalence pharmacocinétique de deux formulations d'une même molécule est une autre application de ce travail. En effet, des essais comprenant plusieurs périodes d'observations successives sont également utilisés pour étudier l'équivalence biologique de nouvelles formulations de molécules (dosages plus élevés, solution buvable, générique, etc.) avec les formulations actuelles.

Afin d'encourager l'utilisation des modèles non-linéaires à effets mixtes dans les essais cliniques, nous avons proposé une méthode de calcul du nombre de sujets nécessaires permettant la planification d'essais testant une différence d'effet entre

deux traitements et reposant sur cette méthode d'analyse. Cet outil devrait permettre d'inciter les instances réglementaires à recommander cette approche de modélisation au lieu des approches actuellement utilisées et homologuées reposant sur une analyse de la dernière valeur recueillie, et ne tenant pas compte de l'ensemble des données. La méthode de calcul du nombre de sujets repose sur une approximation de Monte-Carlo de la matrice de Fisher attendue via la simulation d'un jeu de données comportant plusieurs milliers de sujets. Pour utiliser cette méthode, il faut néanmoins connaître préalablement le modèle non-linéaire utilisé, les valeurs des paramètres, le paramètre sur lequel l'effet traitement est attendue, etc. Cette méthodologie est facilement généralisable au calcul du nombre de sujets nécessaires à inclure pour assurer une puissance donnée dans le cas d'essais à plusieurs périodes d'observations (qui comprennent un niveau de variabilité supplémentaire) comme les essais pharmacocinétiques d'interaction ou de bioéquivalence. Cette méthode est également généralisable au cas de la planification d'un essai évaluant simultanément la pharmacocinétique et la pharmacodynamie d'un médicament.

Pour ces approches, il est important de proposer un protocole de prélèvements conduisant aux plus faibles erreurs d'estimation possibles sur les paramètres du modèle, comme le permet la méthode PFIM-OPT développée par Retout et Mentré [65]. Cette approche est pour le moment basée sur la linéarisation du modèle. Une méthode d'optimisation fondée sur un calcul exact de la matrice de Fisher pourrait être envisagée. Amzal et al. [81] ont par exemple proposé l'optimisation de plans d'expérience de ces modèles en utilisant des méthodes particulaires dans le cadre bayesien.

La modélisation de la dynamique virale à laquelle nous nous sommes spécifiquement intéressés dans cette thèse n'est pas propre à l'infection par le VIH. Plusieurs systèmes similaires ont été proposés pour modéliser la dynamique des infections par le virus des hépatites B et C [82, 83]. La méthodologie développée dans cette thèse pourrait donc être appliquée à ces domaines. De même, l'évolution de la taille de tumeurs cancéreuses peut être modélisée à partir de systèmes dynamiques représentant la multiplication des cellules tumorales, la vascularisation de la tumeur, ainsi que l'effet de traitements radiothérapeutiques ou chimiothérapeutiques [84]. Le travail réalisé dans cette thèse pourrait donc servir pour la modélisation des données de cet autre domaine thérapeutique, présentant un très grand intérêt. Ces modèles pourront également inclure comme covariables des données issues de l'analyse génomique de tumeurs, obtenues par exemple par analyse par puces à ADN.

Bibliographie

[1] A.S. Fauci. HIV and AIDS : 20 years of science. *Nat Med*, 9(7) :839–843, 2003.

[2] A.S. Perelson, A.U. Neumann, M. Markowitz, J.M. Leonard, and D.D. Ho. HIV-1 dynamics in vivo : virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271 :1582–1586, 1996.

[3] M.A. Nowak and C.R. Bangham. Population dynamics of immune responses to persistent viruses. *Science*, 272 :74–79, 1996.

[4] A.S. Perelson, P. Essunger, Y.Z Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, and D.D. Ho. Decay characteristics of HIV-1 infected compartments during combination therapy. *Nature*, 387 :188–191, 1997.

[5] H. Jacqmin-Gadda, R. Thiebaut, G. Chene, and D. Commenges. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, 1 :355–68, 2000.

[6] A.A. Ding and H. Wu. Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics*, 2 :13–29, 2001.

[7] H. Wu and L. Wu. Identification of significant host factors for HIV dynamics modeled by non-linear mixed-effects models. *Stat. Med.*, 21 :753–71, 2002.

[8] H. Putter, S.H. Heisterkamp, J.M. Lange, and F. de Wolf. A bayesian approach to parameter estimation in HIV dynamical models. *Stat. Med.*, 21 :2199–214, 2002.

[9] R. Thiébaut, H. Jacqmin-Gadda, C. Leport, C. Katlama, D. Costagliola, V. Le Moing, P. Morlat, G. Chêne, and APROCO Study Grooup Chêne. Analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left censoring of HIV RNA measures. *J. Biopharm. Stat.*, 13 :271–82, 2003.

[10] M.O. Karlsson and L.B. Sheiner. The importance of modeling interoccasion variability in population pharmacokinetic analyses. *J. Pharmacokinet. Biopharm.*, 21 :735–750, 1993.

[11] J.S. Barrett, L. Labbé, and M. Pfister. Application and impact of population pharmacokinetics in the assessment of antiretroviral pharmacotherapy. *Clin. pharmacokinet.*, 44 :591–625, 2005.

[12] L.B. Sheiner and J.L. Steimer. Pharmacokinetic/pharmacodynamic modeling in drug development. *Annu. Rev. Pharmacol. Toxicol.*, 40 :67–95, 2000.

[13] S.L. Beal and L.B. Sheiner. Estimating population kinetics. *Crit. Rev. Biomed. Eng.*, 8 :195–222, 1982.

[14] M.J. Lindström and D.M. Bates. Nonlinear mixed-effects models for repeated measures data. *Biometrics*, 46 :673–687, 1990.

[15] R Wolfinger. Laplace's approximations for non-linear mixed-effect models. *Biometrika*, 80 :791–795, 1993.

[16] E. F. Vonesh. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83 :447–452, 1996.

[17] Z. Ge, P.J. Bickel, and J.A. Rice. An approximate likelihood approach to nonlinear mixed effects models via spline approximation. *Comput. Statist. Data Anal.*, 46 :747–776, 2004.

[18] U. Wahlby, E.N. Jonsson, and M.O. Karlsson. Assessment of actual significance levels for covariate effects in NONMEM. *J. Pharmacokinet. Pharmacodyn.*, 28 :231–252, 2001.

[19] E. Comets and F. Mentré. Evaluation of tests based on individual versus population modeling to compare dissolution curves. *J. Biopharm. Stat.*, 11 :107–123, 2001.

[20] X. Panhard and F. Mentré. Evaluation by simulation of tests based on non-linear mixed-effects models in interaction and bioequivalence cross-over trials. *Stat. Med.*, 24 :1509–24, 2005.

[21] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85 :398–409, 1990.

[22] A. Racine-Poon. A Bayesian approach to nonlinear random effects models. *Biometrics*, 41 :1015–23, 1985.

[23] J. Wakefield, A. Smith, A. Racine-Poon, and A. Gelfand. Bayesian analysis of linear and non-linear population models by using the gibbs sampler. *J. R. Stat. Soc., Ser. C, Appl. Stat.*, 43 :201–221, 1994.

[24] J. Wakefield. The Bayesian analysis of population pharmacockinetic models. *J. Am. Stat. Assoc.*, 91 :62–75, 1996.

[25] J. E. Bennet, A. Racine-Poon, and J. C. Wakefield. *MCMC for nonlinear hierarchical models*, pages 339–358. Chapman & Hall, London, 1996.

[26] D.J. Spiegelhalter, A. Thomas, N.G. Best, and W.R. Gilks. Bugs : Bayesian inference using gibbs sampling, version 0.5. Technical report, 1996.

[27] C. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.*, 92 :162–170, 1997.

[28] M. G. Gu and H.T. Zhu. Maximum likelihood estimation for spatial models by Markov Chain Monte Carlo stochastic approximation. *J. R. Stat. Soc. B*, 63 :339–355, 2001.

[29] D. Commenges, H. Jacqmin Gadda, C. Proust, and J. Guedj. A Newton like algorithm for likelihood maximization : the robust variance scoring algorithm. *submitted*, 2006.

[30] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49 :1020–1038, 2005.

[31] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, 2002.

[32] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22 :1701–1762, 1994.

[33] C.P. Robert. Simulation of truncated normal variables. *Stat. Comput.*, 5 :121–125, 1995.

[34] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive mcmc. *submitted*, 2006.

[35] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall, London, 1996.

[36] Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. B*, 44 :226–233, 1982.

[37] M. Tanner. *Tools for statistical inference, methods for the exploration of posterior distributions and likelihood functions*. Springer-Verlag, 1996.

[38] H Robbins and S Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22 :400–407, 1951.

[39] Ming Gao Gu and Fan Hui Kong. A stochastic approximation algorithm with Markov Chain Monte-Carlo method for incomplete data estimation problems. *Proc. Natl. Acad. Sci. U. S. A.*, 95 :7270–7274, 1998.

[40] J. Guedj, R. Thiébaut, and Commenges D. Parameter identification of HIV dynamics models with different observational designs. *soumis*, 2006.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39 :1–38, 1977.

[42] C. Wu. On the convergence property of the EM algorithm. *Ann. Statist.*, 11 :95–103, 1983.

[43] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27 :94–128, 1999.

[44] K. Lange. A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5 :1–18, 1995.

[45] G.C. Wei and M. A. Tanner. Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika*, 77 :649–652, 1990.

[46] X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm : a general framework. *Biometrika*, 80 :267–278, 1993.

[47] J.G. Booth and J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. B*, 61 :265–285, 1999.

[48] S. Walker. An EM algorithm for non-linear random effects models. *Biometrics*, 52 :934–944, 1996.

[49] L. Wu. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *J. Am. Stat. Assoc.*, 97 :955–964, 2002.

[50] L. Wu. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *J. Am. Stat. Assoc.*, 99 :700–709, 2004.

[51] R. H. Leary, R. Jelliffe, A. Schumitzky, and R.E. Port. Accurate maximum likelihood estimation for parametric population analysis. In *PAGE 13, Abstr 491*, 2004.

[52] S. Guzy. Monte carlo parametric Expectation Maximization (MC-PEM) method for analyzing population pharmacokinetic/ pharmacodynamic (PK/PD) data. In *PAGE 15, abstr 881*, 2006.

[53] E.N. Jonsson and L.B. Sheiner. More efficient clinical trials through use of scientific model-based statistical tests. *Clin. Pharmacol. Ther.*, 72 :603–14, 2002.

[54] S.L. Beal. Ways to fit a PK model with some data below the quantification limit. *J. Pharmacokinet. Pharmacodyn.*, 28 :481–504, 2001.

[55] J.P. Hughes. Mixed effects models with censored data with applications to HIV RNA levels. *Biometrics*, 55 :625–629, 1999.

[56] H. Wu and L. Wu. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. *Stat. Med.*, 20 :1755–1769, 2001.

[57] C.W. Tornoe, H. Agerso, E.N. Jonsson, H. Madsen, and H.A. Nielsen. Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in nlme using differential equations. *Comput. Methods Programs Biomed.*, 76 :31–40, 2004.

[58] R. Krishna. *Applications of Pharmacokinetic principles in drug development.* Kluwer Academic/Plenum Publishers, New York, 2004.

[59] R.V. Overgaard, N. Jonsson, C.W. Tornoer, and H. Madsen. Non-linear mixed-effects models with stochastic differential equations : Implementation of an estimation algorithm. *J. Pharmacokinet. Pharmacodyn.*, 32 :85–107, 2005.

[60] C.W. Tornoe, R.V. Overgaard, H. Agerso, H.A. Nielsen, H. Madsen, and E.N. Jonsson. Stochastic differential equations in NONMEM : implementation, application, and comparison with ordinary differential equations. *Pharm. Res.*, 22 :1247–58, 2005.

[61] C. Laredo V. Genon-Catalot, T. Jeantheau. Parameter estimation for discretely observed stochastic volatility models. *Bernoulli*, 5 :855–872, 1999.

[62] A. Gloter and J. Jacod. Diffusions with measurement errors. I. local asymptotic normality. *ESAIM : P & S*, 5 :225–242, 2001.

[63] R. Thiebaut, H. Jacqmin-Gadda, A. Babiker, D. Commenges, and CASCADE Collaboration. Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat. Med.*, 24 :65–82, 2005.

[64] R.E. Nettles, T. L. Kieffer, T. Parsons, J. Johnson, J. Cofrancesco, J. E. Gallant, K.A. Carson, R.F. Siliciano, and C. Flexner. Marked intraindividual variability in antiretroviral concentrations may limit the utility of therapeutic drug monitoring. *Clin. Infect. Dis.*, 42 :1189–1196, 2006.

[65] S. Retout, S. Duffull, and F. Mentré. Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs. *Comput. Methods Programs Biomed.*, 65 :141–51, 2001.

[66] O. Launay, L. Gérard, L. Morand-Joubert, P. Flandre, S. Guiramand-Hugon, V. Joly, G. Peytavin, A. Certain, C. Lévy, S. Rivet, C. Jacomet, J.-P. Aboulker, P. Yéni, and Agence Nationale de Recherches sur le SIDA (ANRS) 081 Study Group Yéni. Nevirapine or lamivudine plus stavudine and indinavir : examples of 2-class versus 3-class regimens for the treatment of human immunodeficiency virus type 1. *Clin. Infect. Dis.*, 35 :1096–105, 2002.

[67] W.-Y. Tan and H. Wu. Stochastic modeling of the dynamics of CD4+ T-cell infection by HIV and some Monte Carlo studies. *Math. Biosci.*, 147 :173–205, 1998.

[68] N.I. Stilianakis, C.A. Boucher, M.D. De Jong, R. Van Leeuwen, R. Schuurman, and R.J. De Boer. Clinical data sets of human immunodeficiency virus type 1 reverse transcriptase-resistant mutants explained by a mathematical model. *J. Virol.*, 71 :161–168, 1997.

[69] D. Finzi, M. Hermankova, T. Pierson, L. M. Carruth, C. Buck, R. E. Chaisson, T. C. Quinn, K. Chadwick, J. Margolick, R. Brookmeyer, J. Gallant, M. Markowitz, D. D. Ho, D. D. Richman, and R. F. Siliciano. Identification of a

reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*, 278 :1295–1300, 1997.

[70] M. Di Mascio, M. Markowitz, M. Louie, A. Hurley, C. Hogan, V. Simon, D. Follmann, D.D. Ho, and A.S. Perelson. Dynamics of intermittent viremia during highly active antiretroviral therapy in patients who initiate therapy during chronic versus acute and early human immunodeficiency virus type 1 infection. *J. Virol.*, 78 :10566–10573, 2004.

[71] Y. Huang, S.L. Rosenkranz, and H. Wu. Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. *Math. Biosci.*, 184 :165–186, 2003.

[72] H. Wu, Y. Huang, E.P. Acosta, S.L. Rosenkranz, D.R. Kuritzkes, J.J. Eron, A.S. Perelson, and J.G. Gerber. Modeling long-term hiv dynamics and antiretroviral response : effects of drug potency, pharmacokinetics, adherence, and drug resistance. *J Acquir Immune Defic Syndr*, 39 :272–283, 2005.

[73] J.M.G. Taylor, W.G. Cumberland, and J.P. Sy. A stochastic model for analysis of longitudinal AIDS data. *J. Amer. Stat. Assoc.*, 89 :727–736, 1994.

[74] J.M. Taylor and N. Law. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts ? *Stat. Med.*, 17 :2381–2394, 1998.

[75] A. Hsu, G. R. Granneman, and R. J. Bertz. Ritonavir. Clinical pharmacokinetics and interactions with other anti-HIV agents. *Clin. pharmacokinet.*, 35 :275–291, 1998.

[76] K. Brendel, M. Legrand, A.M. Taburet, G. Baron, C. Goujard, F. Mentré, and Cophar 1-ANRS 102 Trial Group. Population pharmacokinetic analysis of indinavir in HIV-infected patient treated with a stable antiretroviral therapy. *Fundam. Clin. Pharmacol.*, 19 :373–383, 2005.

[77] R. M. Hoetelmans. Pharmacology of antiretroviral drugs. *Antivir. Ther.*, 4 Suppl 3 :29–41, 1999.

[78] M. Lavielle and F. Mentré. Estimation of population pharmacokinetic parameters of saquinavir in HIV patients and covariate analysis with MONOLIX. In *PAGE 14 Abstr 813 [www.page-meeting.org/ ?abstract=813]*, 2005.

[79] E. Comets, C. Verstuyft, and F. Mentré. Building a pharmacogenetic model to describe the pharmacokinetics of digoxin. In *PAGE 14 (2005) Abstr 822 [www.page-meeting.org/ ?abstract=822]*, 2005.

[80] D. Makowski and M. Lavielle. Using SAEM to estimate parameters of models of response to applied fertilizer. *J. Agric. Biol. Environ. Stat.*, to appear, 2006.

[81] B. Amzal, F.Y. Bois, E. Parent, and C.P. Robert. Bayesian optimal design via interacting MCM. *Cahiers du Ceremade 0348.*, 2003.

[82] S.R. Lewin, R.M. Ribeiro, T. Walters, G.K. Lau, S. Bowden, S. Locarnini, and A.S. Perelson. Analysis of hepatitis B viral load decline under potent therapy : complex decay profiles observed. *Hepatology*, 34 :1012–1020, 2001.

[83] R.M. Ribeiro, A. Lo, and A.S. Perelson. Dynamics of hepatitis B virus infection. *Microbes Infect.*, 4 :829–35, 2002.

[84] N.L. Komarova and D. Wodarz. Evolutionary dynamics of mutator phenotypes in cancer : implications for chemotherapy. *Cancer Res.*, 63 :6635–6642, 2003.

Résumé

Cette thèse est consacrée au développement de nouvelles méthodes d'estimation adaptées à l'analyse de données répétées dans l'infection par le virus de l'immunodéficience humaine (VIH). J'ai développé différentes extensions de l'algorithme Stochastic Approximation Expectation Maximisation (SAEM) pour l'estimation par maximum de vraisemblance des paramètres de modèles mixtes complexes, utilisés dans ce contexte de modélisation biologique. J'ai considéré des modèles mixtes dont les observations sont censurées à gauche ainsi que des modèles mixtes ayant un niveau de variabilité supplémentaire, dépendant de la période d'observation. Je me suis également intéressée à des modèles biologiques définis par équations différentielles ordinaires ou stochastiques : la fonction de régression du modèle mixte est alors définie comme solution de ces équations différentielles. Pour ces différents problèmes, j'ai proposé des versions adaptées de SAEM, dont j'ai étudié la convergence théorique. J'ai implémenté ces différents algorithmes et les ai testés sur des données de dynamique du VIH simulées et réelles.

Abstract

This thesis deals with new estimation methods for the analysis of the human immunodeficiency virus (HIV) dynamics by repeated measurements. I developed several extensions of the Stochastic Approximation Expectation Maximisation (SAEM) algorithm for the maximum likelihood estimation of mixed effects model parameters, used in this context. My work was focused on mixed models with left-censored observed data, then with three levels of varibility depending of the observation period. I also considered biological models defined by ordinary or stochastic differential equations : the regression function of the mixed model is then the solution of these differential equations. For these problems, I proposed several versions of the SAEM algorithm and proved their convergence. I implemented these algorithms and evaluated them on HIV dynamic simulated and real datasets.

Mots clés

Modèles mixtes - Modèles à données non observées - Algorithme de Gibbs - Algorithme MCMC - Estimation par maximum de vraisemblance - Algorithme EM - Algorithme d'approximation stochastique - Estimation bayesienne - Données censurées - Equation différentielle ordinaire - Equation différentielle stochastique - Processus de diffusion - Pharmacocinétique - Dynamique du VIH

Classification AMS (2000) : 62F10, 62F15, 62J02, 62J12, 62K99, 62M99, 62P10, 68W01, 92B15