Inférence statistique dans les modèles mixtes à dynamique Markovienne

Maud Delattre

UNIVERSITÉ PARIS SUD | FACULTÉ DES SCIENCES D'ORSAY

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD 11

Spécialité : Mathématiques

par

Maud DELATTRE

Inférence statistique dans les modèles mixtes à dynamique markovienne

---

dirigée par Marc LAVIELLE

Rapporteurs :
M. Jean-Marc AZAIS — Université de Toulouse
M. Eric MOULINES — Télécom ParisTech

Soutenue le 04 juillet 2012 à l'Université Paris-Sud 11 devant le jury composé de

Mme. Elisabeth	GASSIAT	Université Paris-Sud 11	Examinatrice
Mme. Valentine	GENON-CATALOT	Université Paris 5	Examinatrice
M. Marc	LAVIELLE	Université Paris-Sud 11	Directeur de thèse
Mme. France	MENTRÉ	Université Paris Diderot	Examinatrice
M. Eric	MOULINES	Télécom ParisTech	Rapporteur

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

Résumé

La première partie de cette thèse est consacrée à l'estimation par maximum de vraisemblance dans les modèles mixtes à dynamique markovienne. Nous considérons plus précisément des modèles de Markov cachés à effets mixtes et des modèles de diffusion à effets mixtes. Dans le Chapitre 2, nous combinons l'algorithme de Baum-Welch à l'algorithme SAEM pour estimer les paramètres de population dans les modèles de Markov cachés à effets mixtes. Nous proposons également des procédures spécifiques pour estimer les paramètres individuels et les séquences d'états cachés. Nous étudions les propriétés de cette nouvelle méthodologie sur des données simulées et l'appliquons sur des données réelles de nombres de crises d'épilepsie. Dans le Chapitre 3, nous proposons d'abord des modèles de diffusion à effets mixtes pour la pharmacocinétique de population. Nous en estimons les paramètres en combinant l'algorithme SAEM à un filtre de Kalman étendu. Nous étudions ensuite les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans des modèles de diffusion observés sans bruit de mesure continûment sur un intervalle de temps fixé lorsque le nombre de sujets tend vers l'infini. Le Chapitre 4 est consacré à la sélection de covariables dans des modèles mixtes généraux. Nous proposons une version du BIC adaptée au contexte de double asymptotique où le nombre de sujets et le nombre d'observations par sujet tendent vers l'infini. Nous présentons quelques simulations pour illustrer cette procédure.

Mots-clés : maximum de vraisemblance, modèles à effets mixtes, modèles de Markov cachés, équations différentielles stochastiques, algorithme SAEM, sélection de modèles, pharmacologie.

Abstract

The first part of this thesis deals with maximum likelihood estimation in Markovian mixed-effects models. More precisely, we consider mixed-effects hidden Markov models and mixed-effects diffusion models. In Chapter 2, we combine the Baum-Welch algorithm and the SAEM algorithm to estimate the population parameters in mixed-effects hidden Markov models. We also propose some specific procedures to estimate the individual parameters and the sequences of hidden states. We study the properties of the proposed methodologies on simulated datasets and we present an application to real daily seizure count data. In Chapter 3, we first suggest mixed-effects diffusion models for population pharmacokinetics. We estimate the parameters of these models by combining the SAEM algorithm with the extended Kalman filter. Then, we study the asymptotic properties of the maximum likelihood estimate in some mixed-effects diffusion models continuously observed on a fixed time interval when the number of subjects tends to infinity. Chapter 4 is dedicated to variable selection in general mixed-effects models. We propose a BIC adapted to the asymptotic context where both of the number of subjects and the number of observations per subject tend to infinity. We illustrate this procedure with some simulations.

Keywords: maximum likelihood, mixed-effects models, hidden Markov models, stochastic differential equations, SAEM algorithm, model selection, pharmacology.

Un grand merci !

A Marc Lavielle pour avoir accepté d'encadrer mon stage de Master 2 puis mes travaux de thèse. Merci de m'avoir proposé un sujet de recherche aussi riche, alliant développements théoriques et applications, et de m'avoir toujours aidée à mettre en valeur mon travail.

A Eric Moulines et Jean-Marc Azais pour avoir accepté d'examiner mes travaux de thèse en qualité de rapporteurs.

A Elisabeth Gassiat pour avoir accepté de présider mon jury de thèse.

A France Mentré pour avoir accepté de faire partie de mon jury de thèse. Merci également de m'avoir aiguillée vers Marc pour mon stage de Master 2, et de m'avoir si souvent et si généreusement accueillie dans les locaux de ton laboratoire pour me permettre de mener à bien les travaux de ma première année de thèse.

A Valentine Genon-Catalot. Merci d'avoir accepté de participer à mon jury de thèse. Merci surtout de m'avoir proposé cette collaboration fructueuse avec Adeline. Ta confiance, tes conseils, et ta disponibilité m'ont été très précieux.

Aux personnes avec qui j'ai eu l'occasion de collaborer ou plus simplement de discuter ces trois dernières années et qui m'ont permis d'enrichir les travaux de ma thèse. Avec une mention particulière à Adeline pour sa motivation, sa disponibilité et son dynamisme exemplaires et à Marie-Anne pour son oreille attentive !

Aux membres du laboratoire de mathématiques d'Orsay pour avoir su créer un environnement propice au bon déroulement de ma thèse.

A Valérie pour m'avoir aidée, avec beaucoup de gentillesse, de patience et d'efficacité, à trouver la sortie du labyrinthe administratif qui accompagne la fin de la thèse.

Au groupe des ingénieurs de Monolix pour les nombreux dépannages informatiques et leur bonne humeur communicative !

Aux thésards passés et présents de l'école doctorale d'Orsay, pour tous les bons moments que nous avons pu partager durant ces trois années.

A mes camarades Ensaiens et maubeugeois pour leur amitié sincère et leur soutien.

A ma famille pour son soutien inconditionnel, et en particulier à mes parents, mon frère et ma sœur, à qui je tiens à dédier ces travaux.

Table des matières

Chapitre 1

Introduction

Contents

1.1 Préambule

Les travaux de cette thèse portent essentiellement sur l'étude de modèles mixtes à dynamique markovienne. La plupart des développements proposés sont motivés par des problèmes concrets en biologie et s'inscrivent à l'interface entre la statistique et la biostatistique. Quatre grandes catégories de problèmes statistiques parsèment nos travaux.

1. Le premier problème est celui de la **modélisation**. La question est abordée à plusieurs reprises, pour la description de nombres de crises d'épilepsie d'une part (Chapitre 2), et pour la description de données en pharmacocinétique d'autre part (Chapitre 3). L'enjeu est de construire un modèle statistique dont la structure reflète au mieux le mécanisme ayant engendré les données tout en permettant une description fidèle de la variabilité dans les données. Comme nous le verrons de façon plus détaillée dans la suite de l'introduction, dans des contextes de mesures répétées sur plusieurs sujets, les sources de variabilité dans les données sont multiples. La variabilité entre les sujets est prise en compte très classiquement en définissant certains paramètres du modèle comme des variables aléatoires. En revanche, la description de la variabilité intra-sujet ne répond pas à une règle générale et se traite au cas par cas. Dans nos travaux sur la description des nombres de crise d'épilepsie et des données de pharmacocinétique, des processus markoviens permettent de rendre compte des corrélations entre les observations d'un même sujet.
2. Le deuxième problème est purement méthodologique. Une fois le modèle statistique posé, il s'agit d'ajuster le modèle aux données par des méthodes appropriées. L'estimateur du maximum de vraisemblance est un estimateur naturel pour les paramètres des modèles à effets mixtes. La structure complexe des modèles mixtes à dynamique markovienne rend toutefois l'estimation délicate, la vraisemblance étant, comme dans la plupart des modèles à effets mixtes, rarement explicite. Nous reviendrons sur la forme de la vraisemblance dans la suite. Nous proposons pour chacun des modèles considérés une **méthodologie** fondée sur l'algorithme d'estimation par maximum de vraisemblance SAEM, dont nous détaillerons le fonctionnement général dans la suite de l'introduction.
3. La troisième question concerne la **théorie des estimateurs**. Une fois les problèmes numériques liés au calcul des estimateurs du maximum de vraisemblance résolus, nous nous attachons à ses propriétés asymptotiques. En d'autres termes, nous nous intéressons au lien entre la taille de l'échantillon et la confiance que l'on peut accorder à l'estimateur. Nous verrons que comme pour le point précédent, la principale difficulté provient de l'expression de la vraisemblance des observations par des intégrales. Cette question est abordée dans le cadre des modèles de diffusion à effets mixtes dans le Chapitre 3.
4. Enfin, les manières de modéliser un même phénomène étant nombreuses, notre dernier problème est de guider le choix du modèle dont la complexité est la mieux adaptée à l'échantillon. Le quatrième axe de recherche est celui de la **sélection de modèles** dans une approche de population. Dans le Chapitre 4, nous nous intéresserons exclusivement à la sélection de modèles mixtes par des procédures de type BIC.

Dans cette thèse, ces quatre questions sont étroitement mêlées. Après une présentation des modèles à effets mixtes et des principales méthodes d'estimation dans ces modèles, dans un souci de clarté, nous organiserons nos travaux en trois parties : *i)* méthodologie et application des modèles de Markov cachés à effets mixtes (Chapitre 2), *ii)* estimation par maximum de vraisemblance dans les modèles de diffusion à effets mixtes (Chapitre 3), *iii)* sélection de

covariables dans les modèles à effets mixtes par un critère BIC (Chapitre 4). Ces chapitres étant très disjoints, nous choisissons de présenter une bibliographie spécifique à la fin de chaque chapitre plutôt qu'une bibliographie générale en fin de manuscrit.

1.2 Forme générale d'un modèle à effets mixtes

Les modèles à effets mixtes sont essentiellement utilisés pour la modélisation de données répétées obtenues sur plusieurs sujets. Ces modèles trouvent aujourd'hui de nombreuses applications en biologie, notamment pour analyser l'évolution de maladies chroniques, ou encore étudier les caractéristiques d'un médicament en pharmacocinétique.

Pour une présentation générale des modèles à effets mixtes, notons N le nombre de sujets et $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,n_i})$ le vecteur des n_i observations obtenues pour le sujet i, $i = 1, \ldots, N$. $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ désigne l'échantillon global. Les sujets sont supposés indépendants, mais les observations pour un même sujet ne le sont pas nécessairement.

Lorsque l'on dispose de mesures répétées sur plusieurs sujets, la variabilité observée dans les données peut se décliner sous plusieurs formes (Figure 1.1) : *i)* d'abord une variabilité entre les individus qualifiée de variabilité inter-sujets, *ii)* ensuite une variabilité entre les différentes observations de chaque sujet, dite variabilité intra-sujet.

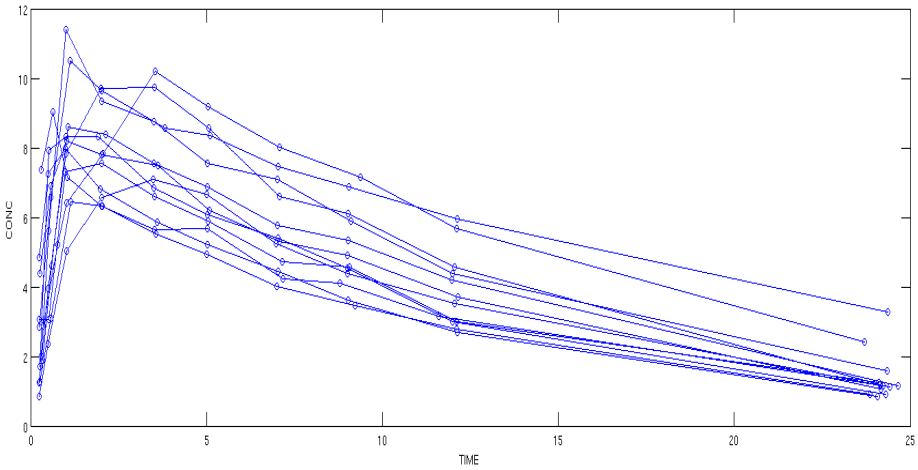

FIGURE 1.1 – Représentation graphique des données du jeu "Theophylline" disponible dans MONOLIX. Ce graphique illustre la variabilité inter-sujets dans des expériences à mesures répétées sur plusieurs sujets.

L'analyse de données répétées sur plusieurs sujets nécessite de pouvoir distinguer ces différentes sources de variabilité, et requiert de ce fait la mise en œuvre d'outils statistiques adaptés. Les modèles à effets mixtes ont été développés pour ce contexte précis. Ils se définissent de façon hiérarchique. Au premier niveau, chaque vecteur d'observations individuelles est décrit par un même modèle paramétrique, et ses propres paramètres :

$$\mathbf{y}_i \sim h(\cdot, \phi_i),$$

où ϕ_i désigne le vecteur des paramètres individuels, et h est une distribution donnée. Le modèle structurel est choisi pour décrire au mieux la variabilité propre à chaque sujet. Les écarts du modèle structurel aux données (ou variabilité résiduelle) sont éventuellement pris en compte en associant à ce modèle structurel un modèle d'erreur résiduelle. Notons que la structure du modèle h n'implique pas nécessairement que les données d'un même sujet soient indépendantes. Nous verrons par exemple dans le Chapitre 2 qu'une structure de chaîne de Markov cachée permet de rendre compte des variations au cours du temps des manifestations symptomatiques chez un sujet épileptique, ou encore dans le Chapitre 3 que les processus de diffusion sont des outils pertinents pour décrire les perturbations observées dans la cinétique d'un médicament. Notons que dans les modèles à dynamique markovienne, les données d'un même sujet sont corrélées.

Le deuxième niveau de spécification des modèles à effets mixtes définit les paramètres individuels comme des variables aléatoires pour décrire la variabilité inter-sujets. Les différences entre les sujets peuvent en partie s'expliquer par des caractéristiques connues des individus (*ex :* l'âge, le poids, ...), intégrées au modèle par des covariables. La part de variabilité non expliquée par ces covariables est prise en compte sous forme d'effets aléatoires. Les paramètres individuels sont alors définis comme des variables aléatoires indépendantes, dont la loi dépend éventuellement de covariables :

$$\phi_i \sim \pi(\cdot, C_i, \theta).$$

Les paramètres $\theta \in \Theta$ de cette loi sont appelés paramètres de population, où Θ est un sous-ensemble de $\mathbb{R}^p$. Dans l'ensemble de nos travaux, nous considérerons un modèle gaussien pour les ϕ_i, de la forme

$$\begin{aligned} \phi_i &= \beta C_i + \eta_i, \\ \eta_i &\underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega), \end{aligned} \tag{1.1}$$

donnant

$$\phi_i \sim \mathcal{N}(\beta C_i, \Omega),$$

où C_i est un ensemble de covariables pour le sujet i. Ici, $\theta = (\beta, \Omega)$. Pour simplifier dans la suite, nous noterons $\pi(\cdot, \theta)$ la loi des paramètres individuels.

***Exemple* 1.1.** Pour illustrer et comprendre la structure générale d'un modèle à effets mixtes, nous présentons un exemple. Celui-ci est très utilisé en pharmacocinétique de population pour décrire l'évolution au cours du temps de la concentration d'un médicament dans le plasma après administration orale d'une dose D_i de la substance médicamenteuse, à partir d'observations y_{ij} recueillies sur N patients en des temps discrets t_{ij}. Le jeu "Theophylline" de MONOLIX, représenté en Figure 1.1 a été récolté lors d'une telle expérience.

1. *Modèle paramétrique pour un individu*

   Dans cet exemple, l'un des modèles usuellement construits pour décrire les données du sujet $i = 1, \ldots, N$ s'écrit :

$$\begin{aligned} y_{ij} &= \frac{D_i}{V_i} \frac{k_{a,i}}{k_{a,i} - k_{e,i}} \left(e^{-k_{e,i} t_{ij}} - e^{-k_{a,i} t_{ij}} \right) + \xi_{ij}, \\ \xi_{ij} &\sim \mathcal{N}(0, \sigma_i^2). \end{aligned}$$

$V_i, k_{a,i}, k_{e,i}$ désignent respectivement le volume, la constante d'absorption du médicament et sa constante d'élimination. Les ξ_{ij} représentent des erreurs de mesure. Les paramètres individuels sont donnés par $V_i, k_{a,i}, k_{e,i}, \sigma_i$.

2. *Modèle pour les paramètres individuels*

   Afin de décrire la variabilité entre les patients, il paraît réaliste de supposer que le volume et les constantes physiologiques diffèrent d'un individu à l'autre et par conséquent de les définir comme des variables aléatoires. De manière raisonnable, on peut également supposer que la variance des erreurs de mesure est la même pour l'ensemble des sujets. Dans ce cas, un modèle possible pour les paramètres individuels est donné par :

$$\begin{pmatrix} \log V_i \\ \log k_{e,i} \\ \log k_{a,i} \end{pmatrix} \underset{i.i.d.}{\sim} \mathcal{N}\left(\begin{pmatrix} \log V \\ \log k_e \\ \log k_a \end{pmatrix}, \Omega \right),$$

   avec $\sigma_i = \sigma$ pour tout $i = 1, \ldots, N$. Les paramètres de population θ du modèle global sont alors V, k_a, k_e, σ et Ω.

Par leur structure hiérarchique, les modèles à effets mixtes autorisent une analyse des données à plusieurs niveaux.

a) L'évaluation de la distribution des paramètres individuels dans la population constitue le point essentiel dans l'interprétation des données. En effet, l'estimation du paramètre θ permet d'établir l'évolution typique du phénomène modélisé au sein de la population et d'évaluer les variations autour de celui-ci (Figure 1.2). La principale difficulté réside dans l'expression complexe de la vraisemblance des modèles à effets mixtes. Diverses méthodes sont utilisées dans la pratique pour estimer les paramètres de population dans les modèles à effets mixtes : des méthodes basées sur une approximation du modèle (linéarisation de la vraisemblance [16], approximation de Laplace ou quadrature de Gauss [31]), des méthodes bayésiennes [19, 28, 27], ou encore par maximisation exacte de la vraisemblance du modèle. Parmi ces dernières, l'algorithme EM (*Expectation - Maximisation*) et ses variantes, dont les algorithmes MCEM (*Monte-Carlo EM*) et SAEM (*Stochastic Approximation EM*), sont détaillées dans la suite.

b) Les modèles à effets mixtes permettent aussi d'analyser les mécanismes propres à chaque sujet (Figure 1.3). Une fois les paramètres de population estimés, il est possible de prédire les profils individuels en étudiant la distribution a posteriori des ϕ_i : $p(\phi_i|\mathbf{y}_i, \theta)$.

1.3 Inférence dans les modèles à effets mixtes

Une fois le modèle déterminé, la question fondamentale est d'évaluer la variabilité inter-sujets ainsi que le profil typique. En d'autres termes, il s'agit d'estimer la valeur du paramètre θ qui décrive le mieux les données. L'estimateur le plus naturel pour les paramètres de population est l'estimateur du maximum de vraisemblance (MLE, *Maximum Likelihood Estimate*). Cependant, le calcul du MLE dans les modèles à effets mixtes est généralement délicat, la vraisemblance étant rarement explicite du fait du caractère aléatoire des paramètres individuels.

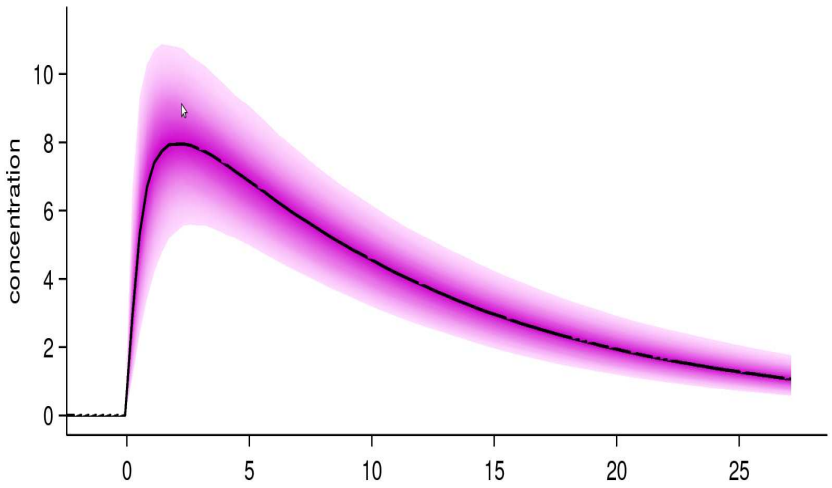

FIGURE 1.2 – Variabilité inter-sujets dans l'évolution de la concentration médicamenteuse au cours du temps, évaluée à partir des estimateurs du maximum de vraisemblance des paramètres de population du modèle oral donné dans l'Exemple 1.1 calculés sur le jeu "Theophylline" dans MONOLIX.

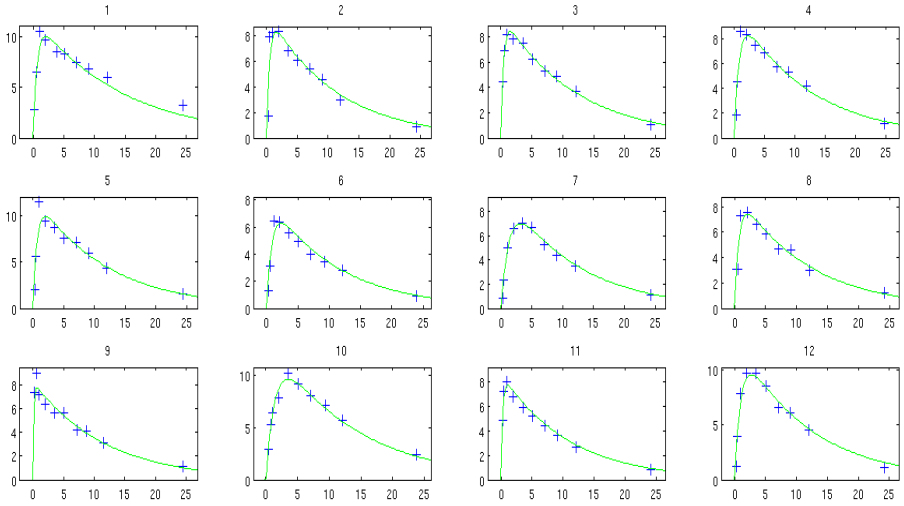

FIGURE 1.3 – Prédiction des profils individuels pour le jeu "Theophylline", obtenue à partir de l'estimation a posteriori des paramètres individuels du modèle oral donné dans l'Exemple 1.1 (MONOLIX).

1.3.1 Expression générale de la vraisemblance

Les modèles à effets mixtes peuvent être considérés comme des modèles à données incomplètes pour lequels le vecteur $\phi = (\phi_1, \ldots, \phi_N)$ des paramètres individuels constitue l'ensemble des données non observées. Dans les modèles à données incomplètes, la distribution marginale des observations s'exprime en intégrant la vraisemblance conditionnelle des

observations par rapport à la distribution des données non observées. La forme générale de la vraisemblance des observations dans un modèle à effets mixtes est la suivante :

$$\begin{aligned} p(\mathbf{y};\theta) &= \prod_{i=1}^{N} p(\mathbf{y}_i;\theta), \\ &= \prod_{i=1}^{N} \int h(\mathbf{y}_i,\phi_i)\pi(\phi_i;\theta)d\phi_i. \end{aligned} \tag{1.2}$$

À l'exception des modèles linéaires à effets mixtes ou de modèles non linéaires à effets mixtes très particuliers, tels que ceux considérés dans la deuxième partie du Chapitre 3 par exemple, les intégrales $\int h(\mathbf{y}_i,\phi_i)\pi(\phi_i;\theta)d\phi_i$ ne sont pas calculables. La vraisemblance des observations n'a alors pas d'expression analytique, et le calcul du maximum de vraisemblance n'est pas direct. Si la vraisemblance des observations est difficilement calculable dans la plupart des modèles à effets mixtes, la vraisemblance complète $p(\mathbf{y},\phi;\theta)$ a généralement une forme explicite. Certaines méthodes d'estimation par maximum de vraisemblance dans les modèles à données incomplètes sont basées sur la vraisemblance des données complètes, et sont donc particulièrement adaptées à l'estimation des paramètres dans les modèles à effets mixtes. Dans la suite, nous décrivons parmi ces algorithmes ceux dont l'utilisation est fréquente dans le cadre des modèles à effets mixtes.

Remarque 1. Dans les modèles à structure markovienne latente étudiés dans cette thèse, nous verrons que les réalisations non observées du système dynamique sous-jacent peuvent être considérées comme des paramètres de nuisance plutôt que des données non observées du modèle.

1.3.2 Estimation par maximum de vraisemblance

1.3.2.1 L'algorithme EM

L'algorithme EM, introduit par Dempster et al. [6] est un algorithme itératif de calcul des estimateurs du maximum de vraisemblance dans les modèles à données non observées. Si toutes les données du modèle $(\mathbf{y},\phi)$ étaient observées, on choisirait d'estimer θ par maximisation de la log-vraisemblance complète $\log p(\mathbf{y},\phi;\theta)$. En l'absence de connaissance sur ϕ, nous ne pouvons plus maximiser la log-vraisemblance complète par rapport à θ. En revanche, il est possible d'optimiser son espérance conditionnelle sachant les données observées $\mathbf{y}$ par rapport au paramètre θ. Cette astuce est le fondement de l'algorithme EM. Pour tout $(\theta,\theta') \in \Theta^2$, nous noterons

$$Q(\theta|\theta') = \mathbb{E}(\log p(\mathbf{y},\phi;\theta)\,|\mathbf{y},\theta').$$

Cette espérance conditionnelle possède la propriété suivante. Tout accroissement de Q augmente la vraisemblance des données observées $p(\mathbf{y},\theta)$: pour tout $(\theta,\theta') \in \Theta^2$,

$$Q(\theta|\theta') \geq Q(\theta|\theta) \Rightarrow \log p(\mathbf{y},\theta') \geq \log p(\mathbf{y},\theta).$$

Cette propriété est appelée propriété de monotonie. Elle justifie le fonctionnement de l'algorithme EM par maximisations successives de Q pour obtenir un estimateur du paramètre θ. Notons $(\theta_k)_{k>0}$ la suite des estimateurs obtenus au cours de l'algorithme. L'itération k de l'algorithme EM se décompose alors en deux phases :

– l'étape *E (Expectation)* qui s'attache à estimer l'espérance conditionnelle $Q(\theta|\theta_{k-1})$ de la log-vraisemblance complète sachant les données observées $\mathbf{y}$ et la valeur courante du paramètre θ_{k-1},
– l'étape M *(Maximisation)* où l'on actualise l'estimateur en maximisant la fonction obtenue en phase E :

$$\theta_k = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta|\theta_{k-1}).$$

L'estimateur retenu pour θ est la valeur calculée à la dernière itération de l'algorithme. La convergence de l'algorithme EM vers un point stationnaire de la log-vraisemblance des observations a été démontrée par plusieurs auteurs [6, 32].

La mise en œuvre de l'algorithme EM pose cependant un certain nombre de problèmes pratiques : lenteur de la convergence ou difficultés liées au calcul de $Q(\theta|\theta')$. Le second problème tient au fait que $Q(\theta|\theta')$ se calcule au moyen d'intégrales qui n'ont pas toujours de forme analytique simple, rendant les étapes E et M particulièrement complexes voire incalculables. Ce problème est résolu par le développement de versions stochastiques de l'algorithme EM, l'idée étant d'approcher la log-vraisemblance conditionnelle $Q(\theta|\theta')$ par Monte-Carlo plutôt que de la calculer de façon exacte.

1.3.2.2 L'algorithme MCEM

L'algorithme MCEM, proposé par Wei and Tanner [30], est l'une des variantes stochastiques de l'algorithme EM. Il s'agit dans l'algorithme MCEM de donner une approximation par Monte Carlo de l'étape E de l'algorithme EM en tirant un échantillon des données non observées dans la distribution conditionnelle $p(\phi|\mathbf{y};\theta_{k-1})$. Notons m_k la taille de l'échantillon $\phi_k = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{k,m_k})$ obtenu à l'itération k de l'algorithme. $Q(\theta|\theta_{k-1})$ est approché par la moyenne empirique :

$$Q(\theta|\theta_{k-1}) \backsimeq \frac{1}{m_k} \sum_{l=1}^{m_k} \log p(\mathbf{y}, \phi_{kl}; \theta).$$

Lorsque la simulation exacte sous la loi $p(\phi|\mathbf{y};\theta_{k-1})$ n'est pas possible, comme c'est souvent le cas dans les modèles à effets mixtes, il est proposé de combiner l'algorithme MCEM à une procédure de Monte-Carlo par chaînes de Markov (MCMC) pour simuler les données non observées [29, 33, 34]. Les méthodes de Monte-Carlo par chaîne de Markov, introduites par Hastings [11], permettent de simuler des échantillons d'une distribution de probabilité donnée par l'intermédiaire d'une chaîne de Markov dont les échantillons sont asymptotiquement distribués selon la distribution requise. Pour des détails pratiques et théoriques sur ces méthodes, le lecteur pourra se référer à l'ouvrage de Robert [20].

Des problèmes numériques sont rapportés dans la littérature, en particulier une convergence très lente de l'algorithme MCEM. Certains auteurs précisent en effet qu'un très grand nombre de simulations lors des dernières itérations est nécessaire pour assurer la convergence de l'algorithme ; c'est par exemple le cas de Booth and Hobert [2]. Ces simulations intensives sont problématiques car elles occasionnent généralement des temps de calculs très importants. Ceci est confirmé par la théorie. En particulier, Fort and Moulines [10] montrent que le taux de convergence de l'algorithme dépend de la taille des échantillons simulés.

1.3.2.3 L'algorithme SAEM

L'algorithme SAEM est une autre variante stochastique de l'algorithme EM, introduite par Delyon et al. [5]. Comme pour l'algorithme MCEM, il s'agit d'obtenir une approximation de la log-vraisemblance à l'étape E grâce à la simulation des données non observées sous leur distribution conditionnelle. L'approximation empirique de la log-vraisemblance $Q(\theta|\theta_{k-1})$ proposée par MCEM est remplacée par une procédure d'approximation stochastique, qui ne nécessite la simulation que d'une réalisation des données non observées à chaque itération.

L'itération $k > 0$ de l'algorithme SAEM est donc la succession de trois étapes :

1. une phase S *(simulation)*, où l'on simule une réalisation des données non observées : $\phi_k \sim p(\cdot|\mathbf{y}; \theta_{k-1})$;
2. une phase SA *(approximation stochastique)*, où les ϕ_k sont combinés aux observations $\mathbf{y}$ pour approcher $Q(\theta|\theta_{k-1})$:
$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\log p(\mathbf{y}, \phi_k; \theta) - Q_{k-1}(\theta)\right),$$
3. une phase M *(maximisation)*, où l'estimateur du paramètre est actualisé en maximisant $Q_k(\theta)$ par rapport à θ :
$$\theta_k = \underset{\theta \in \Theta}{\operatorname{argmax}}\ Q_k(\theta),$$

$(\gamma_k)_{k>0}$ désignant une suite de pas décroissante.

L'algorithme SAEM présente de nombreux intérêts pratiques. En particulier, l'estimation par maximum de vraisemblance dans les modèles à effets mixtes par SAEM requiert des temps de calcul bien plus faibles que la plupart des autres méthodes présentées précédemment. De plus l'algorithme converge vers une valeur proche du maximum de vraisemblance en un faible nombre d'itérations, y compris dans des modèles complexes et incluant un grand nombre d'effets aléatoires. Enfin, le choix de l'initialisation a un impact plus faible sur la convergence de l'algorithme que pour les autres méthodes. Ces propriétés pratiques sont en lien avec la procédure d'approximation stochastique de la log-vraisemblance des données complètes mise en œuvre par l'algorithme : à chaque itération, les données simulées lors des itérations précédentes sont utilisées pour affiner l'approximation de la log-vraisemblance des données complètes. Ceci évite le recours à des simulations intensives des données non observées à chaque itération, coûteuses en temps de calcul.

La simulation exacte de la loi conditionnelle des données manquantes à l'étape S de l'algorithme SAEM n'est pas toujours réalisable. Dans ce cas, l'étape S est combinée à une procédure de Monte-Carlo [13]. À l'itération k, une réalisation ϕ_k des données non observées est donc obtenue par une chaîne de Markov $\tilde{\phi}$ de loi stationnaire $p(\cdot|\mathbf{y}, \theta_{k-1})$. Notons $q(\cdot|\cdot)$ la loi instrumentale choisie. L'étape S consiste alors en M itérations d'un algorithme de Metropolis-Hastings :

1. Initialisation : $\tilde{\phi}_0 = \phi_{k-1}$;
2. Pour $m = 1, \ldots, M$, un candidat $\tilde{\phi}_m$ est tiré sous la distribution $q(\phi_{k-1}|\cdot)$, et retenu avec probabilité
$$\min\left(\frac{p(\tilde{\phi}_m|\mathbf{y}, \theta_{k-1})}{p(\phi_{k-1}|\mathbf{y}, \theta_{k-1})} \frac{q(\tilde{\phi}_m|\phi_{k-1})}{q(\phi_{k-1}|\tilde{\phi}_m)}\right),$$

3. $\phi_k = \tilde{\phi}_M$.

Remarque 2. Le calcul des probabilités d'acceptation nécessite de connaître l'expression de $p(\phi|\mathbf{y};\theta)$ pour tout $(\mathbf{y},\phi)$ et tout $\theta \in \Theta$. Notons que $p(\phi|\mathbf{y};\theta) \propto p(\mathbf{y}|\phi;\theta)p(\phi|\theta)$. La connaissance de $p(\mathbf{y}|\phi;\theta)$ et $p(\phi;\theta)$ pour tout $(\mathbf{y},\phi)$ et tout $\theta \in \Theta$ est donc requise. Dans les modèles à effets mixtes considérés dans cette thèse, la distribution des données non observées est connue de façon explicite puisque nous choisissons les paramètres individuels gaussiens (équation (1.1)). En revanche, dans certains modèles, les distributions conditionnelles des observations $h(\cdot,\phi_i)$ n'ont pas une expression simple. La clé pour adapter l'algorithme MCMC-SAEM à ces modèles résidera dans une procédure efficace de calcul des densités conditionnelles $h(\cdot,\phi_i)$, comme nous le soulignerons dans les Chapitres 2 et 3.

La convergence des suites $(\theta_k)_{k>0}$ générées par les algorithmes SAEM et MCMC-SAEM vers un maximum local de la vraisemblance observée est avérée d'un point de vue théorique. Delyon et al. [5] ont démontré la convergence presque sûre de $(\theta_k)_{k>0}$ vers un maximum local de la vraisemblance observée dans les situations où il est possible de simuler les données non observées de façon exacte à chaque itération, sous des hypothèses de régularité du modèle et de compacité du support de la suite des approximations stochastiques proposées par l'algorithme. Leur résultat est étendu à l'algorithme MCMC-SAEM par Kuhn and Lavielle [13], et la condition de compacité est levée par les travaux de Allassonniere et al. [1].

1.3.3 Estimation de la matrice d'information de Fisher

Estimer la matrice d'information de Fisher est primordial pour évaluer la loi asymptotique des estimateurs de θ. Dans les modèles à effets mixtes, il est possible d'évaluer la matrice d'information de Fisher par linéarisation du modèle au premier ordre autour de la moyenne des effets aléatoires. Cette méthode est implémentée dans le logiciel PFIM, mais ne s'applique pas aux modèles à observations discrètes. L'algorithme SAEM permet d'estimer la matrice d'information de Fisher dans n'importe quel modèle à effets mixtes par une procédure détaillée dans Delyon et al. [5]. Celle-ci est basée sur la formule de Louis, selon laquelle

$$\frac{\partial^2}{\partial\theta\partial\theta'}\log p(\mathbf{y};\theta) = \mathbb{E}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log p(\mathbf{y},\phi;\theta)|\mathbf{y},\theta\right) - \mathrm{cov}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)|\mathbf{y},\theta\right),$$

où

$$\begin{aligned}\mathrm{cov}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)|\mathbf{y},\theta\right) &= \mathbb{E}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)'|\mathbf{y},\theta\right)\\ &\quad - \mathbb{E}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)\right)\mathbb{E}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{y},\phi;\theta)\right)'.\end{aligned}$$

La matrice d'information de Fisher est alors approchée par la séquence (H_k) définie comme suit :

$$\begin{aligned}\Delta_k &= \Delta_{k-1} + \gamma_k\left[\frac{\partial\log p(\mathbf{y},\phi_k;\theta_k)}{\partial\theta} - \Delta_{k-1}\right],\\ D_k &= D_{k-1} + \gamma_k\left[\frac{\partial^2\log p(\mathbf{y},\phi_k;\theta_k)}{\partial\theta\partial\theta'} - D_{k-1}\right],\\ G_k &= G_{k-1} + \gamma_k\left[\frac{\partial\log p(\mathbf{y},\phi_k;\theta_k)}{\partial\theta}\frac{\partial\log p(\mathbf{y},\phi_k;\theta_k)}{\partial\theta}' - G_{k-1}\right],\\ H_k &= D_k + G_k - \Delta_k\Delta_k'.\end{aligned}$$

Sous des hypothèses de régularité du modèle, la séquence $(H_k)_{k>0}$ converge presque sûrement vers la matrice d'information de Fisher observée [5].

1.3.4 Estimation de la vraisemblance

La vraisemblance d'un modèle à effets mixtes, dont l'expression générale est rappelée en équation (3.6), est rarement explicite du fait des intégrales par rapport aux effets aléatoires. Le calcul de la vraisemblance des observations est pourtant important pour la mise en œuvre de tests du rapport de vraisemblance ou encore pour le calcul de critères de sélection de modèles tels que le BIC. Il est possible d'évaluer la vraisemblance des observations par des méthodes reposant sur la linéarisation du modèle. Par cette technique, le modèle est approché par un modèle gaussien, mais ce type d'approximation n'est pas toujours valable. Des méthodes de calcul par quadrature permettent également d'approcher l'intégrale exprimant la vraisemblance. Ces méthodes posent néanmoins des problèmes de convergence lorsque le modèle comprend un grand nombre d'effets aléatoires. Des méthodes alternatives, par Monte-Carlo, ont été suggérées. Kuhn and Lavielle [14] ont d'abord proposé de calculer la vraisemblance des observations par la moyenne empirique :

$$\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{y}, \phi^{(t)}),$$

où les $\phi^{(t)}$ forment un T-échantillon généré selon la distribution $\pi(\cdot;\theta)$. Samson et al. [22] ont proposé d'évaluer la vraisemblance par des méthodes d'échantillonnage préférentiel, en simulant des échantillons sous une distribution instrumentale $\tilde{\pi}(\cdot,\theta)$ et en approchant la vraisemblance des observations par la moyenne empirique

$$\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{y}, \phi^{(t)})\frac{\pi(\phi^{(t)};\theta)}{\tilde{\pi}(\phi^{(t)};\theta)}.$$

1.4 Motivations de la thèse

L'algorithme SAEM est un outil puissant et performant pour l'estimation des paramètres dans une approche populationnelle. Il est aujourd'hui implémenté dans des logiciels dédiés à l'analyse de modèles à effets mixtes en pharmacologie. L'algorithme SAEM a d'abord été implémenté dans le logiciel MONOLIX, développé par le groupe de travail Inria du même nom et depuis 2011 par la société Lixoft. Le logiciel de référence dans l'industrie pharmaceutique, NONMEM, privilégie les méthodes fondées sur la linéarisation de la vraisemblance (FO, FOCE) pour l'analyse de modèles à effets mixtes, mais permet également d'utiliser l'algorithme SAEM dans sa dernière version NONMEM7, disponible depuis 2011. Il est également possible d'utiliser l'algorithme SAEM sous Matlab via la fonction `nlmefitsa` de la "Statistics Toolbox", et depuis peu sous R en utilisant le package `saemix`. Grâce à ces développements logiciels, les modèles à effets mixtes sont très largement diffusés aussi bien dans le milieu académique que dans le milieu industriel, et l'algorithme SAEM est très facilement utilisable en pratique. Le logiciel MONOLIX a été le support de nombreuses études de population, notamment en pharmacocinétique (PK) et pharmacondynamique (PD) [15, 25, 3], en pharmacogénétique [4], en génétique [12] et en agronomie [17].

Plusieurs adaptations de l'algorithme SAEM ont été proposées en réponse aux besoins liés aux différentes applications des modèles à effets mixtes. Samson et al. [21] ont proposé une version spécifique de l'algorithme SAEM pour l'analyse de données censurées à gauche dans une approche populationnelle, en lien avec la description de données de dynamique virale pour le VIH. La question de l'évaluation d'effets traitement étant récurrente en pharmacologie, Samson et al. [22] ont également développé des tests de Wald et du rapport de vraisemblance pour évaluer les effets de covariables dans les modèles non linéaires à effets mixtes en estimant la matrice d'information de Fisher par l'algorithme SAEM et en estimant la vraisemblance des observations par échantillonnage préférentiel. Panhard and Samson [18] ont étendu l'algorithme SAEM aux modèles incluant un niveau supplémentaire d'effets aléatoires pour décrire la variabilité inter-occasions dans l'analyse pharmacocinétique d'un traitement antirétroviral. Beaucoup de modèles dynamiques en pharmacologie sont définis par des systèmes d'équations différentielles ordinaires, dont la solution n'est pas nécessairement connue de façon explicite. Donnet and Samson [7] ont proposé une version de l'algorithme SAEM couplée à une procédure de linéarisation locale des équations différentielles ordinaires pour estimer les paramètres de ces modèles dans une approche populationnelle. Ces algorithmes sont mis à profit par Tao et al. [26] dans le cadre de modèles à effets mixtes reposant sur des systèmes d'équations différentielles ordinaires de grande dimension en génomique. Des modèles de diffusion à effets mixtes sont parfois utilisés comme une alternative aux modèles dynamiques définis par des équations différentielles ordinaires en modélisation compartimentale. Nous y reviendrons dans le Chapitre 3. Deux versions de l'algorithme SAEM ont été proposées pour estimer les paramètres de ces modèles, l'une faisant appel à l'approximation du processus de diffusion par la méthode d'Euler-Maruyama [8], la seconde associant des méthodes de filtrage particulaire à l'algorithme SAEM [9]. D'autres adaptations spécifiques de SAEM ont également été proposées, pour des modèles mixtes à observations discrètes [23] et des modèles mixtes à observations catégorielles [24].

Ainsi, l'industrie pharmaceutique est demandeuse de méthodes performantes pour l'analyse de modèles de plus en plus complexes dans une approche de population. Dans cette thèse, nous généralisons l'algorithme SAEM aux modèles à effets mixtes à dynamique markovienne : les modèles de Markov cachés à effets mixtes puis les modèles de diffusion à effets mixtes. Ces modèles ont des applications pratiques en pharmacologie. Nous verrons dans le Chapitre 2 que les modèles de Markov cachés à effets mixtes sont de bons candidats pour la description de nombres de crises d'épilepsie et l'évaluation de traitements anti-épileptiques. Les modèles de diffusion à effets mixtes sont quant-à eux justifiés en pharmacocinétique, comme nous le verrons dans le Chapitre 3. Ces algorithmes sont destinés à être implémentés dans le logiciel MONOLIX. Nous aborderons ensuite la question de la sélection de modèles dans une approche de population. Plus précisément, dans le Chapitre 4, nous proposerons de clarifier l'usage du BIC en sélection de covariables dans des modèles non linéaires mixtes généraux, la pénalité de ce critère étant équivoque dans le cadre des modèles à effets mixtes.

1.5 Bibliographie

[1] S. Allassonniere, E. Kuhn, and A. Trouvé. Construction of bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16(3) :641–678, 2010.

[2] J.G. Booth and J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, 61 :265–285, 1999.

[3] P.L.S. Chan, P. Jacqmin, M. Lavielle, L. McFadyen, and B. Weatherley. The use of the saem algorithm in monolix software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic hiv subjects. *Journal of Pharmacokinetics and Pharmacodynamics*, 38 :41–61, 2011.

[4] E. Comets, C. Verstuyft, M. Lavielle, P. Jaillon, L. Becquemont, and F. Mentré. Modelling the influence of mdr1 polymorphism on digoxin pharmacokinetic parameters. *European Journal of Clinical Pharmacology*, 63 :437–449, 2007.

[5] B. Delyon, M. Lavielle, and E. Moulines. Convergence of stochastic approximation version of the em algorithm. *Annals of Statistics*, 27 :94–128, 1999.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27 :94–128, 1977.

[7] S. Donnet and A. Samson. Estimation of parameters in incomplete data models defined by dynamical systems. *Journal of Statistical Planning and Inference*, 137 :2815–2831, 2007.

[8] S. Donnet and A. Samson. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM P& S*, 12 :196–218, 2008.

[9] S. Donnet and A. Samson. Em algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. 2010. submitted paper.

[10] G. Fort and E. Moulines. Convergence of the monte-carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4) :1220–1259, 2003.

[11] W.H. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57 :97–109, 1970.

[12] F. Jaffrézic, C. Meza, M. Lavielle, and J.L. Foulley. Genetic analysis of growth curves using the saem algorithm. *Genetics Selection Evolution*, 38 :583–600, 2006.

[13] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

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

[15] M. Lavielle and F. Mentré. Estimation of pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. *Journal of Pharmacokinetics and Pharmacodynamics*, 34 :229–249, 2007.

[16] M.J. Lindstrom and D.M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46 :673–687, 1990.

[17] D. Makowski and M. Lavielle. Using saem to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11 (1) :45–60, 2006.

[18] X. Panhard and A. Samson. Extension of the saem algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10 :121–135, 2009.

[19] A. Racine-Poon. A bayesian approach to nonlinear random effects models. *Biometrics*, 41 :1015–1023, 1985.

[20] C. Robert. *Méthodes de Monte-Carlo par chaînes de Markov*. Statistique Mathématique et probabilité, paris, economica edition, 1996.

[21] A. Samson, M. Lavielle, and F. Mentré. Estension of the saem algorithm to left-censored data in nonlinear mixed-effects model : application to hiv dynamics model. *Computational Statistics and Data Analysis*, 53 :1562–1574, 2006.

[22] A. Samson, M. Lavielle, and F. Mentré. The saem algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Statistics in Medicine*, 26 :4860–4875, 2007.

[23] R.M. Savic and M. Lavielle. Performance in population models for count data, part ii : a new saem algorithm. *Journal of Pharmacokinetics and Pharmacodynamics*, 36 :367–379, 2009.

[24] R.M. Savic, F. Mentré, and M. Lavielle. Implementation and evaluation of the saem algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics and pharmacodynamics. *The AAPS Journal*, 13 :44–53, 2010.

[25] E. Snoeck, P. Chanu, M. Lavielle, P. Jacqmin, E.N. Jonsson, K. Jorga, T. Goggin, J. Grippo, N.L. Jumbe, and N. Frey. A comprehensive hepatitis c viral kinetic model explaining cure. *Clinical Pharmacology and Therapeutics*, 87 :706–713, 2010.

[26] L. Tao, L. Hua, L. Hongzhe, and W. Hulin. High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106(496) :1242–1258, 2011.

[27] J. Wakefield. The bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association*, 91 :62–75, 1996.

[28] J.C. Wakefield, A. F. M. Smith, A. Racine-Poon, and A. E. Gelfand. Bayesian analysis of linear and non-linear population models by using the gibbs sampler. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43 :201–221, 1994.

[29] S. Walker. An em algorithm for non-linear random effects models. *Biometrics*, 52 : 934–944, 1996.

[30] G.C.G. Wei and M.A. Tanner. A monte carlo implementation of the em algorithm and the poor's man's data augmentation algorithms. *Journal of American Statistical Association*, 85 :699–704, 1990.

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

[32] C.F.J. Wu. On the convergence properties of the em algorithm. *Annals of Statistics*, 11 : 95–103, 1983.

[33] L. Wu. 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*, 97 :955–964, 2002.

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

Chapitre 2

Modèles de Markov cachés à effets mixtes

Contents

2.1 Introduction

Les résultats établis dans ce chapitre sont inspirés des besoins liés à la description de nombres de crises d'épilepsie sur plusieurs patients. L'épilepsie est une maladie neurologique qui se manifeste par des crises associées à une hyperactivité cérébrale. Ces crises se produisent généralement de façon soudaine et imprévisible, et présentent des caractéritiques très différentes selon les patients et l'instant des crises : hallucinations, perte de connaissance, convulsions, ... Actuellement, les éléments déclencheurs des crises ne sont pas établis avec certitude. Chez la plupart des sujets épileptiques, des traitements médicamenteux sont prescrits en prévention, pour ralentir la fréquence des crises ou diminuer leur intensité. Ce chapitre s'articule autour de la description de données d'essai clinique pour l'un d'entre eux, la gabapentine. La conduite de l'essai clinique est détaillée dans la troisième partie de ce chapitre. Dans le cadre de ce travail, l'échantillon est composé des nombres des crises d'épilepsie obtenus chaque jour de l'essai clinique sur plusieurs centaines de patients. Quelques exemples d'observations individuelles sont présentés en Figure 2.1. Cette représentation graphique illustre la dissociation de la variabilité des données en plusieurs composantes :

- tant au niveau du nombre quotidien moyen de crises d'épilepsie que de l'évolution de la maladie au cours du temps, il existe une grande variabilité entre les patients (*variabilité inter-sujets*),
- pour un sujet donné, la quantité des crises peut être très fluctuante dans le temps (*variabilité intra-sujet*).

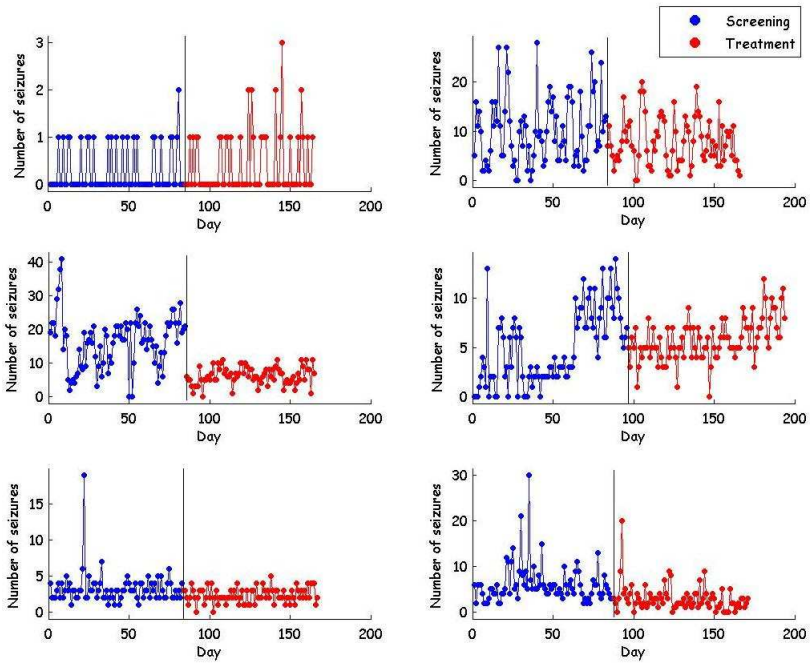

FIGURE 2.1 – Évolution du nombre de crises d'épilepsie pour six sujets au cours des deux phases de l'essai clinique sur la gabapentine.

Pour un patient épileptique donné, décrire de telles fluctuations nécessite l'utilisation de modèles statistiques adéquats. Il a été proposé, notamment dans l'article de Albert [1], de

décrire l'évolution du nombre de crises d'épilepsie au cours du temps au moyen de modèles de Markov cachés (*HMM, Hidden Markov Model*). Les modèles de Markov cachés sont en effet utiles pour modéliser des phénomènes décomposables en phases ou états lorsque la séquence des états visités n'est pas directement observable.

Rappelons rapidement la définition d'un modèle de Markov caché. Un HMM de paramètres φ est formé de l'association de deux processus stochastiques (Figure 2.2). Le premier, que l'on note $z = (z_j)_{j\in\mathbb{N}^*}$, est une chaîne de Markov à temps discrets et à espace d'états discrets $\{1, 2, \ldots, S\}$, non observée et de matrice de transition $\Pi(\varphi) = (p(z_{j+1} = s'|z_j = s; \varphi))_{1\leq s,s'\leq S}$. Le second, noté $y = (y_j)_{j\in\mathbb{N}^*}$, est observable et généré par des distributions d'émission qui diffèrent selon les états :

$$y_j|z_j = s \sim p_s(\cdot; \varphi) \ , \ s \in \{1, 2, \ldots, S\}.$$

Pour une présentation plus détaillée des modèles de Markov cachés et des méthodologies spécifiques à ces modèles, le lecteur pourra se référer aux travaux de Baum ([8, 7, 9]), au tutoriel de Rabiner [23], ou encore à l'ouvrage de Cappé et al. [10].

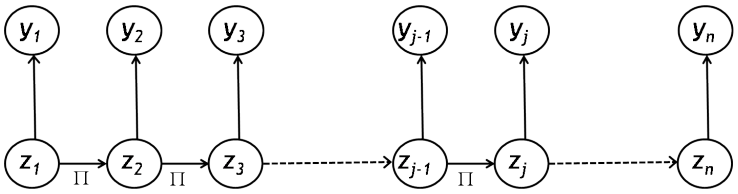

FIGURE 2.2 – Modèle de Markov caché.

Outre leur souplesse vis à vis de l'irrégularité des manifestation épileptiques, les HMM proposent une interprétation simple et réaliste de la maladie. Ces modèles supposent que les observations - les nombres quotidiens de crises d'épilepsie - sont des variables aléatoires distribuées selon les réalisations d'un processus de Markov discret, où l'espace d'états symbolise différents stades dans la maladie. Les crises sont rythmées par les transitions du patient d'un état à un autre. Par exemple, dans nos travaux, nous supposerons l'existence de deux états, un état à faible activité épileptique et un état à forte activité épileptique (Figure 2.3). D'autre part, les états ne sont pas indépendants, de sorte que la nature de l'état à un instant donné dépend de l'état à l'instant qui le précède directement. Cette hypothèse, dite propriété de Markov, est naturelle pour modéliser l'évolution dans le temps d'une maladie chronique. D'autres applications des modèles de Markov cachés à des pathologies différentes de l'épilepsie ont aussi été proposées [2, 4, 6].

Nous étendons les modèles de Markov cachés dans une approche populationnelle (*MHMM, Mixed Hidden Markov Model*) pour décrire les nombres de crises d'épilepsie des patients soignés par gabapentine. En d'autres termes, le modèle pour le sujet i, $i = 1, \ldots, N$, est un HMM de paramètres ϕ_i, et les ϕ_i sont des variables aléatoires de même distribution paramétrée par θ. Notons $\mathbf{z}_i = (z_{i1}, \ldots, z_{i,n_i})$ la séquence d'états cachés pour le sujet i. Le but sera d'estimer θ par maximum de vraisemblance à partir des seules observations $\mathbf{y}$. L'expression de la vraisemblance est, au premier abord, particulièrement complexe. Comme dans tout modèle à effets mixtes, on obtient $p(\mathbf{y}; \theta)$ en intégrant les vraisemblances conditionnelles par rapport à

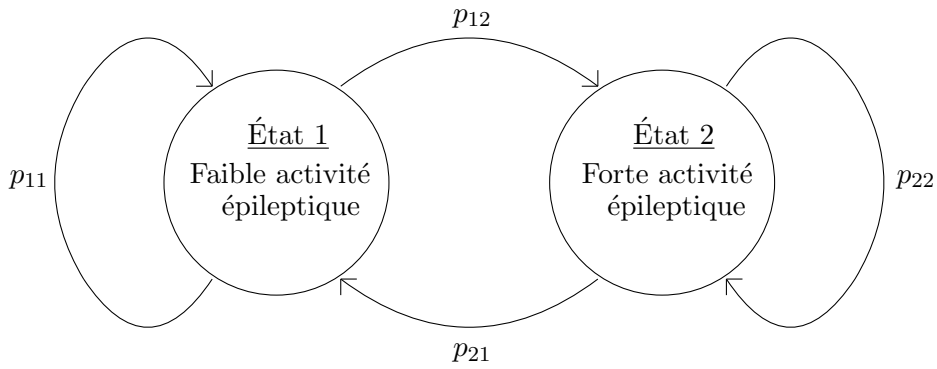

FIGURE 2.3 – Dynamique des crises d'épilepsie. Exemple d'une chaîne de Markov à deux états représentant des phases de faible et de forte activité épileptique.

la distribution des effets aléatoires. De plus, $\mathbf{z}_i$ étant inconnue, $p(\mathbf{y}_i|\phi_i)$ s'exprime comme une somme sur l'ensemble des séquences d'états possibles :

$$p(\mathbf{y}_i|\phi_i) = \sum_{z_{i1}} \dots \sum_{z_{i,n_i}} \left[p(z_{i1}|\phi_i) \left(\prod_{j=2}^{n_i} p(z_{ij}|z_{i,j-1},\phi_i) \right) \left(\prod_{j=1}^{n_i} p(y_{ij}|z_{ij},\phi_i) \right) \right]. \tag{2.1}$$

Pour autant, la somme multiple en (2.1) reste facilement calculable puisque la procédure forward en permet un calcul récursif rapide et exact [23]. En revanche, les intégrales par rapport aux ϕ_i n'ont pas d'expression analytique. Plusieurs procédures ont été proposées pour estimer les paramètres de population dans les HMM à effets mixtes. Certaines sont fondées sur l'algorithme EM [3, 24, 21]. Le calcul de l'étape E est le point clé de cet algorithme. Altman [3] propose de le résoudre par des méthodes d'intégration numérique, et Rijmen et al. [24] par un algorithme de l'arbre de jonction. Maruotti and Rydén [21] adoptent quant-à eux une approche semi-paramétrique. Toutefois, Altman [3] souligne le coût important en temps de calcul de son algorithme : dès lors que le modèle contient plus de deux effets aléatoires, le temps nécessaire pour obtenir la convergence de l'algorithme est de l'ordre de plusieurs heures, voire plusieurs jours, y compris pour des jeux de données de taille raisonnable (39 sujets et 24 observations par sujet). L'auteur propose également une version spécifique de l'algorithme MCEM. Néanmoins, pour en assurer la convergence, il est nécessaire de calibrer le nombre de simulations des effets aléatoires à chaque itération en fonction du nombre d'effets aléatoires dans le modèle, ce qui est également très coûteux en temps de calcul. Ainsi, trois jours sont nécessaires pour ajuster un MHMM à trois effets aléatoires au même jeu de données, des échantillons de taille 50000 étant simulés lors des dernières itérations.

Nous avons proposé une méthode d'estimation des paramètres dans les modèles de Markov cachés à effets mixtes fondée sur l'algorithme SAEM. L'algorithme de Baum-Welch est combiné à SAEM et permet de traiter efficacement les problèmes liés à l'expression de la vraisemblance complète lors de la mise en œuvre de l'algorithme de Metropolis-Hastings à l'étape de simulation. Nous proposons également une procédure d'estimation a posteriori des paramètres individuels, et utilisons l'algorithme de Viterbi pour estimer les séquences d'états cachés les plus probables au vu des données de chaque sujet. Les propriétés de la méthode d'estimation

par maximum de vraisemblance SAEM sont illustrées par une étude sur données simulées sous un modèle de Markov caché à effets mixtes à deux états, qui nous a ensuite servi de point de départ à l'analyse des données réelles. Ce travail a fait l'objet d'une publication dans un numéro spécial du *Journal de la Société Française de Statistiques* paru suite à la tenue du *GDR Statistique et Santé 2009* [12], et d'un article publié dans *Computational Statistics and Data Analysis* [13]. Pour limiter les redondances, seul le deuxième article sera présenté dans ce manuscrit.

Nous avons ensuite utilisé cette méthodologie pour analyser les données de nombres de crises d'épilepsie. Les propriétés du traitement antiépileptique par gabapentine ont été étudiées lors d'un essai contre placebo incluant 788 patients épileptiques, chez qui l'on a dénombré les crises quotidiennes pendant une durée totale de 24 semaines. L'utilisation de MHMM a permis de montrer une différence significative entre le placebo et la nouvelle substance médicamenteuse. La mise en pratique de l'algorithme SAEM sur données réelles a également permis de mettre en lumière le faible coût en temps de calcul des méthodes proposées, les rendant ainsi très attractives dans la pratique, notamment dans des situations où l'analyse porte sur des jeux de données de grande taille et fait intervenir des modèles complexes. Ce travail appliqué, réalisé en collaboration avec Radojka M. Savic, post-doctorante à l'UMR INSERM 738 de Université Paris Diderot, Mats O. Karlsson, professeur à l'université d'Uppsala (Suède), Raymond Miller, responsable Recherche et Développement chez Pfizer, et Marc Lavielle, est développé dans un article accepté dans *Journal of Pharmacokinetics and Pharmacodynamics* [14].

2.2 Premier article : Maximum Likelihood Estimation in Discrete Mixed Hidden Markov Models using the SAEM algorithm

Ce travail, co-écrit avec Marc Lavielle, a été publié dans la revue Computational Statistics and Data Analysis [13].

Abstract

Mixed hidden Markov models have been recently defined in the literature as an extension of hidden Markov models for dealing with population studies. The notion of mixed hidden Markov models is particularly relevant for modeling longitudinal data collected during clinical trials, especially when distinct disease stages can be considered. However, parameter estimation in such models is complex, especially due to their highly nonlinear structure and the presence of unobserved states. Moreover, existing inference algorithms are extremely time consuming when the model includes several random effects. New inference procedures are proposed for estimating population parameters, individual parameters and sequences of hidden states in mixed hidden Markov models. The main contribution consists of a specific version of the stochastic approximation EM algorithm coupled with the Baum-Welch algorithm for estimating population parameters. The properties of this algorithm are investigated via a Monte-Carlo simulation study, and an application of mixed hidden Markov models to the description of daily seizure counts in epileptic patients is presented.

Keywords: Nonlinear Mixed Effects Model, SAEM algorithm, Forward, Backward Algorithm, Epileptic Seizures Count.

2.2.1 Introduction

Markov chains are a useful tool for analyzing time-series data. However, the Markov process sometimes cannot be directly observed but we can assume that some output, dependent on the state, is visible. More precisely, we assume that the distribution of this continuous or discrete observable data depends on the underlying hidden state. Such models are called hidden Markov models (HMMs); for a review of their properties, see [8, 7, 9, 23, 10]. HMMs can be applied in many contexts, and particularly appear to be pertinent in several biological contexts. For example, they are useful when characterizing diseases for which the existence of several discrete illness stages is a realistic assumption. Applications to epilepsy and migraines can be found in [1] and [6] respectively.

Recently, [3] extended hidden Markov models to population studies. However, inference in mixed hidden Markov models is a complex issue; see for example [3, 24, 21, 11]. Indeed, due to the highly nonlinear structure of these models, computing the maximum likelihood estimate (MLE) of the parameters is challenging. The aim of the present paper is to propose a new mixed HMM inference methodology. In particular, we suggest adapting the MCMC-SAEM algorithm [19] to get the MLE of the population parameters in mixed HMMs.

The article is organized as follows. We start with a general definition of mixed HMMs. The second section is devoted to the estimation method, of which the most original part is the use of the MCMC-SAEM algorithm for estimation of population parameters. The last section presents an application of mixed HMMs on epileptic seizure count data after a Monte Carlo simulation study aiming at investigating the properties of the estimates given by the SAEM algorithm. General conclusions follow.

2.2.2 Mixed Hidden Markov Models

Mixed HMMs were recently defined by [3] as an extension of HMMs to population studies. In the present section, some brief details on HMMs will help to introduce mixed HMMs. Here, we consider a parametric framework with homogeneous Markov chains on a discrete and finite state space $\mathbf{S} = \{1, \ldots, S\}$ and discrete observations in $\mathbb{N}$.

2.2.2.1 Hidden Markov models

HMMs have been developed to understand how a given system moves from one state to another over time in situations where the successive visited states are unknown and a set of observations is the only available information to describe the dynamics of the system. HMMs can be seen as a variant of mixture models, allowing for possible memory in the sequence of hidden states. An HMM is thus defined as a pair of processes $\{z_j, y_j\}$, where the observation process $\{y_j\}_{j=1}^n$ enables inference on the latent one $\{z_j\}_{j=1}^n$. More precisely, $\{z_j\}_{j=1}^n$ is a homogeneous Markov chain with initial state distribution $\pi(\cdot)$ and transition probabilities

$$a_{s,s\prime} = \mathbb{P}\left(z_{j+1} = s\prime | z_j = s\right),$$

for all $(s, s\prime) \in \mathbf{S}^2$ with

$$\sum_{s\prime \in \mathbf{S}} a_{s,s\prime} = 1,$$

for all $s \in \mathbf{S}$. The observation process is related to the latent one by the response distributions given by the probabilities

$$b_{\ell,s} = \mathbb{P}\left(y_j = \ell | z_j = s\right),$$

$(\ell, s) \in \mathbb{N} \times \mathbf{S}$, with

$$\sum_{\ell \in \mathbb{N}} b_{\ell,s} = 1,$$

for all $s \in \mathbf{S}$. Additionally, the observations are assumed to be conditionally independent, given the states. Denote ψ the set of parameters of the model (initial state distribution, transition probabilities, response probabilities), $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$. Then, the HMM likelihood is given by:

$$L(\mathbf{y}; \psi) = \sum_{\mathbf{z} \in \mathbf{S}^n} \pi(z_1) \prod_{j=1}^{n-1} a_{z_j, z_{j+1}} \prod_{j=1}^{n} b_{y_j, z_j}. \tag{2.2}$$

Note that an extension of this model to continuous observations is straightforward. See [23] and [10] for further details about HMMs.

2.2.2.2 Mixed HMMs

HMMs are a powerful modeling tool to describe a single sequence of observations when the existence of a discrete latent process is assumed. However, when longitudinal data from several individuals need to be described simultaneously, the model needs to account for different variability sources in the data. In this case, mixed HMMs can extend HMMs to deal with this specific context of the population approach. Before defining mixed HMMs, let us fix some notation. Let N be the number of subjects, n_i the number of observations for individual i ($i = 1, \ldots, N$), and y_{ij} and z_{ij} denote respectively the j^{th} outcome and the j^{th} hidden state for individual i, $j = 1, \ldots, n_i$. The sequences of observations and hidden states for each individual are respectively denoted by $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,n_i})$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{i,n_i})$.

Mixed HMMs involve several levels of construction. An HMM is first specified for each individual's data. The N individual HMMs are supposed independent with the same structure, but their parameters are expected to vary from one individual to another. Let ψ_i denote the set of parameters of the i^{th} HMM, $i = 1, \ldots, N$. ψ_i is basically composed of the transition probabilities

$$a_{s,s\prime}^{(i)} = \mathbb{P}\left(z_{i,j+1} = s\prime | z_{ij} = s\right) \ \forall (s, s\prime) \in \mathbf{S}^2,$$

and the response probabilities

$$b_{\ell,s}^{(i)} = \mathbb{P}\left(y_{ij} = \ell | z_{ij} = s\right) \ \forall (\ell, s) \in \mathbb{N} \times \mathbf{S},$$

of subject i.

In a second step, the individual parameters are assumed to be random variables with a shared probability distribution, possibly including covariates. We will consider a Gaussian model, i.e., assume that there exists some known transformation g such that:

$$\phi_i = g\left(\psi_i\right), \tag{2.3}$$

$$\phi_i = \mu + C_i\beta + \eta_i, \tag{2.4}$$

and

$$\eta_i \underset{i.i.d.}{\backsim} \mathcal{N}\left(0, \Omega\right).$$

For example, natural choices for g would be the logit transformation for the transition probabilities, and log-transformation for the response probabilities. C_i is a known matrix of covariates for individual i, μ and β are unknown vectors of fixed effects and Ω is the variance-covariance matrix of the random effects. The parameters of the ϕ_i distribution in the population are called the population parameters and are indicated here by $\theta = (\mu, \beta, \Omega)$.

The likelihood for mixed HMMs has a nontrivial form. Using independence of the N individuals and the fact that the ϕ_i are unobserved, it is given by

$$\begin{aligned} L\left(\mathbf{y}_1, \ldots, \mathbf{y}_N; \theta\right) &= \prod_{i=1}^{N} \int p(\mathbf{y}_i, \phi_i; \theta) d\phi_i, \\ &= \prod_{i=1}^{N} \int p(\mathbf{y}_i | \phi_i) p(\phi_i; \theta) d\phi_i. \end{aligned} \tag{2.5}$$

Here $p(\phi_i; \theta)$ is a Gaussian density, and $p(\mathbf{y}_i|\phi_i)$ is the marginal distribution of the observations of an HMM with parameters $\psi_i = g^{-1}(\phi_i)$. As the sequences of states are hidden and the

observations for each subject are assumed to be independent conditional on the states, the expression of $p(\mathbf{y}_i|\phi_i)$ follows (2.2).

2.2.3 Estimation in Mixed HMM

In this section, a new estimation methodology for mixed HMMs is described. Our main contribution consists of a specific version of the stochastic approximation EM (SAEM) algorithm coupled with the Baum-Welch algorithm for estimating population parameters of the model. Specific procedures for estimating the individual parameters and the sequences of hidden states are also suggested.

2.2.3.1 Estimation of the population parameters

a) Background

The maximum likelihood approach is the most intuitive approach for estimating the population parameters of mixed HMMs. However, expressing the mixed HMM likelihood involves integration over the ϕ_i's and multiple sums over the states, making its direct maximization with respect to θ intractable. Thus, MLE in mixed HMMs is a challenging issue. In fact, mixed HMMs can be seen as missing data models, with the hidden states and the individual parameters as unobserved data. Therefore, the EM algorithm [16] is a natural parameter estimation method for estimating the population parameters θ from the observations. However, the E-step can not be performed in a closed form in mixed HMMs. This difficulty has been tackled by several authors. [3] suggests evaluating the E-step of the EM algorithm by using numerical integration methods, but the proposed algorithms become time-intensive when the model includes more than two random effects. In [24], the model is described as a directed acyclic graph and the E-step is carried out with a junction tree algorithm. In [3] and [11], the MCEM algorithm is implemented as an alternative to the EM algorithm to perform parameter estimation in HMMs with random effects. [21] also proposed a non-parametric maximum likelihood approach.

b) Estimation of the population parameters in mixed HMMs via the SAEM algorithm

i) General description of the SAEM algorithm

Linear and nonlinear mixed models form a specific sub-class of incomplete data models in which the random effects $\phi = (\phi_1, \ldots, \phi_N)$ are the non-observed data and the population parameters are the parameters of the model that need to be estimated from the observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,n_i})$. In many situations – particularly in nonlinear mixed models – the E-step of the EM algorithm has no closed form. Some variants of the algorithm get around this difficulty. In particular, in [15], a stochastic version of EM, the SAEM algorithm, is proposed in which the E-step is evaluated by a stochastic approximation procedure. Let $\theta^{(k-1)}$ denote the current estimate for the population parameters; iteration k of the SAEM algorithm involves three steps:

- In the simulation step, $\theta^{(k-1)}$ is used to simulate the missing data $\phi_i^{(k)}$ under the conditional distribution $p(\phi_i|\mathbf{y}_i, \theta^{(k-1)})$, $i = 1, \ldots, N$.
- In the stochastic approximation step, the simulated data $\phi^{(k)}$ and the observations $\mathbf{y}$ are used together to update the stochastic approximation $Q_k(\theta)$ of the conditional

expectation $E\left(\log p(\mathbf{y},\phi;\theta)|\mathbf{y},\theta^{(k-1)}\right)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log p(\mathbf{y},\phi^{(k)};\theta) - Q_{k-1}(\theta)\right], \tag{2.6}$$

where $(\gamma_k)_{k>0}$ is a sequence of positive step sizes decreasing to 0 and starting with $\gamma_1 = 1$.

– In the maximization step, an updated value of the estimate $\theta^{(k)}$ is obtained by maximization of $Q_k(\theta)$ with respect to θ:

$$\theta^{(k)} = \underset{\theta}{\operatorname{argmax}}\, Q_k(\theta).$$

This procedure is iterated until numerical convergence of the sequence $\left(\theta^{(k)}\right)_{k>0}$ to some estimate $\hat{\theta}$ is achieved. Convergence results can be found in [15].

Let us now make some practical remarks concerning the implementation of the SAEM algorithm in the general context of nonlinear mixed models, before focusing on the specific adaptation of the algorithm to mixed HMMs.

1. Using the fact that
$$p(\mathbf{y}_i,\phi_i^{(k)};\theta) = p(\mathbf{y}_i|\phi_i^{(k)},\theta) \times p(\phi_i^{(k)};\theta), \tag{2.7}$$
we note that when the conditional distribution of $\mathbf{y}_i$ does not depend on θ, the approximation step reduces to approximating the conditional expectation of the marginal log-likelihood of (ϕ_i):
$$T_k(\theta) = (1-\gamma_k)T_{k-1}(\theta) + \gamma_k \sum_{i=1}^{N} \log p(\phi_i^{(k)};\theta), \tag{2.8}$$
and the maximization step consists in maximizing $T_k(\theta)$. This maximization step is straightforward since only Gaussian distributions are involved.
2. Direct sampling from $p(\phi_i|\mathbf{y}_i,\theta^{(k-1)})$ is not always feasible in the simulation step of SAEM. [19] suggested combining the SAEM algorithm with an MCMC procedure. Then, the simulated values $\phi_i^{(k)}$ are obtained through a Markov chain for which it is supposed that $p(\phi_i|\mathbf{y}_i,\theta^{(k-1)})$ is the unique stationary distribution.
3. When we make inference from small samples, *ie* less than 50 subjects, it is possible in the MONOLIX software to simulate several realizations of ϕ_i instead of a single one in the simulation step of SAEM to improve convergence of the algorithm [27].
4. This algorithm was shown to be very efficient for maximum likelihood estimation in nonlinear mixed models [20]. SAEM was first implemented in MONOLIX, a software mainly dedicated to pharmacokinetics-pharmacodynamics (PKPD) applications [27]. The algorithm has also been recently implemented in NONMEM, MATLAB (`nlmefitsa.m`) and R (`saemix` package).

ii) Simulation of the missing data in mixed HMM

We propose to adapt the SAEM algorithm to estimate population parameters of mixed HMMs. Although the unobserved data in these models consist of both the sequences of

hidden states $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ and the (transformed) individual parameters $\phi = (\phi_1, \ldots, \phi_N)$, we would rather avoid simulating the hidden states $(\mathbf{z}_i)$ at each iteration of the algorithm. Indeed, the conditional likelihoods do not depend on θ in our models. Therefore, according to (2.8), simulated sequences $(\mathbf{z}_i^{(k)})$ are not required for the algorithm since the M-step of SAEM maximizes $T_k(\theta)$, which is only a function of the simulated individual parameters $(\phi_i^{(k)})$.

The difficulty here lies in the simulation of the conditional distribution of the individual parameters (ϕ_i). The simulation step of the MCMC-SAEM algorithm consists of M iterations of the Metropolis-Hastings algorithm: let $\phi_{i,0} = \phi_i^{(k-1)}$, then for $1 \leq m \leq M$ and $1 \leq i \leq n$, $\widetilde{\phi}_{i,m}$ is drawn with some proposal distribution $q_{\theta_{k-1}}(\phi_{i,m-1}, \cdot)$ and accepted with probability:

$$\begin{aligned} \alpha(\phi_{i,m-1}, \widetilde{\phi}_{i,m}) &= \frac{p(\widetilde{\phi}_{i,m}|\mathbf{y}_i;\theta_k) q_{\theta_{k-1}}(\widetilde{\phi}_{i,m}, \phi_{i,m-1})}{p(\phi_{i,m-1}|\mathbf{y}_i;\theta_k) q_{\theta_{k-1}}(\phi_{i,m-1}, \widetilde{\phi}_{i,m})}, \\ &= \frac{p(\mathbf{y}_i, \widetilde{\phi}_{i,m};\theta_k) q_{\theta_{k-1}}(\widetilde{\phi}_{i,m}, \phi_{i,m-1})}{p(\mathbf{y}_i, \phi_{i,m-1};\theta_k) q_{\theta_{k-1}}(\phi_{i,m-1}, \widetilde{\phi}_{i,m})}. \end{aligned} \tag{2.9}$$

After M iterations, we set $\phi_i^{(k)} = \phi_{i,M}$.

Then, for any individual i, the Metropolis-Hastings algorithm requires a quick computation of $p(\mathbf{y}_i, \phi_i; \theta)$ for any ϕ_i and any θ. Note that

$$p(\mathbf{y}_i, \phi_i; \theta) = p(\mathbf{y}_i|\phi_i) \times p(\phi_i; \theta). \tag{2.10}$$

Computing $p(\phi_i;\theta)$ is straightforward since ϕ_i has been defined as a Gaussian variable. On the other hand, $p(\mathbf{y}_i|\phi_i)$ represents the likelihood of the n_i observations $\mathbf{y}_i$ in an HMM with parameter $\psi_i = g^{-1}(\phi_i)$. The forward recursions of the Baum-Welch algorithm provide a quick way to numerically compute $p(\mathbf{y}_i|\phi_i)$ [23].

2.2.3.2 Estimation of the variance of the estimates

When an estimate $\hat{\theta}$ of θ has been obtained with the SAEM algorithm, computing the Fisher information matrix $I(\hat{\theta}) = -\frac{\partial^2 \log(p(\mathbf{y};\theta))}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}}$ allows us to derive the standard errors of the components of $\hat{\theta}$. We propose to estimate $I(\hat{\theta})$ using a stochastic approximation procedure. This methodology is based on the Louis formula, and was initially proposed in [20] for computing the Fisher information matrix in nonlinear mixed effects models for continuous responses. This methodology is implemented in MONOLIX [27], in the Matlab function `nlmefitsa` and in the `saemix` R package. An extension for count data models was proposed in [25] and implemented in MONOLIX. The stochastic approximation procedure implemented for computing $I(\hat{\theta})$ in mixed HMMs requires simulation of the conditional distribution $p(\phi_i|\mathbf{y}_i;\hat{\theta})$ via the Metropolis-Hastings algorithm described above. This algorithm needs $p(\mathbf{y}_i|\phi_i)$ to be calculated in closed form for any subject i and any individual parameter ϕ_i. This is straightforward thanks to the Baum-Welch algorithm.

2.2.3.3 Estimation of the likelihood and model selection

Standard model selection criteria such as the Bayesian Information Criteria (BIC) require computation of the observed log-likelihood $\log(p(\mathbf{y};\hat{\theta}))$. The mixed HMM likelihood defined in (2.5) can not be computed in a closed form. Therefore, $\log(p(\mathbf{y};\hat{\theta}))$ is approximated using

an Importance Sampling integration procedure as initially suggested in [20]. Once again, this procedure requires $p(\mathbf{y}_i|\phi_i)$ in closed form for all ϕ_i, $i = 1, \ldots, N$, thus necessitating the Baum-Welch algorithm. The present procedure for computing the model likelihood and deriving the BIC is also implemented in MONOLIX.

2.2.3.4 Estimation of the individual parameters

When an estimate $\widehat{\theta}$ of the population parameters has been obtained, the prior distribution of $(\psi_1, \ldots \psi_N)$ is fully defined. Several methods are implemented in MONOLIX to estimate ψ_i, $i = 1, \ldots, N$.

1. The first consists in estimating the individual parameters using the Maximum A Posteriori (MAP) approach. The estimate $\hat{\psi}_i$ of ψ_i, $i = 1, \ldots, N$ is then given by:

$$\begin{aligned} \hat{\psi}_i &= \underset{\psi_i}{\operatorname{argmax}}\, p(\psi_i|\mathbf{y}_i, \hat{\theta}), \qquad (2.11) \\ &= \underset{\psi_i}{\operatorname{argmax}}\, p(\mathbf{y}_i|\psi_i)p(\psi_i, \hat{\theta}). \qquad (2.12) \end{aligned}$$

   Maximization of the right-hand term in (2.12) is generally not directly feasible, especially in mixed HMMs, and requires a numerical optimization procedure.

2. The second consists in estimating ψ_i with the mean of the conditional distribution of ψ_i given the observations $\mathbf{y}_i$ and $\hat{\theta}$:

$$\hat{\psi}_i = \mathbb{E}(\psi_i|\mathbf{y}_i, \hat{\theta}). \qquad (2.13)$$

   Once again, the expression of $\mathbb{E}(\psi_i|\mathbf{y}_i, \hat{\theta})$ is not explicit in a mixed HMM, and the conditional mean of ψ_i is estimated with a Metropolis-Hastings algorithm.

2.2.3.5 Estimation of the individual parameters

Using the individual parameter estimate $\hat{\psi}_i$ obtained either as the conditional mode or as the conditional mean of $p(\psi_i|\mathbf{y}_i, \hat{\theta})$, we look for the most likely sequence of states $\mathbf{z}_i$, $i = 1, \ldots, N$. In other words, we estimate the hidden sequence $\mathbf{z}_i$ using MAP:

$$\hat{\mathbf{z}}_i = \underset{\mathbf{z}_i}{\operatorname{argmax}}\, p(\mathbf{z}_i|\mathbf{y}_i, \hat{\psi}_i).$$

The Viterbi algorithm is a very efficient and fast decoding algorithm for computing the MAP estimate of the sequence of states in an HMM [23]. Note that conditional on $\hat{\psi}_i$, each set $\{\mathbf{z}_i, \mathbf{y}_i\}$ is a "classical" HMM. Then, the Viterbi algorithm can be easily implemented for computing $\hat{\mathbf{z}}_i$, $i = 1, \ldots, N$.

2.2.4 Describing daily seizures counts with mixed HMMs

The goal of the present section is to investigate the properties of the population parameter estimates obtained with the SAEM algorithm through simulation and to present a concrete application of mixed HMMs. As our application of interest is epilepsy, the simulation study is conducted in this context. The application to epileptic seizures follows.

2.2.4.1 Model development

As the supposition of discrete Markovian disease stages for describing evolution of seizures in epileptic patients is a common assumption in the literature – see for example [1] – we suggest developing a mixed HMM to describe seizure dynamics in a cohort of epileptic patients. We assume that epileptic subjects go through alternating periods of low and high epileptic susceptibility, and therefore consider a two-state Poisson mixed-HMM. The data include two groups of patients: one group are treated with a new anti-epileptic drug while the other is given a placebo.

Let y_{ij} denote the number of seizures recorded by patient i, $i = 1, \ldots, N$ on day j, $j = 1, \ldots, n_i$. Let z_{ij} be the state of patient i at day j. We assume that z_{ij} takes its values in $\{1, 2\}$, where state 1 and state 2 respectively stand for the low and high epileptic susceptibility stages. The conditional distributions of daily seizures for subject i are assumed to be Poisson distributions with parameters λ_{1i} and λ_{2i} in state 1 and state 2 respectively, where $\lambda_{1i} < \lambda_{2i}$. Let x_i be an indicator variable equaling 1 if patients receive a treatment against epilepsy, and 0 if they receive a placebo.

The sequence of states for patient i can be considered several ways. The simplest model is a mixture model where the consecutive states for each subject are assumed to be independent of one another. Then, the probability to be in state 1 or in state 2 stays constant over time and is not influenced by the nature of the previous states. Let p_{1i} and p_{2i} be the mixture proportions for the sequence of hidden states of subject i, that is: $p_{1i} = \mathbb{P}(z_{ij} = 1) = 1 - \mathbb{P}(z_{ij} = 2) = 1 - p_{2i}$ for all $j = 1, \ldots, n_i$. Then, the model becomes:

Model MIXT

$$\begin{aligned} \log(\lambda_{1i}) &= \mu_1 + \beta_1 x_i + \eta_{1i}, \\ \log(\alpha_i) &= \mu_2 + \beta_2 x_i + \eta_{2i}, \\ \lambda_{2i} &= \lambda_{1i} + \alpha_i, \end{aligned}$$

and

$$\text{logit}(p_{1i}) = \mu_3 + \beta_3 x_i + \eta_{3i},$$

where the random effects are i.i.d. centered Gaussian random variables:

$$\eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})' \underset{i.i.d.}{\backsim} \mathcal{N}(0, \Omega).$$

The anti-epileptic treatment is likely to influence the value of the mixture proportion and of the response parameters. Typically, it is expected to reduce the number of seizures in the two states ($\beta_1 < 0$, $\beta_2 < 0$) or even decrease the probability to be in the state of high epileptic activity ($\beta_3 > 0$).

If on the contrary epileptic patients are more likely to stay in the same state than to switch to the other state, a two-state HMM would incorporate the dependence between successive states as a homogeneous Markov chain. Each individual sequence of observations is then assumed to follow a two-state HMM. Let $p_{11,i}$ and $p_{21,i}$ be the state 1 to state 1 and the state 2 to state 1 transition probabilities for subject i ($p_{12,i} = 1 - p_{11,i}$ and $p_{22,i} = 1 - p_{21,i}$). λ_{1i}, λ_{2i}, $p_{11,i}$ and $p_{21,i}$ fully specify individual i's HMM. Then, the model is given by:

Model HMM

$$\begin{aligned}
\log(\lambda_{1i}) &= \mu_1 + \beta_1 x_i + \eta_{1i}, \\
\log(\alpha_i) &= \mu_2 + \beta_2 x_i + \eta_{2i}, \\
\lambda_{2i} &= \lambda_{1i} + \alpha_i, \\
\text{logit}(p_{11,i}) &= \mu_3 + \beta_3 x_i + \eta_{3i},
\end{aligned}$$

and

$$\text{logit}(p_{21,i}) = \mu_4 + \beta_4 x_i + \eta_{4i},$$

where the random effects are i.i.d. centered Gaussian random variables:

$$\eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i}, \eta_{4i})' \underset{i.i.d.}{\backsim} \mathcal{N}(0, \Omega).$$

Like above, the anti-epileptic treatment is expected to decrease the amount of seizures ($\beta_1 < 0$, $\beta_2 < 0$), and to increase transition probabilities towards the state of low epileptic activity ($\beta_3 > 0$, $\beta_4 > 0$). In model **HMM**, $\psi_i = (p_{11,i}, p_{21,i}, \lambda_{1i}, \alpha_i)'$,
$\phi_i = (\text{logit}(p_{11,i}), \text{logit}(p_{21,i}), \log(\lambda_{1i}), \log(\alpha_i))'$ and the population parameter vector is composed of μ_1, μ_2, μ_3, μ_4, β_1, β_2, β_3, β_4 and Ω's components.

2.2.4.2 Simulation study

We first present a simulation study aimed at investigating the properties of the SAEM algorithm to estimate the population parameters in the mixed HMM framework.

a) Design of the Monte Carlo study

The present study is inspired by the description of epileptic symptoms. The models used for the simulations derive from the mixed HMM described above. Our Monte-Carlo study envisages two scenarios based on the model **HMM**.

1. In the first series of simulations, the transition structure is chosen to be exactly the same for all the individuals. In other words, there is no inter-patient variability and no treatment effect on the transition probabilities $p_{11,i}$ and $p_{21,i}$:

   $$\text{logit}(p_{11,i}) = \mu_3,$$

   and

   $$\text{logit}(p_{21,i}) = \mu_4.$$

   Here, Ω is a 2×2 diagonal covariance matrix with diagonal terms ω_1^2 and ω_2^2.
2. The second series of simulations uses the full model **HMM** described in previous section, which includes four independent random effects. Then Ω is a 4×4 diagonal matrix with diagonal terms ω_1^2, ω_2^2, ω_3^2 and ω_4^2.

100 datasets including $N = 30$ individuals with $m = 20$ observations each, and 100 datasets with $N = 200$ subjects and $m = 100$ observations per subject are simulated in both situations. Each simulated dataset includes the same number of subjects in placebo ($x_i = 0$) and treatment ($x_i = 1$) groups. The values of the fixed effects and the variance of the random effects used in simulations are given in Tables 2.1 and 2.2.

For each simulated dataset, the population parameters and the standard errors (s.e.) of the estimated parameters are estimated.

Remark 3. It is assumed here that the treatment starts on the first day of the experiment. We use a non-informative distribution for the first state, i.e., $\mathbb{P}(z_{i1} = 1) = \mathbb{P}(z_{i1} = 2) = 0.5$. If the experiment starts after several days of treatment, we assume that each hidden Markov chain has reached its own stationary distribution on the first day. Then, these stationary distributions are used as initial distributions.

b) Results

Table 2.1 and Table 2.2 respectively display the results of the Monte Carlo study for the models with two and four random effects. For each set of 100 simulations and for each of the model parameters, we display:

i) the sample mean:

$$\bar{\theta} = \frac{1}{100}\sum_{k=1}^{100}\hat{\theta}_k,$$

ii) the sample standard deviation:

$$\text{sd} = \sqrt{\frac{1}{100}\sum_{k=1}^{100}(\hat{\theta}_k - \bar{\theta})^2},$$

iii) the mean estimated standard error:

$$\bar{\text{se}} = \frac{1}{100}\sum_{k=1}^{100}\widehat{\text{se}}_k.$$

As expected, the estimations are poorer when the amount of data is small ($N = 30$, $m = 20$). Even if the mean values of the estimates of all the population parameters are quite close to the true values, a non-negligible bias still exists. Above all, the sample standard deviation of the estimates is very large, especially for the variance parameters. As an example, regarding the model with two random effects, $\text{sd}(\omega_1^2)$ is greater than $\bar{\omega^2}_1$. This suggests that even in models with a limited number of random effects, the variance parameters are not well estimated from small datasets. Moreover, the SAEM algorithm sometimes fails in estimating the Fisher information matrix when the amount of data is not large enough. Indeed, the estimation of the Fisher Information matrix could not be performed on 9 occasions for the model with two random effects and on 64 occasions for the model with four random effects.

When increasing the number of subjects ($N = 200$, $m = 100$), we are closer to asymptotic conditions and get much better results. On the one hand, the sample mean of the estimates are closer to the corresponding true values whatever the number of random effects in the model. The estimated standard errors derived from the estimation of the Fisher Information Matrix provide an accurate approximation of the standard deviations of $\hat{\theta}$'s components since the estimated standard errors reported in Tables 2.1 and 2.2 are similar to the standard deviations when the datasets include $N = 200$ individuals. In both cases, the standard errors for all parameters of the model are low, which suggests that the asymptotic variance of the population parameter estimates are small.

We can also notice that adding random effects and a treatment effect to the transition probabilities does have an impact on the estimation of parameters μ_3 and μ_4. The bias of

$\widehat{\mu}_3$ and $\widehat{\mu}_4$ slightly increases in comparison with the scenario without any variability on these two parameters, while the standard errors are also twice as large.

Lastly, only 300 seconds of CPU time is required on a laptop (processor Intel Core Duo 2.94 GHz) for estimating the whole set of population parameters in the most complex of the above simulated models (including a binary covariate and four random effects) from a dataset with 200 subjects and 100 observations per subject.

[3] used similar mixed HMMs to illustrate the performance of their estimation method. Their Monte Carlo study was based on datasets with 30 subjects and 20 observations per individual, but only models with zero, one or two random effects were investigated. They reported little bias in their estimates in general, as well as good similarities between the estimated standard errors and the standard deviations of the parameter estimates. The quality of their results slightly decreases with the number of random effects included in the mixed HMM. However, when the number of random effects exceeds three, they reported estimation times of the order of one day for a single dataset. Our estimation algorithm runs much faster, and allows fitting complex models to relatively large datasets.

Remark 4. Although the model chosen in the present section for the simulations is a two-state Poisson mixed HMM, the inference methodology presented in Section 2.2.3 also applies to more general models – including more than two states and other response distributions – and successfully estimates the parameters from large datasets whatever the number of random effects in the model.

c) Application to daily seizure counts data

i) The data

The initial data consisted of sequences of daily seizures in a sample of 507 epileptic patients. The observations were collected during a randomized placebo-controlled clinical trial aimed at studying the potential efficacy of a given anti-epileptic treatment.

The clinical trial consisted of two consecutive phases, each of a duration of 12 weeks on average. The included epileptic patients were first observed during the screening phase before being randomly attributed either a placebo or the active medication. The follow-up duration widely varied from one patient to another, with screening periods from 23 days to 230 days and treatment periods from 4 to 130 days.

For the present application, we only consider observations during the active treatment phase. This dataset contains a total of 41026 seizure from 307 patients attributed the placebo and 200 patients receiving the active form. Figure 2.4 displays the observed seizure counts of four typical subjects.

We now aim to fit the models **MIXT** and **HMM** to these data.

ii) Results

Results obtained with models **MIXT** and **HMM** are displayed in Tables 2.3 and 2.4. Based on the comparison of the Bayesian Information Criteria (BIC) of the two fitted models (78318 vs 76589), we come to the conclusion that the two-state mixed HMM clearly allows a better description of the seizure counts data than the mixed mixture model. Assuming the existence of hidden states appears to be a natural way to describe the evolution of seizure

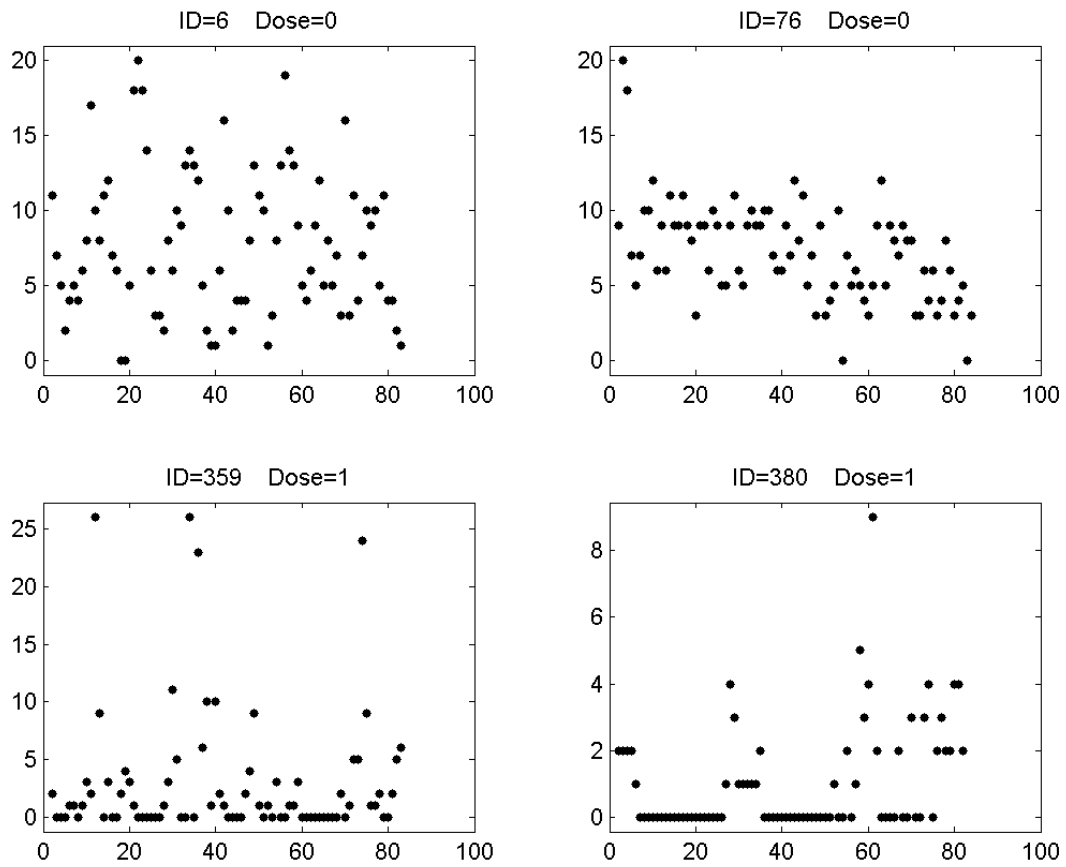

Figure 2.4: Observed seizure counts of four typical subjects.

counts in epileptic patients over time, and considering that these sequences of hidden states have Markovian dynamics establishes a real contribution in modeling epileptic symptoms.

Table 2.4 displays the MLEs of the parameters of model **HMM**. We see that the estimated standard errors of the parameters are small, suggesting that the population parameters are estimated accurately. Parametrization of model **HMM** involves a log-transform and logit-transform of the original parameters of the model. It is convenient for the estimation step, but the original parametrization (λ_1, λ_2, $p_{11,}$, $p_{21,}$) should be used for a comparison of the two groups. Table 2.5 displays the MLEs of the original parameters in both groups of epileptic patients as well as the stationary distribution ($\mathbb{P}(z_{ij}=1), \mathbb{P}(z_{ij}=2)$) of the hidden Markov chain, the mean number of seizures in each state $\mathbb{E}(y_{ij}|z_{ij}=1)$, $\mathbb{E}(y_{ij}|z_{ij}=2)$, and the global mean number of seizures $\mathbb{E}(y_{ij})$. Table 2.5 suggests a beneficial effect of the drug on the mean number of daily seizures in both states. Indeed, the population values of λ_1 and λ_2 are lower in the group of treated patients as well as $\mathbb{E}(y_{ij})$. However, the drug does not have the expected effect on the dynamics of the Markov chain, the fraction of time spent in the state of high epileptic activity being slightly increased in treated patients.

As well as this, estimates of the variances ω_m^2, $m = 1, \ldots, 4$ indicate that the dynamics and

intensity of seizures are largely heterogeneous among epileptic patients. The mean number of daily seizures in the state of low epileptic activity as well as the difference between the counts of seizures in the two pre-supposed epilepsy stages appear to be particularly variable between subjects.

Next, we estimated individual parameters by computing the MAP estimates for each subject and each individual state sequence with the Viterbi algorithm. Figure 2.5 displays the estimated states obtained for four typical subjects. On each graph, the observations (daily seizures) are represented as a function of time (number of days). The estimated sequences of states show that the patients actually alternate periods in two states characterized by low and high counts of seizures. These four patients also illustrate the high inter-patient variability in the data. This variability clearly justifies the use of a mixed model.

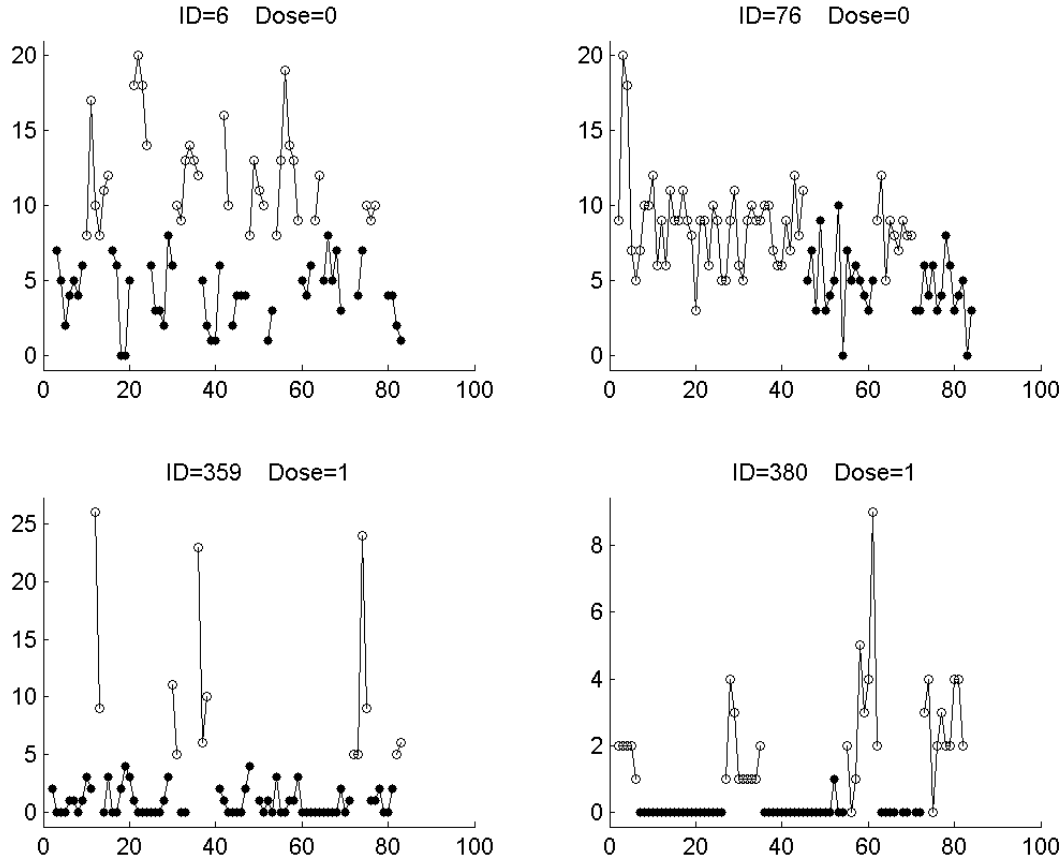

Figure 2.5: Observed seizure counts and estimated sequences of states of four typical subjects ; • : state 1, ∘ : state 2.

Finally, we show some specific Visual Predictive Checks (VPC) for model assessment. Model diagnostics are based on the comparison of some statistics computed from the observed data with the theoretical distribution of these statistics under the model(s) to assess. More precisely, a confidence interval of the statistics is computed and displayed together with the

observed statistics. When this theoretical distribution cannot be computed in closed form, it is estimated by Monte-Carlo using simulated datasets.

The Marginal distribution of the daily seizures count (y_{ij}) is displayed in Figure 2.6. We compare the empirical distribution of the observations with the theoretical distributions assuming a single Poisson model (one state model) and a mixture of two Poisson models (two state model). Here, 100 datasets were simulated to derive 90% confidence intervals of the cumulative distribution functions under both models. These datasets were simulated using the design of the original dataset (same number of subjects, same numbers of observations per subject, same treatments). The parameters of the single Poisson model used for the simulation were estimated from the dataset ($\lambda = 0.465$ in the placebo group, $\lambda = 0.369$ in the treatment group and $\omega = 1.01$). The parameters of the mixture model used for the simulation are those given in Table 2.3. It is obvious that the two state model describes much better the observed distribution of the data than the one state model.

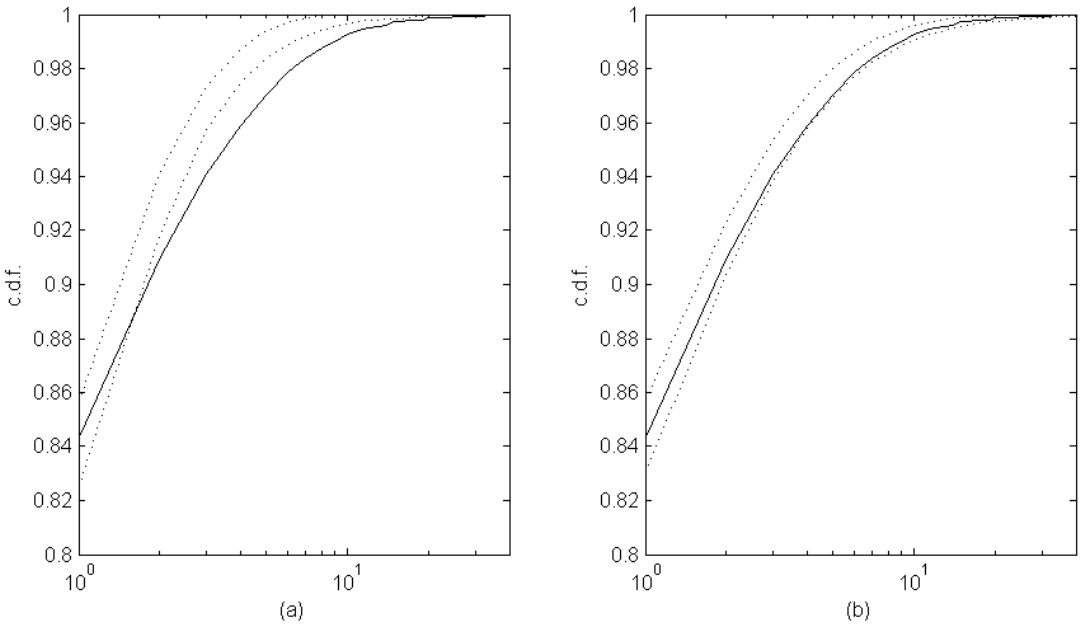

Figure 2.6: Marginal distribution of daily seizures counts. The distribution of the observations (solid line) is displayed together with a 90% confidence interval (dotted line), assuming: (a) a one state model, (b) a two state model.

The sole marginal distribution of the observations doesn't allow us to assess the underlying dynamics of the states. Indeed, both mixture (**MIXT**) and hidden Markov (**HMM**) models assume a two state model. Since the observations are assumed to be independent under (**MIXT**) but not under (**HMM**), we propose to base our second model diagnostic on the correlations between consecutive observations. Let ρ_i be the empirical correlation for subject i:

$$\rho_i = \frac{\sum_{j=1}^{n_i-1}(y_{i,j} - \bar{y}_i)(y_{i,j+1} - \bar{y}_i)/(n_i - 1)}{\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_i)^2/n_i}.$$

The distribution of the correlations (ρ_i) is displayed Figure 2.7. For the VPCs, we simulated

100 datasets under models **MIXT** and **HMM** using the parameters given in Tables 2.3 and 2.4 respectively. We observe that the distribution of the observed correlations is better described assuming a Markov chain (model **HMM**) than assuming a sequence of independent states (model **MIXT**).

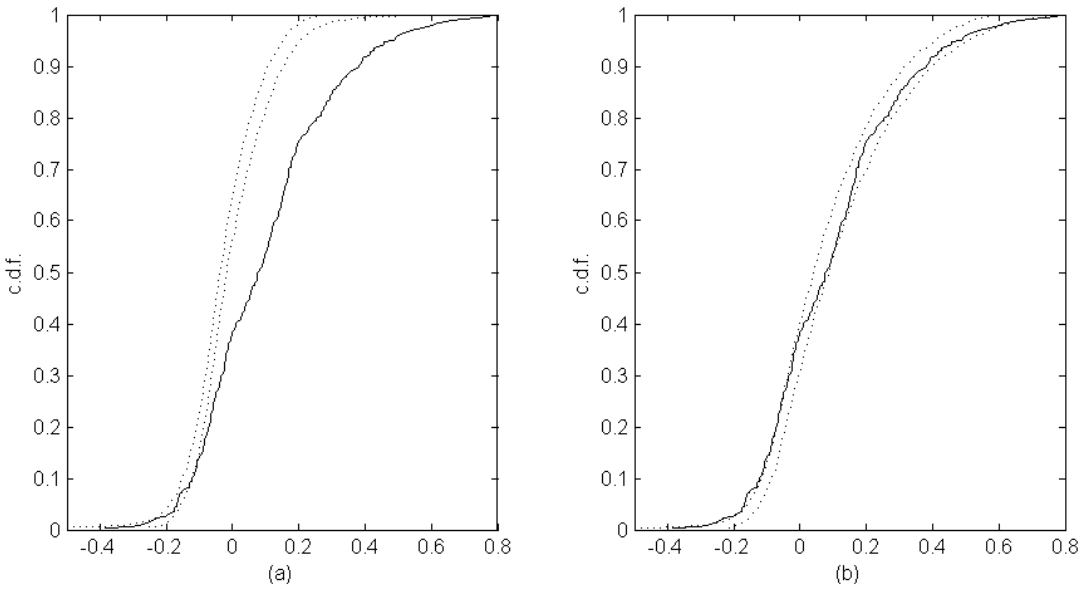

Figure 2.7: Distribution of the correlations between consecutive observations. The distribution of the observed correlations (solid line) is displayed together with a 90% confidence interval (dotted line), assuming: (a) a mixture model **MIXT**, (b) a hidden Markov model **HMM**.

2.2.5 Discussion

Mixed HMMs have recently been defined as the extension of HMMs to population studies [3]. We have proposed a new inference method for these models for which maximum likelihood estimation is a complex issue. In particular, the SAEM algorithm has been successfully adapted to deal with the estimation of the population parameters of mixed HMMs. The Baum-Welch algorithm, used for computing the individual conditional likelihoods, is the key to the developed estimation method. Our methods are fast, even for models with a large number of random effects. Moreover, we have shown that the population parameters estimated with the SAEM algorithm have negligible bias and that the standard errors of the estimates obtained with SAEM correctly evaluate the standard deviations of the population parameter estimates. We have developed a two-state Poisson mixed HMM including a treatment effect to describe the daily seizure count in epileptic patients. The assumption of distinct disease stages is realistic and easily interpretable. Estimation has been successfully performed under this model using our new algorithms, giving satisfying initial results.

Further improvements of the present application of mixed HMMs could be considered in future work. The first questions of interest would be about the general nature of the latent pro-

cess. The main assumption of the study is that the evolution of epilepsy in epileptic patients follows a two-state first-order Markov chain. We have shown that the first-order Markovian dependence between states leads to a better description of the data than a simple mixture structure. However, a thorough analysis of the seizure data would require comparing the present HMM model against models assuming more than two states or higher-order Markov dependence. Then, other response distributions could be considered for handling substantial overdispersion in the observations of some subjects or underdispersion in others. A generalized Poisson distribution could be considered instead of a Poisson distribution. Correlations between the model's random effects in the population model could also be studied.

The present article also raises many theoretical and methodological questions for future studies. First, the good practical properties of the population parameter estimate obtained with the SAEM algorithm suggest good theoretical properties of the maximum likelihood population parameter estimates in mixed HMMs. The asymptotic properties of the MLE is the subject of a parallel work. Second, model selection in the context of mixed HMMs may require more complex criteria than the BIC for a better consideration of the number of hidden states, for example. Finally, as this work is devoted to mixed models with discrete and finite state space, a natural development would be to extend the inference methodology to models with continuous state spaces.

Acknowledgements

The authors would like to thank Pfizer Inc. for making the pregabalin data set available for applying the proposed methodology to a real data example, and Mats Karlsson (Uppsala University), Rada Savic (Standford University) and Raymond Miller (previously Pfizer Inc.) for fruitful discussions about mixed HMMs and applications to daily seizure counts modelling, and Kevin Bleakley for his help in improving the quality of the paper.

2.2.6 Tables

		N=30 ; m=20			N=200 ; m=100		
Parameter	True value	$\bar{\theta}$	sd	$\bar{\text{se}}$	$\bar{\theta}$	sd	$\bar{\text{se}}$
μ_1	-1.00	-1.23	0.30	0.35	-1.02	0.06	0.07
β_1	-2.00	-2.89	1.09	1.42	-2.09	0.24	0.18
μ_2	1.40	1.36	0.12	0.11	1.39	0.03	0.03
β_2	-1.00	-1.06	0.17	0.18	-1.00	0.05	0.05
μ_3	0.85	0.74	0.19	0.20	0.84	0.03	0.03
μ_4	-0.40	-0.42	0.21	0.21	-0.40	0.03	0.03
ω_1^2	0.30	0.49	0.53	0.75	0.31	0.06	0.06
ω_2^2	0.10	0.12	0.06	0.06	0.10	0.01	0.01

Table 2.1: Monte-Carlo experiment: estimation of the population parameters in the model **HMM** with two random effects. The table displays the true values of the parameters, the means and standard deviations of the estimated parameters, and the mean estimated standard errors. The estimation of the standard errors failed on 9 of the simulated datasets with $N = 30$ individuals and $m = 20$ observations per subject.

		N=30 ; m=20			N=200 ; m=100		
Parameter	True value	$\bar{\theta}$	sd	$\bar{\text{se}}$	$\bar{\theta}$	sd	$\bar{\text{se}}$
μ_1	-1.00	-1.21	0.28	0.35	-1.03	0.06	0.07
β_1	-2.00	-2.59	1.50	1.20	-2.01	0.15	0.15
μ_2	1.40	1.36	0.13	0.13	1.39	0.03	0.03
β_2	-1.00	-1.05	0.20	0.24	-1.00	0.05	0.05
μ_3	0.85	0.89	0.23	0.42	0.88	0.07	0.06
β_3	1.00	1.09	0.51	0.78	0.97	0.10	0.10
μ_4	-0.40	-0.55	0.25	0.32	-0.43	0.06	0.06
β_4	-1.00	-0.87	0.55	0.70	-0.95	0.10	0.10
ω_1^2	0.30	0.47	0.40	0.70	0.30	0.06	0.07
ω_2^2	0.10	0.12	0.07	0.62	0.10	0.01	0.01
ω_3^2	0.30	0.29	0.13	0.52	0.26	0.05	0.05
ω_4^2	0.30	0.31	0.17	0.08	0.26	0.05	0.05

Table 2.2: Monte-Carlo experiment: estimation of the population parameters in the model **HMM** with four random effects. The table displays the true values, the means and standard deviations of the estimated parameters, and the mean estimated standard errors. The estimation of the standard errors failed on 64 of the simulated datasets with $N = 30$ individuals and $m = 20$ observations per subject.

Parameter	Estimate	s.e.	r.s.e. (%)	P-value
μ_1	-1.45	0.08	6	
β_1	-0.50	0.14	28	3.10 10^{-4}
μ_2	0.14	0.11	78	
β_2	-0.24	0.14	60	0.09
μ_3	2.28	0.16	7	
β_3	-0.43	0.23	53	0.058
ω_1^2	1.60	0.13	8	
ω_2^2	1.26	0.14	11	
ω_3^2	1.57	0.27	17	
BIC=78318				

Table 2.3: Estimation of the population parameters for the seizure counts data with the mixture model. The table displays the parameter estimates, the absolute and relative estimated standard errors, and the $p-values$ of the tests $\beta_m = 0$, $m = 1, \ldots, 3$ (Wald test).

Parameter	Estimate	s.e.	r.s.e. (%)	P-value
μ_1	-1.64	0.09	6	
β_1	-0.60	0.15	26	1 10^{-5}
μ_2	-0.19	0.09	48	
β_2	-0.15	0.13	90	0.27
μ_3	2.31	0.08	3	
β_3	-0.25	0.12	50	0.047
μ_4	-0.20	0.12	60	
β_4	-0.24	0.18	74	0.18
ω_1^2	1.79	0.17	9	
ω_2^2	1.51	0.15	10	
ω_3^2	0.33	0.07	20	
ω_4^2	1.22	0.17	14	
BIC=76589				

Table 2.4: Estimation of the population parameters for the seizure counts data in the mixed HMM. The table displays the parameter estimates, the absolute and relative estimated standard errors, and the $p-values$ of the tests $\beta_m = 0$, $m = 1, \ldots, 4$ (Wald test).

Parameter	Placebo	Treatment
λ_1	0.19	0.11
λ_2	1.03	0.82
p_{11}	0.91	0.89
p_{21}	0.45	0.39
π_1	0.83	0.78
π_2	0.17	0.22
$\mathbb{E}(y_{ij}\|z_{ij}=1)$	0.48	0.26
$\mathbb{E}(y_{ij}\|z_{ij}=2)$	2.23	1.77
$\mathbb{E}(y_{ij})$	0.84	0.67

Table 2.5: Comparison of the placebo and treatment arms using the two-state mixed HMM. The table displays the maximum likelihood estimates of the parameters of both distributions as well as a summary of both stationary distributions. These are derived from the estimates of Table 2.4 by Monte Carlo methods.

2.3 Deuxième article: Analysis of exposure-response of CI-945 in patients with epilepsy: application of novel Mixed Hidden Markov Modeling Methodology

Cet article, réalisé en collaboration avec Radojka Savic, Raymond Miller, Mats Karlsson et Marc Lavielle est accepté pour publication dans la revue Journal of Pharmacokinetics and Pharmacodynamics [14].

Abstract

We propose to describe exposure - response relationship of an antiepileptic agent, using mixed hidden Markov modeling methodology, to reveal additional insights in the mode of the drug action which the novel approach offers. Daily seizure frequency data from six clinical studies including patients who received gabapentin were available for the analysis. In the model, seizure frequencies are governed by underlying unobserved disease activity states. Individual neighbouring states are dependent, like in reality and they exhibit their own dynamics with patients transitioning between low and high disease states, according to a set of transition probabilities. Our methodology enables estimation of unobserved disease dynamics and daily seizure frequencies in all disease states. Additional modes of drug action are achievable: gabapentin may influence both daily seizure frequencies and disease state dynamics. Gabapentin significantly reduced seizure frequencies in both disease activity states; however it did not significatively affect disease dynamics. Mixed hidden Markov modeling is able to mimic dynamics of seizure frequencies very well. It offers novel insights into understanding disease dynamics in epilepsy and gabapentin mode of action.

Keywords: Mixed hidden Markov model, epilepsy, CI-945, maximum likelihood estimation, Monolix.

2.3.1 Introduction

Characterizing, understanding and quantifying the time course of disease progression in epilepsy and treatment effect for number of antiepileptic agents have posed a significant challenge in the area of quantitative clinical pharmacology. The challenges are rooted both in the complex nature of the collected seizure data itself as well as limited methodology available to analyse such data. The epileptic activity is most often summarized as number of seizures per time unit, with time unit being a day, a week, or a month. Individual time profiles of epileptic seizures are often erratic with distinct periods of low or no activity followed by the periods of high epileptic activity, leading to the substantial variability within a patient. The count data analysis for repeated measurements has been most commonly used methodology for the analysis of seizure data. Miller *et al.* used Mixed Poisson model to describe exposure response relationship of pregabalin [22]. Similar was the case in the work by Snoeck *et al.*, describing dose response relationship for levetiracetam [26]. Trocoñiz *et al.* pointed out to the important drawback of the commonly used Mixed Poisson model, illustrated on the analysis of the epileptic data [28]. The first drawback is that the model is not able to handle common

large variance in individual counts (e.g. overdispersion phenomena). Authors suggested using an alternative negative binomial model which is an elegant way to deal with observed high individual variances (overdispersion) by introduction of an additional parameter; however it remains empirical. The second drawback of current methodology is related to the assumption that neighbouring observations, e.g. seizure frequencies at two consecutive days, are independent. This is not the case in reality, since the days with low seizure frequencies are often preceded by the days with the similar pattern. This aspect of seizure frequencies data analysis represents the key challenge. Previous authors suggested use of Markov elements, where the seizure frequency was conditioned whether the previous day was a seizure day or not. However, more mechanistic understanding of the seizure patterns is needed. The likely reason for the observed erratic profiles in epileptic patients is the fact that the disease itself consists of two disease states, a state of low and a state of high epileptic activity. Patients are transitioning between these two disease states and transitions may be triggered by number of factors. When at low state, patients are experiencing no or low number of seizures, and when in high active state, the seizures are frequent. Since epilepsy cannot be cured, the objective of the current therapies is to reduce the average number of daily seizures. This can be achieved by prolonging the periods in low disease activity, or by reducing the expected number of daily seizures in each state. Even though clearly present, the states of the disease activity are not directly observed, but the consequential events (seizures) are. To handle the analysis of such data type, an appropriate methodology is needed. Hidden Markov Model (HMM) is a methodology where biological system is seen as a Markov process where the states are not directly visible, however output, which is dependent on the state, is observable. This model structure is directly applicable to the epilepsy type of data, where the observed seizure frequencies would represent the observed output, while the unobserved state would represent disease activity. The hidden Markov methodology has been successfully utilized in different areas, such as bioinformatics [17], speech recognition [23] and gene prediction [18]. However, the analysis of the large epilepsy trials requires the extended methodology which can also provide the estimates of the population (between subject) variability in the baseline, placebo and drug effect parameters, often referred to as mixed effect analysis. To date, the methodology for the implementation of the mixed hidden Markov model (MHMM) has been proposed and illustrated on the example of modelling the longitudinal MRI counts in multiple sclerosis patients [3]. Mixed-effects hidden Markov models have then been used with other methodologies by Rijmen *et al.* for the analysis of the treatment effect on temporal pattern of symptom burden in brain cancer [24], or by Maruotti *et al.* for the description of the number of episodes of side effects [21]. However no other application to the field of quantitative clinical pharmacology has been presented up to date. Our aim was (i) to implement the mixed hidden Markov methodology using an SAEM algorithm for the wider use within an area of quantitative clinical pharmacology and (ii) to describe exposure response relationship of an antiepileptic agent CI 945 using the mixed Hidden Markov Model and to reveal additional insights in the mode of the drug action which the novel approach offers.

2.3.2 Methods

2.3.2.1 Data base

The data base consisted of six clinical studies: Phase 2 and Phase 3 double blind placebo controlled parallel group multicenter studies. All recruited patients were on standard anti-

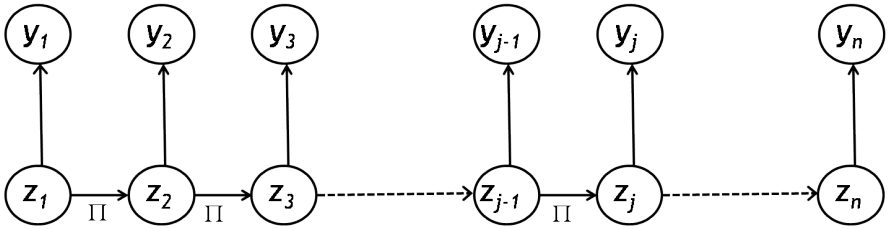

Figure 2.8: Architecture of the Hidden Markov Model: (z_j) is a non observed 2-states Markov chain with transition matrix $\Pi = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$. Conditionally to the sequence (z_j), (y_j) is a sequence of independent Poisson random variables with intensity λ_1 if $z_j = 1$ and λ_2 if $z_j = 2$.

epileptic therapy and they completed 12 weeks baseline screening phase. Thereafter, patients were randomized to parallel treatment groups receiving placebo or active treatment (gabapentin 0.45, 0.6, 0.9, 1.2 and 1.8g). Overall, time profiles from 788 patients were included into the data base. The data consisted of baseline daily counts of epileptic seizures measured over 12 weeks followed by 12 weeks of active treatment administration. As there were only two patients receiving 0.45g of gabapentin, these were excluded from the model building part. A brief description of the trials is given in Table 2.6.

2.3.2.2 Mixed hidden Markov Model

A first basic model proposed in [28] assumes that the daily number of seizures are independent Poisson random variables with intensity λ. In this model, λ is both the mean and the variance of the number of daily seizures, but this approach is not satisfactory because of overdispersion in the data [28]. In order to better take into account this overdispersion, the authors proposed a mixture of two Poisson distributions with parameters λ_1 and λ_2 with $\lambda_1 < \lambda_2$. In other terms, the model postulated existence of two disease states of epileptic activity, low (state 1 with parameter λ_1) and high (state 2 with parameter λ_2). This mixture model fits well the marginal distribution of the observations. However, assuming the states to be independent is not realistic. indeed, epileptic patients are reasonably expected to stay in the same state several days rather than switching randomly every day between states. We propose to describe the dynamics of the states by using a Markov chain. We have used first order Markov chains (i.e. with memory 1) indicating that it is enough to know the state on day $j-1$ for predicting the state on day j. Then, each patient is transitioning between two states according to individual transition probabilities p_{11} and p_{21}. Here, p_{11} is the probability of remaining in low state, while p_{21} is the probability of transitioning from high state to low state. From p_{11} and p_{21}, probability of transitioning to the high state (p_{12}) and probability of remaining in an active state (p_{22}) are easily calculated. The distribution of the number of daily seizures in each state remains a Poisson distribution with parameters λ_1 and λ_2. The architecture of the mixed Hidden Markov Model is shown in Figure 2.8.

It is straightforward to extend this model to any number of states and higher order Markov chains. Nevertheless, identification of such complex models can be a strong issue in practice. An other possible extension is to consider that the Poisson intensity is a continuous stochastic process $\lambda(t)$. Such extension is beyond the scope of the paper.

Model building procedure involved at the first stage development of the baseline model where the 12-week baseline data were utilized. Thereafter followed the development of the placebo and drug model where different mechanisms of drug actions were explored. The baseline model for patient i involved a constant baseline defined with four parameters: p^B_{11i}, p^B_{21i}, λ^B_{1i} and λ^B_{2i}, defined as:

$$\begin{aligned} \text{logit}(p^B_{11i}) &= \beta_1 + \eta_{1i}, && (2.14)\\ \text{logit}(p^B_{21i}) &= \beta_2 + \eta_{2i}, && (2.15)\\ \log(\lambda^B_{1i}) &= \log(\lambda_1) + \eta_{3i}, && (2.16)\\ \lambda^B_{2i} &= \lambda^B_{1i} + \alpha^B_i, \end{aligned}$$

where

$$\log(\alpha^B_i) = \log(\alpha) + \eta_{4i}. \quad (2.17)$$

Subscript i refers to an individual while superscript B to the baseline parameter. Actual estimated population parameters were (β^B_1, β^B_2, λ^B_1, α^B) and Ω, the diagonal covariance matrix defining random effects for four baseline parameters, which after appropriate transformations define basic parameters for the Hidden Markov Model. Additional models describing baseline mean counts (λ^B_1, α^B) as a linear and nonlinear function of time were also investigated. Once baseline model was developed and evaluated, development of the placebo and drug model followed. Placebo effect (δ_{1-4}) was introduced as a step function after the treatment onset on all 4 parameters defining the seizures baseline. Similarly, gabapentin drug effect was introduced as a linear function of dose ($\gamma_{1-4}Dose_i$) on log or logit transformations of all 4 parameters of the baseline model. The drug effect is assumed to be fixed in the population. In order to handle effect delay, additional functions describing time course of the effect onset were introduced. This was modeled as a first-order delay with estimated half lives τ_1 for λ_1 and τ_2 for λ_2. Therefore, the entire mixed Hidden Markov model may be summarized by following equations:

$$\begin{aligned} \text{logit}(p^T_{11i}) &= \text{logit}(p^B_{11i}) + (\delta_1 + \gamma_1 Dose_i + \eta_{5i}), && (2.18)\\ \text{logit}(p^T_{21i}) &= \text{logit}(p^B_{21i}) + (\delta_2 + \gamma_2 Dose_i + \eta_{6i}), && (2.19)\\ \log(\lambda^T_{1i}) &= \log(\lambda^B_{1i}) + (\delta_3 + \gamma_3 Dose_i + \eta_{7i})(1 - 2^{-t/\tau_1}), && (2.20)\\ \log(\alpha^T_i) &= \log(\alpha^B_i) + (\delta_4 + \gamma_4 Dose_i + \eta_{8i})(1 - 2^{-t/\tau_2}), && (2.21)\\ \lambda^T_{2i} &= \lambda^T_{1i} + \alpha^T_i. \end{aligned}$$

2.3.2.3 MHMM's estimation methods

The analysis was performed in Monolix 3.1 and Matlab software. Following estimation methods and sequences were used: (i) First step involved estimation of population parameters (means and variances) where the maximum likelihood approach was utilized. Mixed hidden Markov model belongs to the group of incomplete data models, where the likelihood cannot be explicitly computed. Therefore, the forward procedure of the Baum-Welch algorithm [23] linked with SAEM [27] algorithm was utilized. (ii) Individual parameters were computed using MAP (Maximum A Posteriori) method [27]. (iii) Thereafter, most likely sequences of the Hidden Markov States were decoded for each individual by use of Viterbi algorithm [23]

and individual parameter estimates. Finally, the likelihood was estimated using an importance sampling method [27].

2.3.2.4 Model evaluation

Individual fits combined with individual estimates of the underlying disease state dynamics were used to visualize the model fits. Visual predictive checks were performed, and these were stratified by important trial features (treatment effect, treatment period). Additionally, in order to assess benefit of Mixed Hidden Markov Model over mixture of two Poisson distribution models without Markov elements, posterior predictive checks were performed to assess the number of transitions between different seizure frequencies.

2.3.3 Results

The overall data base consisted of 134396 daily seizure counts from 788 patients divided into a placebo group (307 individuals) and an active treatment group. Total of 71162 (resp. 63234) observations were in the screening (resp. treatment) phase, with 28158 (resp. 25400) observations from the placebo and 43004 (resp. 37834) from the active treatment. Model fit with Mixed Hidden Markov model was significantly better (Bayesian Information Criteria (BIC) =260881) compared to the standard Poisson model (BIC=296907) and mixture Poisson model (BIC=268266). A constant baseline model best described screening phase in epileptic patients, based on data from the screening phase (788 individuals, 71162 observations) indicating no time related changes, e.g. disease progression, in the baseline frequencies during 12 weeks of the screening period. Inspection of the raw data revealed features such as: substantial noise, variable baseline, negligible baseline changes and similarities between six different studies. The other baseline models tested included linear baseline model, exponential baseline model and a mixture model, without underlying Markov chain, however with no substantial improvements.

Based on the established baseline model, following was revealed using novel Mixed Hidden Markov Model methodology: probability of remaining in the low activity disease state was high (mean = 0.85, CV = 12%) translating into low probability of transitioning into the high seizure activity state conditioned on the current low seizure activity state. On the other hand, probability of transitioning from high activity state to the low activity state was also high (mean = 0.79, CV = 17%) indicating only 0.21 risk of remaining in the high activity on average.

Remark 5. The typical values $1/(1+e^{-\beta_1}) = 0.88$ and $1/(1+e^{-\beta_2}) = 0.83$ are the medians of the distributions of p_{11} and p_{21}. They are computed using the maximum likelihood estimations of β_1 and β_2 (1.96 and 1.56) indicated in Table 2.7. The mean and standard deviations of these distributions cannot be computed in a closed form. They are estimated by Monte Carlo method.

Using the Maximum Likelihood Estimates of $\log(\lambda_1)$ and $\log(\alpha)$, which are -2.01 and -0.08 and their variances (see Table 2.7), we can also compute by Monte-Carlo method the mean of the seizure counts intensities in each state. These were 0.34 and 2.25 for low and high activity state respectively associated with large inter patient variability (CV = 190% and CV = 140% respectively). The placebo model was best described with the simple step function where placebo effect was estimated for all major 4 parameters of the Hidden

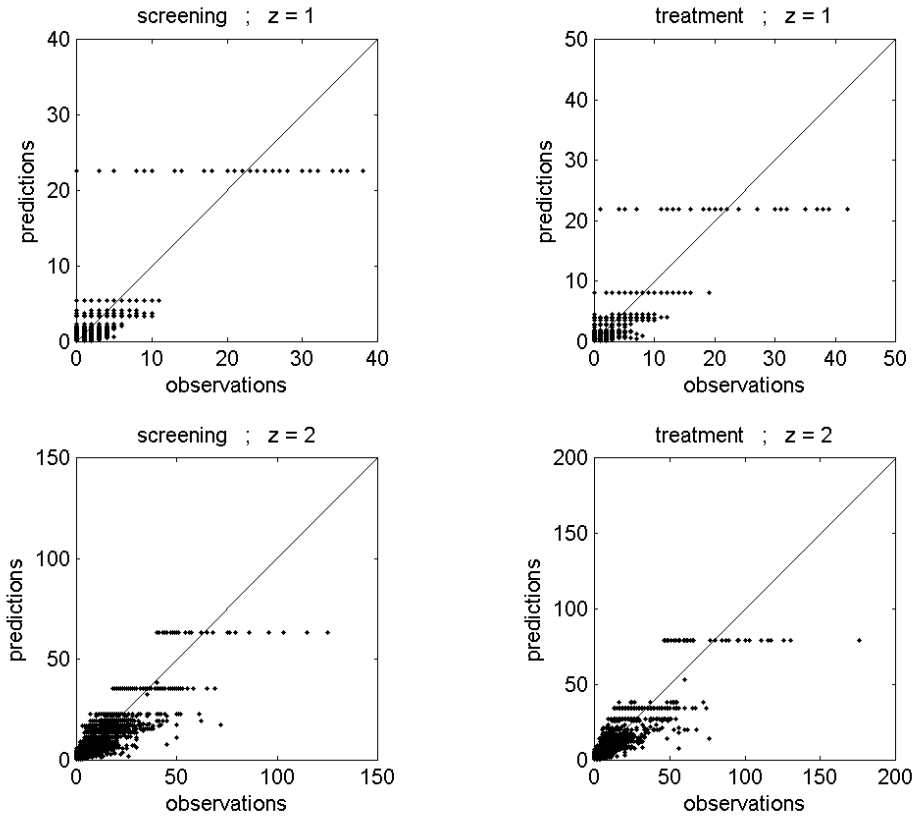

Figure 2.9: Comparison of estimated mean counts with the observed individual mean counts in both states in baseline and treatment periods.

Markov Model (p_{11}, p_{21}, λ_1, λ_2). Even though all placebo effects on all four parameters were statistically significant, the impact on the model was extremely modest: mean probability of remaining in low state slightly increased to 0.86, while probability of transitioning from high activity state to the low activity state also increased to 0.82. Small estimated values of δ_3 and δ_4 which are -0.05 and 0.07 indicate that median count intensities in both states remain stable, but due to the inter-patient variability of the placebo effect, mean count intensities in low and high activity states slightly increased to 0.39 and 2.83 respectively. Placebos effects were accompanied with rather modest interindividual variabilities ($< 15\%$ CV). Introduction of gabapentin treatment significantly decreased seizure frequencies in both disease states in dose-dependent fashion. The decrease in seizure frequencies with the highest dose (1800 mg) was substantial (15% and 29% for low and high activity state in average) translating into the seizure frequencies of 0.08 and 0.69 for low and high disease activity state respectively. On the other hand, the drug effect on the transitions between states was not significant ($< 1\%$ with the highest dose). The half maximal effect was reached within 1.7 and 16 days for low and high disease activity state respectively. Figure 2.9 shows comparison of estimated mean counts with the observed individual mean counts in both states and both periods (baseline and treatment).

Left panel of Figure 2.10 indicates few observed individual profiles while right panel showing "individual fits" with different colors indicating two predicted seizure activity states.

Figure 2.11 shows posterior predictive check for daily transitions between different seizure counts. The empirical transition probabilities derived from the screening observations respectively belong to the MHMM confidence intervals but not to those of the Poisson model and the mixture model.

And finally, Figure 2.12 demonstrates that marginal distribution of the daily seizure counts is better described with a HMM with two Poisson distributions than a single Poisson model.

2.3.4 Discussion

To our knowledge, this is a first application of Mixed Hidden Markov Modeling methodology to describe development of an exposure-response model from a large clinical data base. In a previous published application, Altman applied MHMM to model MRI time course in MS patients [3]. Even though this author pioneered the work in the area of Mixed Hidden Markov Model, she also pointed out that she was able to estimate only single or double random effect and the work regarding the estimation of more random effects was ongoing. Here, we were able to develop rather complex model with several random effects which were successfully and precisely estimated. Our model indicated that patients already in the screening phase had high probability of remaining in the low activity state. Also, patients had high probability of transitioning from high activity state to the low activity state, favouring stable conditions in these patients. Start of the therapy in the placebo group increased somewhat these probabilities. The major change in the placebo group was somewhat increased epileptic frequency in the high activity state, which may be associated with the increased stress of the clinical trial onset. Application of the MHMM also revealed significant decrease in seizure frequencies in both disease activity states, in dose dependent fashion. However, gabapentin does not exhibit any effect on the time course of disease states sequence and transition probabilities. This may suggest that gabapentin may not have any effect on changing the course or dynamics of the disease, however, the number of seizures will be decreased in periods of both low and high epileptic activities. We can consider disease state dynamics to be representative of the natural course of the disease progression accompanied with certain seizure frequency distributions in both states, which may be observed as indicator of symptoms severity of the diseases. Therefore, one may conclude that gabapentin does not exhibit disease modifying effect in epileptic patients; however it demonstrates significant symptomatic effect with decreased seizure frequencies in dose and time dependent fashion. In other words, patients will keep transitioning from one to another disease state governed by natural course of the disease irrespective of the treatment; however, each state will be accompanied with less severe symptoms, e.g. lower numbers of seizures. Application of the Mixed Hidden Markov Model allowed us to decompose the rough summary level data (daily frequencies) into several complex underlying mechanisms. The unobserved disease state dynamics may be assessed as well as specificities of each disease state (seizure frequencies, time course of drug effect, covariate effects, etc.). Using the proposed model, we can estimate what is the individual probability/risk of transitioning to the high/low activity state given the current characteristics and individual data. Similarly, individual probability of remaining in the same state can be estimated. Furthermore, this advanced approach allows introduction of covariate effects at different levels. In our example, treatment effect was significant when introduced to decrease seizure frequencies, but not transition probabilities. Other covariates did not turn

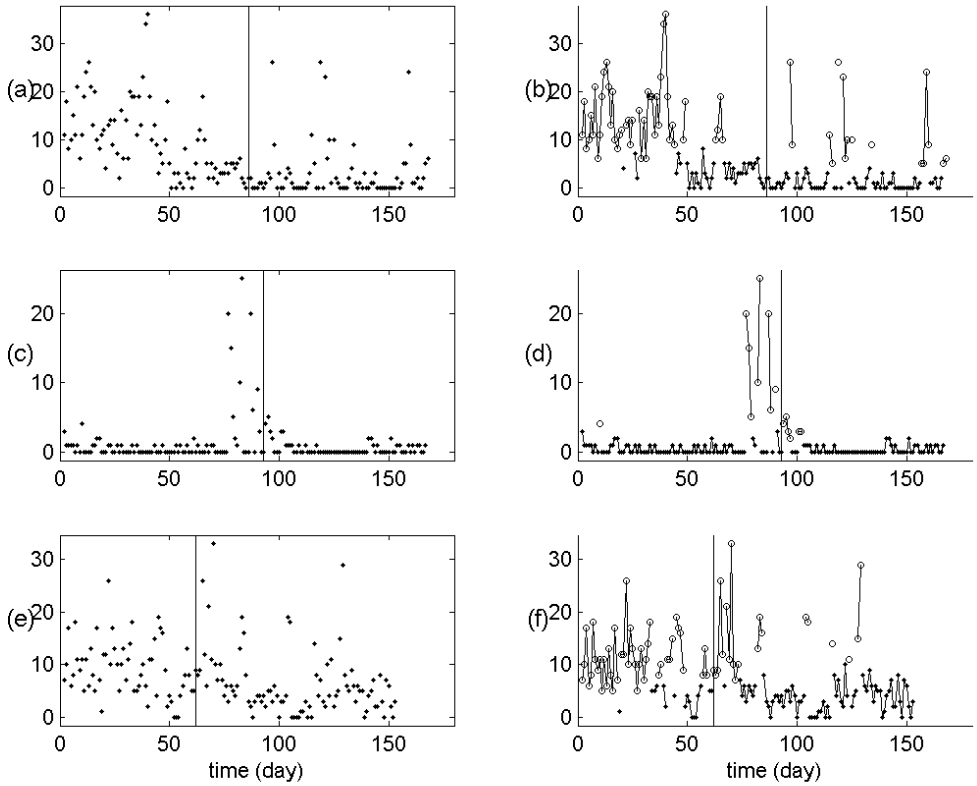

Figure 2.10: Estimation of the sequences of hidden states for three typical profiles: left panel of figures displays observed seizure counts of three typical subjects while right panel of figures shows the estimated sequences of hidden states for these three selected subjects obtained with the Viterbi algorithm. On each figure, x-axis represents time in days and y-axis provides the number of seizures. Vertical line marks the start of the treatment phase. Different symbols indicate the predicted seizure activity states on right panel: ∘ state of high epileptic activity; ▪ state of low epileptic activity.

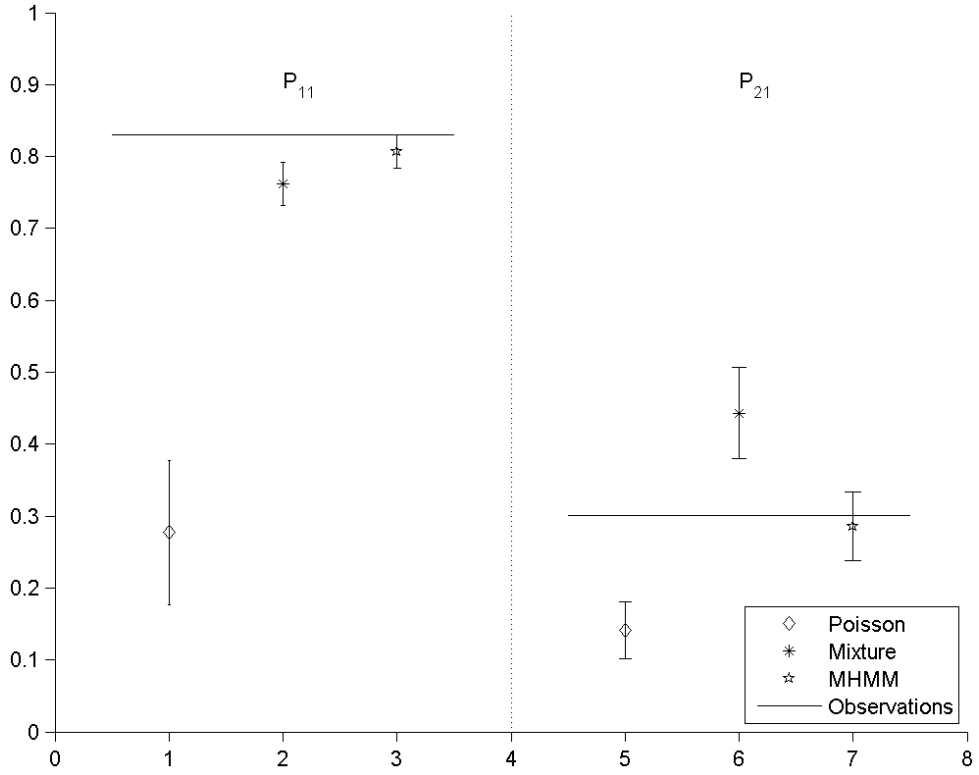

Figure 2.11: Posterior predictive check for daily transitions between different seizure counts in baseline period. The black horizontal line depicts some empirical values for p_{11} and p_{21} estimated by attributing each individual observation to either state of low epileptic activity or state of high seizure activity according to whether the observed count is lower or upper than a threshold value. To account for the between subjects variability, the thresholds value varies from an individual to another. 95% confidence interval of the empirical transition probabilities given by the same procedure from data simulated under a mixed-effects Poisson model ($\diamond$), a mixed-effects two-state Poisson mixture ($*$) and a mixed-effects two-state Poisson hidden Markov model ($\star$) are displayed. The parameters used for the simulations in each models were estimated from the screening observations of the dataset. The threshold value used for the definition of the two states was arbitrarily defined independently on the model used for simulations.

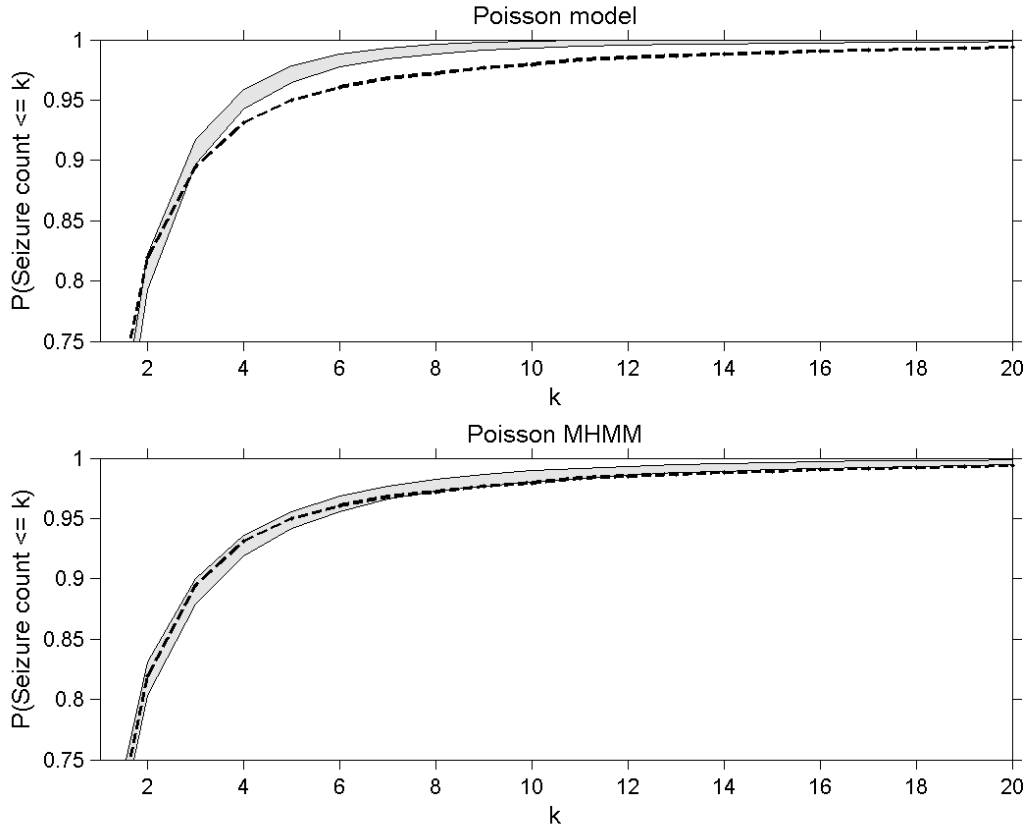

Figure 2.12: Cumulative distribution function of the daily seizure counts: the marginal distribution of the daily seizure counts (dotted line) is compared with the theoretical distributions assuming a single Poisson model (upper figure) and a HMM with two Poisson distributions (lower figure). 95% confidence intervals of the cumulative distribution function under both models are derived from simulated datasets. The parameters used for the simulations in each models were estimated from the screening observations of the dataset. In order to make the visual comparison easier, 10 times the logarithm of the cumulative distribution function is displayed.

out significant in our analysis, however only limited number of basic demographic covariates were available in our exercise. Nevertheless, MHMM allows us to investigate in more details specific covariate effects, and therefore to better understand the underlying processes and mechanisms/mode of drug actions. For example, in our analysis, we would particularly be interested into finding which predictor would increase probability of remaining in low disease activity state even further and which predictor would increase probability of transitioning from the high to the low state using data driven approach.

The additional key value of Mixed Hidden Markov Model is in improved simulation properties of the model. Mixed Hidden Markov Model allows us to simulate realistic patient profiles, not only for seizure frequencies but also for disease status. This is very important when it comes to any instance where a model is to be used for simulation purposes, which is the case when novel scenarios with respect to dose and schedule are to be evaluated. Unless Markov elements are included, number of transitions cannot be realistically generated which was also demonstrated both in work by Trocoñiz *et al.* [28] and in our exercise. From the data analysis point of view, Mixed Hidden Markov Model offers two advantages: (i) data autocorrelation issues can be elegantly handled, which is essential point in order to generate a model with good simulation properties and (ii) high individual variances are well described in a mechanistic manner. These two issues posed a significant challenge in the area of mixed effect modelling of count data. And finally, implementation of the Mixed Hidden Markov Model in Monolix software and coupling it with the SAEM algorithm, enabled the computational burden to be minimal (the full estimation for this large dataset and advanced model was less than 2 hours), which is yet another example of usefulness of SAEM algorithms to handle challenging problems.

In conclusion, the first application of the Mixed Hidden Markov model within the area of clinical pharmacology has been conducted. The proposed methodology offers a new insight in the analysis of epilepsy data where novel knowledge can be generated: gabapentin effect decreases seizure frequencies in both low and high activity state, but has no effect on transition probabilities between states. Additionally, with the novel methodology, we have addressed several important limitations that have been current in the analysis of the discrete repeated count data. Finally, this novel methodology is implemented into the Monolix software.

2.3.5 Tables

Table 2.6: Summary of the 6 multicenter clinical trials

Trial	Sex Nb of subjects	Dose Nb of subjects	Age mean (std)	Weight mean (std)	Height mean (std)
945-005	Male: 200 Female: 101 Overall: 301	Placebo : 98 600mg/day: 53 1200mg/day: 98 1800mg/day:52	34.87 (10.20)	76.82 (17.57)	171.92 (10.87)
945-006	Male: 150 Female: 118 Overall: 268	Placebo : 107 900mg/day:110 200mg/day:51	32.24 (11.92)	68.79 (13.97)	168.46 (9.01)
945-009	Male: 20 Female: 23 Overall: 43	Placebo : 18 900mg/day:16 1200mg/day:9	37.53 (10.85)	68.93 (12.81)	168.02 (10.14)
945-010	Male: 20 Female: 22 Overall: 42	Placebo : 16 900mg/day:18 1200mg/day:8	37.40 (12.20)	68.86 (14.19)	171.45 (9.09)
945-011	Male: 6 Female: 3 Overall: 9	Placebo : 4 450mg/day:2 600mg/day:3	10.56 (4.03)	33.08 (13.10)	137.13 (18.34)
945-210	Male: 53 Female: 72 Overall: 125	Placebo : 65 1200mg/day:60	30.82 (12.03)	67.10 (16.86)	165.92 (9.34)

Table 2.7: Final Parameter estimates

	Fixed effect		Variance term	
Model parameter	Estimate	RSE (%)	Estimate	RSE (%)
β_1^B	1.96	3	0.72	10
β_2^B	1.56	5	0.78	11
$\log(\lambda_1^B)$	-2.01	3	1.95	7
$\log(\alpha^B)$	-0.08	65	1.5	7
δ_1	0.17	20	0.09	29
δ_2	0.27	16	0.07	26
δ_3	-0.05	78	0.41	9
δ_4	0.07	57	0.37	12
γ_3	-1.6. 10^{-4}	18	-	-
γ_4	-2.0. 10^{-4}	11	-	-
τ_1 (day^{-1})	1.70	10	-	-
τ_2 (day^{-1})	15.65	13	-	-

2.4 Résultat complémentaire

Dans l'application que nous avons proposée des modèles de Markov cachés à effets mixtes à l'évolution du nombre de crises d'épilepsie, nous avons choisi de définir le nombre de crises se produisant dans chaque état de la maladie comme une variable aléatoire poissonnienne. Nous avons évalué ce modèle par des simulations et avons mis en évidence de bonnes propriétés d'adéquation du modèle aux données. En particulier, les données réelles et les données simulées sous le MHMM à distributions poissonniennes présentent une répartition très proche.

Toutefois, le modèle de Markov caché à effets mixtes et émissions poissonniennes ne tient pas compte de la sous-dispersion des observations de certains sujets (Figure 2.13).

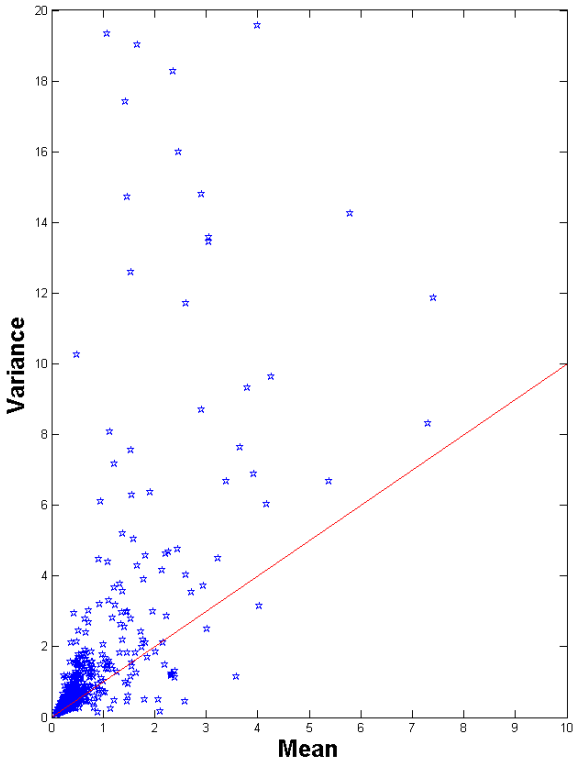

FIGURE 2.13 – Dispersion des données individuelles sur la période de screening.

La loi de Poisson généralisée (GPD, *Generalized Poisson Distribution*, [5]) étend la loi de Poisson à des données non nécessairement équidispersées, et se définit au moyen de deux paramètres : λ et θ. Selon le signe du paramètre θ, les données sont alors sur-dispersées ($\theta > 0$) ou sous-dispersées ($\theta < 0$), le cas particulier $\theta = 0$ correspondant à la loi de Poisson de paramètre λ.

Afin de mieux décrire les écarts de dispersion dans les données des patients épileptiques, nous proposons donc un modèle de Markov caché à effets mixtes à deux états, dont la loi d'émission dans chaque état est une distribution de Poisson généralisée, et dont le paramètre de dispersion est identique dans les deux états, et aléatoire. Nous avons ensuite comparé les résultats obtenus en ajustant le MHMM à émissions poissonniennes et le MHMM à émissions

GPD aux données de screening. Le modèle à émissions GPD s'écrit alors :

$$\begin{aligned}
\text{logit}(p_{11i}) &= \beta_1 + \eta_{1i}, \\
\text{logit}(p_{21i}) &= \beta_2 + \eta_{2i}, \\
\log \lambda_{1i} &= \log \lambda_1 + \eta_{3i}, \\
\log \alpha_i &= \log \alpha + \eta_{4i}, \\
\lambda_{2i} &= \lambda_{1i} + \alpha_i, \\
\theta_i &= 2/(1 + \exp(-\beta_3 - \eta_{5i})) - 1,
\end{aligned}$$

où les $\eta_i = (\eta_{1i}, \ldots, \eta_{5i})'$ sont supposés Gaussiens centrés et de matrice de covariance Ω diagonale. Les résultats, présentés dans les tableaux 2.8 et 2.9, mettent en évidence une diminution notable du BIC lorsque le nombre de crises d'épilepsie est supposé de loi GPD dans chacun des deux états de la maladie. Le développement de MHMM à émissions poissonniennes généralisées pourrait donc améliorer l'analyse de l'essai clinique portant sur l'anti-épileptique CI-945.

Parameter	Estimate	s.e.	r.s.e. (%)
β_1	1.410	0.053	4
β_2	0.148	0.011	7
β_3	-0.195	0.036	19
λ_1	0.062	0.009	14
α	0.907	0.035	4
ω_{β_1}	0.897	0.043	5
ω_{β_2}	0.053	0.006	10
ω_{β_3}	0.916	0.030	3
ω_{λ_1}	2.570	0.110	4
ω_α	0.815	0.029	4
BIC = 139522			

TABLE 2.8 – Estimation des paramètres de population d'un MHMM à deux états et émissions GPD sur les données de crises d'épilepsie. Dans ce tableau figurent les estimateurs des paramètres, ainsi que leurs s.e. relatifs et absolus, obtenus par l'algorithme SAEM, et la valeur du BIC.

Parameter	Estimate	s.e.	r.s.e. (%)
β_1	1.860	0.057	3
β_2	0.135	0.009	6
λ_1	0.168	0.011	7
α	0.760	0.043	6
ω_{β_1}	0.834	0.052	6
ω_{β_2}	0.047	0.005	11
ω_{λ_1}	1.600	0.053	3
ω_α	1.290	0.047	4
BIC = 142936			

TABLE 2.9 – Estimation des paramètres de population d'un MHMM à deux états et émissions poissonniennes sur les données de crises d'épilepsie sur la période de screening. Dans ce tableau figurent les estimateurs des paramètres, ainsi que leurs s.e. relatifs et absolus, obtenus par l'algorithme SAEM, et la valeur du BIC.

2.5 Bibliographie

[1] P. S. Albert. A two state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 37(4) :1371–1381, 1991.

[2] P.S. Albert, H.F. McFarland, M.E. Smith, and J.A. Frank. Time series for modelling counts from relapsing remitting disease : application to modelling disease activity in multiple sclerosis. *Statistics in Medicine*, 13(5-7) :453–466, 1994.

[3] R.M. Altman. Mixed hidden markov models : an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102 : 201–210, 2007.

[4] R.M. Altman and A.J. Petkau. Application of hidden markov models to multiple sclerosis lesion count. *Statistics in Medicine*, 24(15) :2335–2344, 2005.

[5] R.S. Ambagapistiya and N. Balakrishnan. On the compound generalized poisson distributions. *ASTIN Bulletin*, 24(2) :255–263, 1994.

[6] V.V. Anisimov, H.J. Maas, M. Danhof, and O. Della Pasqua. Analysis of responses in migraine modelling using hidden markov models. *Statistics in Medicine*, 26(22) :4163–4178, 2007.

[7] L.E. Baum and J.A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3) :360–363, 1967.

[8] L.E. Baum and T.P. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37(6) :1554–1563, 1966.

[9] L.E. Baum, T.P. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41(1) :164–171, 1970.

[10] O. Cappé, T. Rydén, and E. Moulines. *Inference in hidden Markov models*. Springer, 2005.

[11] F. Chaubert-Pereira, Y. Guédon, C. Lavergne, and C. Trottier. Markov and semi-markov switching linear mixed models used to identify forest tree growth components. *Biometrics*, 66(3) :753–762, 2010.

[12] M. Delattre. Inference in mixed hidden markov models and applications to medical studies. *Journal de la Société Française de Statistique*, 151(1) :90–105, 2010.

[13] M. Delattre and M. Lavielle. Maximum likelihood estimation in discrete mixed hidden markov models using the saem algorithm. *Computational Statistics & Data Analysis*, 56 (6) :2073–2085, 2012.

[14] M. Delattre, R.M. Savic, R. Miller, M.O. Karlsson, and M. Lavielle. Analysis of exposure-response og ci-945 in patients with epilepsy : application of novel mixed hidden markov modelling methodology. *Journal of Pharmacokinetics and Pharmacodynamics*, 39(1), 2012. doi : 10.1007/s10928-012-9248-2.

[15] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1) :94–128, 1999.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society - Series B*, 39 :1–38, 1977.

[17] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequences analysis*. Cambridge University Press, 1998.

[18] D. Haussler, A. Krogh, and I. Mian. A hidden markov model that finds genes in ecoli dna. *Nucleic Acids Res.*, 22(22) :4768–4778, 1994.

[19] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

[20] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49(4) :1020–1038, 2005.

[21] A. Maruotti and T. Rydén. A semiparametric approach to hidden markov models under longitudinal observations. *Statistics and Computing*, 19(4) :381–393, 2009.

[22] R. Miller, B. Frame, B. Corrigan, P. Burger, H. Bockbrader, E. Garofalo, and R. Lalonde. Exposure-response analysis of pregabalin add-on treatment of patients with refractory partial seizures. *Clin. Pharmacol. Ther.*, 73(6) :491–505, 2003.

[23] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 :257–286, 1989.

[24] F. Rijmen, E. H. Ip, S. Rapp, and E. G. Shaw. Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumours. *Journal of the Royal Statistical Society - Series A.*, 171(3) :739–753, 2008.

[25] R. Savic and M. Lavielle. A new SAEM algorithm : Performance in population models for count data. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(4) :367–379, 2009.

[26] E. Snoeck and A. Stockis. Dose-response population analysis of levetiracetam add-on treatment in refractory epileptic patients with partial onset seizures. *Epilepsy Res.*, 73 (3) :284–291, 2007.

[27] The MONOLIX software. Monolix user guide, release 3.1R2. *http ://software.monolix.org/sdoms/software/*, 2010.

[28] I.F. Trokoñiz, E. Plan, R. Miller, and M.O. Karlsson. Modelling overdispersion and markovian features in count data. *J. Pharmakinet. Pharmacodyn.*, 36 :461–477, 2009.

Chapitre 3

Modèles de diffusion à effets mixtes

Contents

3.1 Introduction

Les modèles à effets mixtes sont très largement utilisés en pharmacométrie, pour décrire par exemple la concentration d'un médicament au cours du temps dans l'organisme. Certains modèles à dynamique markovienne sont pertinents en pharmacocinétique. En particulier, les modèles définis à partir d'équations différentielles stochastiques permettent une meilleure représentation de la variabilité intra-sujet que les modèles déterministes classiques. Nous nous intéressons à l'estimation par maximum de vraisemblance dans les modèles de diffusion à effets mixtes. Les deux contributions présentées dans ce chapitre sont indépendantes.

3.1.1 Modèles de diffusion en pharmacocinétique de population et méthode d'estimation

3.1.1.1 Construction du modèle

Les modèles développés en pharmacocinétique (*Pharmacokinetics, PK*) permettent d'étudier l'évolution d'un ou plusieurs médicaments dans l'organisme humain. Ces modèles reposent sur une représentation schématique du corps humain, comme un ensemble de compartiments communiquant entre eux et à travers lesquels circule(nt) la (les) substance(s) médicamenteuse(s). Les transferts entre compartiments sont communément intégrés au modèle sous la forme d'un système d'équations différentielles ordinaires. Le lecteur pourra se référer à l'ouvrage de Gabrielsson and Weiner [10] pour davantage de détails sur cette approche. Prenons l'exemple simple du bolus : une *Dose* de médicament est administrée par injection rapide dans un unique compartiment de volume V et éliminée avec une constante d'élimination k (Figure 3.7).

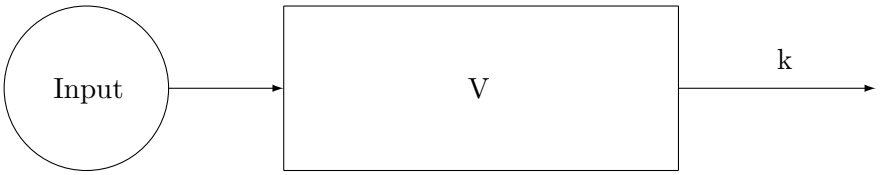

FIGURE 3.1 – Modèle Bolus.

La concentration C du médicament dans le sang est alors décrite par l'équation différentielle :

$$\frac{dC(t)}{dt} = -kC(t),$$

de solution

$$C(t) = \frac{Dose}{V} e^{-kt}.$$

La concentration du médicament modélisée par cette fonction est représentée en Figure 3.2. Des modèles plus complexes, reposant sur plusieurs équations différentielles ordinaires sont également utilisés en PK [10], notamment lorsque l'expérience est modélisée au moyen de

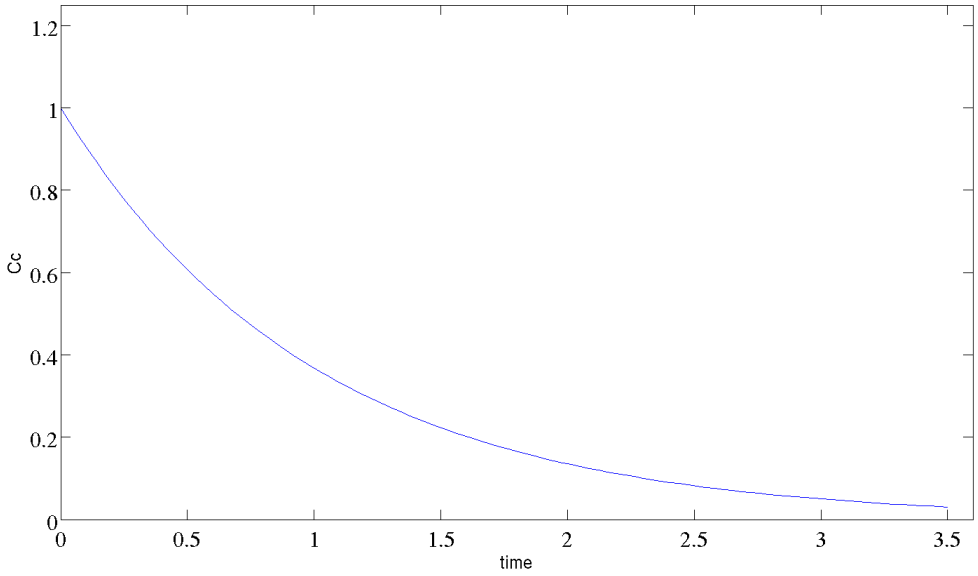

FIGURE 3.2 – Courbe de concentration d'un bolus modélisée par une équation différentielle ordinaire.

plusieurs compartiments. Dans la réalité, la cinétique des médicaments n'est pas complètement régulière et peut connaître de petites perturbations. Ces fluctuations ne sont pas prises en compte par les modèles compartimentaux déterministes, et nécessitent l'ajout d'une variabilité supplémentaire dans le système différentiel. Certains auteurs ont montré les limites de la modélisation PK par des systèmes différentiels ordinaires [22, 27]. Ils proposent de remplacer le système d'équations différentielles ordinaires par un système d'équations différentielles stochastiques pour décrire l'évolution du médicament dans les différents compartiments humains. En général, un terme de volatilité est ajouté aux équations traditionnelles des modèles compartimentaux, et les concentrations sont définies comme des processus stochastiques. Reprenons l'exemple du bolus, Overgaard et al. [22] proposent par exemple de définir la concentration du médicament comme une réalisation d'un processus d'Ornstein-Uhlenbeck :

$$dC(t) = -kC(t)dt + \gamma dW(t), \tag{3.1}$$

et [7] comme une réalisation d'un mouvement Brownien géométrique

$$dC(t) = -kC(t)dt + \gamma C(t)dW(t),$$

où W est un processus de Wiener. Les dynamiques de la concentration modélisée par ces deux processus respectifs sont représentées en Figure 3.3.

La description que ces nouveaux modèles proposent des systèmes biologiques n'est pas réaliste. En effet, les réalisations des concentrations médicamenteuses modélisées de cette façon sont très irrégulières et non monotones dans le temps. De plus, selon le choix des fonctions de volatilité, les solutions de ces systèmes peuvent prendre des valeurs négatives ; c'est le cas par exemple du modèle (3.1) proposé dans [22].

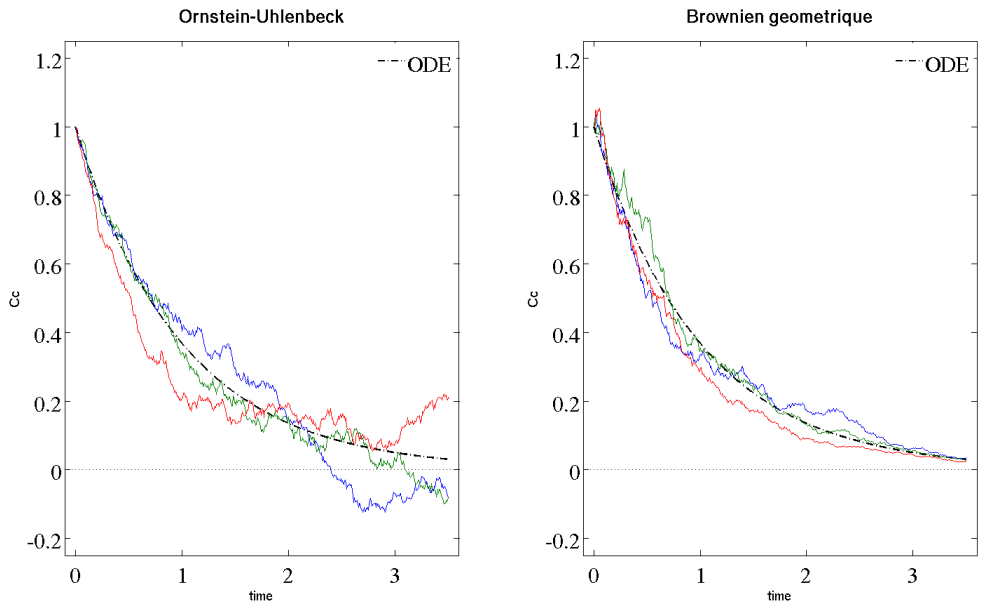

FIGURE 3.3 – Trois réalisations des courbes de concentration d'un bolus modélisée par un processus d'Ornstein-Uhlenbeck (gauche) et par un mouvement Brownien géométrique (droite).

Nous proposons une représentation plus raisonnable de la cinétique des médicaments, en supposant que les perturbations observées sur le système proviennent de petites fluctuations aléatoires des constantes physiologiques (ou taux de transfert). Ces perturbations aléatoires peuvent être intégrées au modèle en définissant les taux de transfert comme des processus discrets autocorrélés (Figure 3.4) ou comme des processus continus dont les valeurs oscillent autour d'une valeur seuil (Figure 3.5).

Dans le premier cas, les taux de transferts changent de valeur un nombre fini de fois, en des temps $t_1 < \ldots < t_j < t_{j+1} < \ldots < t_m$ inconnus, et sont constants sur chaque intervalle de temps $]t_j, t_{j+1}]$. Les sauts de valeurs sont supposés aléatoires. Disposant d'observations bruitées du système dynamique, il s'agirait alors d'estimer les paramètres du modèle dynamique ainsi que les temps de rupture t_j. Écrivons le modèle bolus avec un taux d'élimination autorégressif d'ordre 1 :

$$\begin{aligned} k(]t_j, t_{j+1}]) &= k_{j+1}, \\ dC(t) &= -k_{j+1}C(t)dt \text{ sur }]t_j, t_{j+1}], \\ k_{j+1} &= \rho k_j + b\nu_{j+1}, \nu_{j+1} \sim \mathcal{N}(0,1). \end{aligned}$$

La concentration d'un bolus modélisé de cette manière est représentée en Figure 3.4.

Il est néanmoins plus réaliste de supposer que les taux de transfert entre compartiments évoluent continûment dans le temps. Nous les définissons alors au moyen d'équations différentielles stochastiques, pour lesquelles une fonction de drift raisonnable est de la forme $\alpha(k^* - k_t)$ où α représente la force de rappel du processus à une valeur seuil k^*. Un modèle stochastique logique pour le bolus pourrait finalement s'obtenir en définissant le taux de

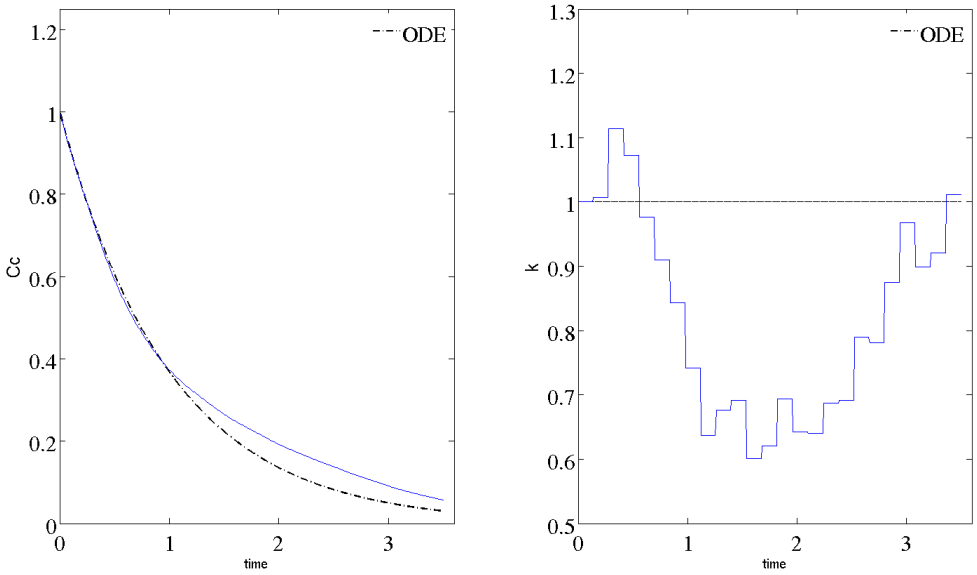

FIGURE 3.4 – Allure de la courbe de concentration d'un bolus modélisée avec un taux d'élimination constant par morceaux.

transfert comme un processus de Cox-Ingersoll-Ross

$$\begin{aligned} dk(t) &= \alpha(k^* - k(t))dt + \gamma\sqrt{k(t)}dW(t), \\ dC(t) &= -k(t)C(t)dt, \end{aligned}$$

ou son logarithme comme un processus d'Ornstein-Uhlenbeck :

$$\begin{aligned} d\log k(t) &= \alpha(\log k^* - \log k(t))dt + \gamma dW(t), \\ dC(t) &= -k(t)C(t)dt. \end{aligned}$$

Une réalisation de la courbe de concentration décrite par ce modèle est représentée en Figure 3.5.

Remarque 6. Nous pouvons voir le modèle continu comme une extension du modèle discret lorsque les variations des taux de transferts sont très rapprochées dans le temps ($t_{j+1} - t_j \to 0$). Prenons l'exemple simple d'un processus d'Ornstein-Uhlenbeck :

$$dX(t) = aX(t)dt + \gamma dW(t),\ a < 0,$$

dont la solution est explicite. Supposons les temps t_j régulièrement espacés et notons δ le pas de temps $t_{j+1} - t_j$ pour tout $j = 1, \ldots, m$. Pour toute valeur de δ et de t, $X(t+\delta)$ et $X(t)$ sont liés par la relation

$$X(t+\delta) = e^{a\delta}X(t) + b(t+\delta),$$

où $b(t+\delta)$ est une variable aléatoire définie par

$$b(t+\delta) = \sigma \int_t^{t+\delta} e^{2a(t+\delta-u)}dW(u),$$

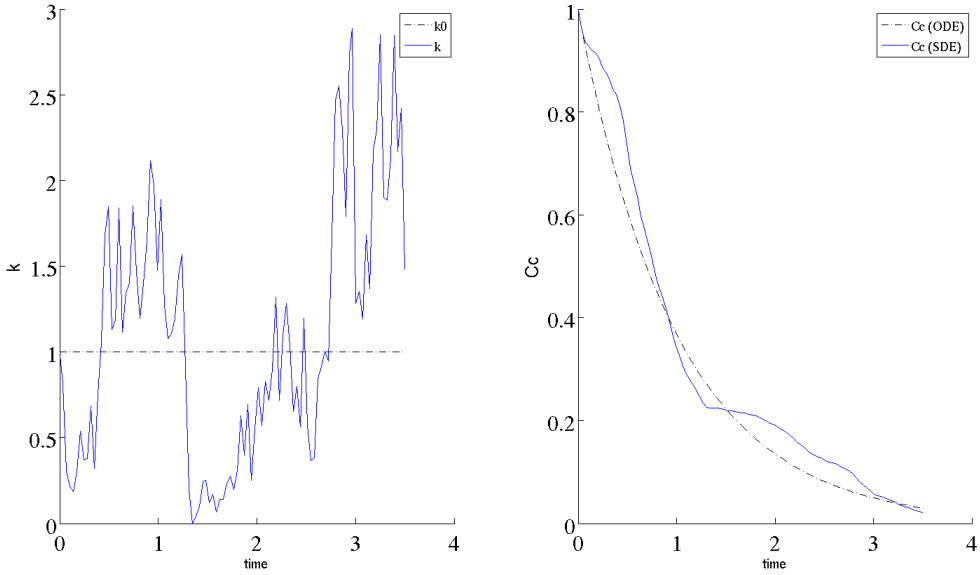

FIGURE 3.5 – Une réalisation de la courbe de concentration d'un bolus modélisée avec une constante d'élimination définie comme un processus de Cox-Ingersoll-Ross.

gaussienne, d'espérance nulle et de variance $\frac{\sigma^2}{2a}(e^{2a\delta}-1)$. Dans ce cas, les variables aléatoires $X(t_j)$, $j=1,\ldots,m$, forment un processus autorégressif d'ordre 1

$$\begin{aligned} X(t_{j+1}) &= \rho_\delta X(t_j) + b_\delta \nu_{j+1},\ \rho_\delta = e^{a\delta},\ b_\delta = \sigma\sqrt{\frac{e^{2a\delta}-1}{2a}}, \\ \nu_{j+1} &\sim \mathcal{N}(0,1). \end{aligned}$$

3.1.1.2 Contribution méthodologique

Dans une première contribution, nous nous intéressons à des modèles de diffusion à effets mixtes, à observations discrètes et bruitées :

$$\begin{aligned} dX_i(t) &= b(X_i(t),\phi_i)dt + \gamma(X_i(t),\phi_i)dW_i(t), \\ y_{ij} &= g(X_i(t_{ij}),\phi_i) + \xi_{ij}, \\ \xi_{ij} &\underset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2(\phi_i)), \\ \phi_i &\underset{i.i.d.}{\sim} \pi(\cdot, C_i, \theta), \end{aligned} \tag{3.2}$$

où les W_i sont des processus de Wiener indépendants et les fonctions de drift b et γ sont connues. Nous supposons que les conditions assurant l'existence d'une solution de ces systèmes sont vérifiées. Le principal objectif de ce travail est de proposer une version spécifique de l'algorithme SAEM pour l'estimation du paramètre θ de ces modèles à partir des observations bruitées y_{ij}.

L'estimation par maximum de vraisemblance des paramètres de ces systèmes différentiels à $\phi_i = \varphi$ fixé n'est pas standard. Du fait de la structure markovienne sous-jacente, la vraisemblance a une forme complexe puisque les observations ne sont pas indépendantes. En général, les systèmes différentiels stochastiques n'ont pas de solution explicite. Dans ce cas, l'estimation des paramètre du modèle repose sur une approximation de la distribution du processus. Le filtre de Kalman étendu propose d'approcher la distribution du processus par une distribution gaussienne en linéarisant le modèle, et donc d'estimer les paramètres du modèle à partir de la distribution ainsi approchée. Une fois les paramètres estimés, il est aussi possible a posteriori de reconstruire les trajectoires des processus par une procédure de lissage (Kalman smoother).

Plusieurs auteurs se sont déjà intéressés à l'estimation des paramètres de population dans les modèles (3.2). Overgaard et al. [22] ont proposé de combiner un algorithme FOCE (*First Order Conditional Estimate*) à un filtre de Kalman étendu pour approcher les processus de diffusion. Donnet and Samson [8] ont adapté l'algorithme d'estimation par maximum de vraisemblance SAEM en approchant les vraisemblances des données complètes par une méthode d'Euler-Maruyama. Donnet and Samson [9] ont couplé un algorithme SAEM à des méthodes de filtrage particulaire. L'étape de simulation de ces deux versions de SAEM est complexe, puisqu'elle demande de simuler à la fois les paramètres individuels ϕ_i et les trajectoires des processus sous-jacents X_i sous leurs lois conditionnelles sachant les observations. Nous proposons une nouvelle version de l'algorithme SAEM pour ces modèles couplé au filtre de Kalman étendu par approcher les vraisemblances des données complètes. Dans cette variante de l'algorithme SAEM, les trajectoires des processus de diffusion non observées sont traitées comme des paramètres de nuisance et non comme des données cachées, ce qui simplifie l'étape de simulation de l'algorithme. Ce travail est actuellement en préparation pour soumission.

3.1.2 Propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans des modèles de diffusion à effets mixtes

Nous nous sommes ensuite intéressés aux propriétés théoriques du MLE dans les modèles de diffusion à effets mixtes. Étant donné la complexité de ces modèles, nous nous sommes contentés de considérer des modèles à observations non bruitées.

L'étude théorique des estimateurs des paramètres dans les modèles stochastiques à effets fixes est déjà un problème difficile, mais la question a déjà été largement investiguée dans la littérature [15, 16]. L'estimateur du maximum de vraisemblance des paramètres a rarement une expression explicite, et ses propriétés ne sont pas toujours standard.

En revanche, les résultats asymptotiques sur le maximum de vraisemblance dans les modèles de diffusion à effets mixtes sont à ce jour peu nombreux dans la littérature. À notre connaissance, les seuls résultats concernent l'estimateur du maximum de vraisemblance des paramètres d'un mouvement Brownien, dont la fonction de drift dépend d'un effet aléatoire unidimensionnel, à partir d'observations discrètes et non bruitées [7]. Dans ce cas particulier, l'estimateur du maximum de vraisemblance est explicite et il est possible d'en étudier les propriétés asymptotiques à partir de son expression lorsque le nombre de sujets tend vers l'infini. Nous généralisons ce résultat à des processus de diffusion réels dont le drift est une fonction linéaire d'effets aléatoires gaussiens. Le modèle considéré est de la forme

$$d\,X_i(t) = \phi_i' \, b(X_i(t))dt + \sigma(X_i(t))\, dW_i(t), \quad X_i(0) = x,$$

où les W_i sont des processus de Wiener indépendants, les $\phi_i = (\phi_i^1, \ldots, \phi_i^d)'$ sont indépendants et identiquement distribués selon une loi gaussienne de paramètres θ, $b(x) = (b^1(x), \ldots, b^d(x))'$, et la fonction de volatilité $\sigma(x)$ est connue. Nous supposons que les N processus individuels sont observés continûment sur un intervalle de temps $[0, T]$. Dans cette situation, plusieurs cadres asymptotiques peuvent être considérés selon que N et/ou T tend(ent) vers l'infini. Ici, nous travaillons à T fixé et supposons que le nombre de sujets tend vers l'infini. Nous montrons que l'estimateur du maximum de vraisemblance pour θ est fortement consistant et asymptotiquement gaussien. Nous adaptons ensuite les estimateurs au cadre plus réaliste d'observations discrètes. Ce travail, réalisé en collaboration avec Adeline Samson et Valentine Genon-Catalot (MAP5, Université Paris-Descartes) est détaillé dans la deuxième partie de ce chapitre sous forme d'un article soumis à *Scandinavian Journal of Statistics* et disponible sur le HAL [4].

3.2 Premier article : Maximum likelihood estimation for mixed-effects diffusion models using the SAEM algorithm

Abstract

We consider mixed-effects diffusion models observed at discrete time points up to a Gaussian noise. These models are relevant in population pharmacokinetics. We propose new procedures for estimating the population parameters, the individual parameters and the individual latent processes in these models. In particular, we combine the SAEM algorithm with the extended Kalman filter to estimate the population parameters. The properties of this algorithm are investigated via a simulation study using models having direct applications in pharmacokinetics.

Keywords: Stochastic differential equations, Nonlinear mixed-effects models, SAEM algorithm, Extended Kalman filter, Pharmacokinetics.

3.2.1 Introduction

Mixed-effects models are standard tools for describing data collected over time in several subjects. In mixed-effects models, the same structural model is used for each individual sequence of observations, but the parameters of this model are random variables. Thus, the parameter values can change from one individual to another and the global model for the whole dataset properly reflects the variability in the data. In a population approach, variability is split into two categories: the variability between subjects which is tackled by the randomness of the individual parameters, and the intra-individual variability. In mixed-effects diffusion models, the model for each subject is based on stochastic differential equations. Diffusion models are relevant for describing random variability in dynamical systems, and have applications in many domains, such as finance, neuronal, pharmackinetics/pharmacodynamics,... The population extension of these models leads to the development of specific inference methodologies. People are mainly interested in estimating the population parameters (*ie* the parameters of the distribution of the individual parameters). Maximum likelihood estimation in general mixed-effects diffusion models is nevertheless a very complex issue. Except in very specific classes of mixed-effects diffusion models, like those considered in [7], the likelihood of the observations and the maximum likelihood estimate of the population parameters do not have any closed-form expression. Computation of the observations likelihood involves integrals over the random individual parameters; and the conditional distribution of the observations given the individual parameters is generally not explicit, as the transition densities of the underlying diffusion processes are rarely known. Several authors tackled maximum likelihood estimation of parameters in mixed-effects diffusion models, either in the case of noise-free observations of the diffusions or in the case of noisy observations of the diffusions. One of the main approach developed for that purpose consists in approximating the likelihood of the observations, then maximizing the approximated likelihood with respect to the parameters. For example, it has been suggested to compute the likelihood of the observations by combining the First-Order Conditional Estimation (FOCE) method with the extended Kalman filter [22, 27, 18]. Other approximations of the likelihood in mixed-

effects diffusion models can be found in [24] and [23]. An alternative to methods based on an approximation of the observations density is given by EM-type algorithms which iteratively perform maximum likelihood estimation based on the complete log-likelihood rather than the marginal log-likelihood. In particular, specific versions of the SAEM algorithm have been proposed for estimating parameters in mixed-effects diffusion models. In [8] the SAEM algorithm is combined with an Euler-Maruyama approximation of the individual processes and in [9], it is coupled with some particle Markov Chain Monte-Carlo methods. In these two versions of SAEM, simulation of both the random individual parameters and the individual latent processes is required at simulation step, which is computationally cumbersome. In the present work, we develop a new version of SAEM, coupled with the (continuous-discrete) extended Kalman filter. In this algorithm, we only need to simulate the random parameters. We also give tools for estimating the individual parameters and the individual diffusion trajectories. The organisation of the present paper is as follows. In Section 3.3.5, we introduce mixed-effects diffusion models and some notations. In Section 3.2.3, we detail inference methodolody based on the (continuous-discrete) extended Kalman filter and the SAEM algorithm. The model and the methods are illustrated and motivated in Section 3.2.4 in population pharmacokinetics (PK) through a brief simulation study. Section 3.2.5 summarizes the results and discusses the properties of the proposed methodology.

3.2.2 Model

3.2.2.1 Formulation of the model

a) Diffusion model

Models involving diffusion processes are relevant statistical models for describing random variability in dynamical systems. Assume that a trajectory of the system is observed, up to a noise, at discrete time points $t_0 < t_1 < \ldots < t_j < \ldots < t_n$. Let $X \in \mathbb{R}^d$ denote the dynamical process and y_j the observation of process X at time t_j, $j = 0, \ldots, n$. The diffusion model has following general form:

$$\begin{cases} dX(t) = b(X(t), \varphi)dt + \gamma(X(t), \varphi)dW(t), \\ y_j = g(X(t_j), \varphi) + \xi_j \quad , \quad \xi_j \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \quad , \quad j = 0, \ldots, n, \end{cases} \tag{3.3}$$

where functions $b(\cdot)$, $\gamma(\cdot)$ and $g(\cdot)$ are assumed known up to parameter φ, and $W(t)_{t\geq 0}$ stands for a standard Wiener process. Diffusion models have numerous applications, of which the description of the course of financial assets, neuronal, population growth,... Main statistical problem in model (3.3) is to estimate φ from the observations. Estimation procedures in such models are not standard as the likelihood of model (3.3) has generally no closed form. As the observations depend on a Markovian latent process, $y_0, \ldots, y_n$ are not independant. Thus, the expression of the likelihood is:

$$p(\mathbf{y}, \varphi) = p(y_0; \varphi) \prod_{j=1}^{n} p(y_j | y_0, \ldots, y_{j-1}, \varphi). \tag{3.4}$$

An exact evaluation of (3.4) is generally intractable as each conditional density involves integral over the latent process. Main estimation procedures are based on an approximation

of the likelihood, such as the extended Kalman filter (EKF). A concise description of the EKF is provided in next section.

b) Mixed-effects diffusion model

Let us now consider model (3.3) with observations resulting from several subjects. An adequate adaptation of model (3.3) in a population approach consists in considering the parameters of each dynamical system as independent random variables, in a way to correctly reflect the variability occuring between the different trajectories. Let us consider N different subjects randomly chosen from a population. Let $n_i + 1$ denote the number of observations for individual i, $i = 1, \ldots, N$, y_{ij} the observation for individual i at time t_{ij}, $j = 0, \ldots, n_i$ ($t_{i0} < t_{i1} < \ldots < t_{i,n_i}$) and $\mathbf{y}_i = (y_{i0}, \ldots, y_{i,n_i})$ the data vector of subject i. The y_{ij}'s are now governed by a mixed-effects model based on a d-dimensional real-valued stochastic differential equation (SDE) system defined as follows:

$$\begin{cases} dX_i(t) = b(X_i(t), \phi_i)dt + \gamma(X_i(t), \phi_i)dW_i(t), \\ y_{ij} = g(X_i(t_{ij}), \phi_i) + \xi_{ij} \quad , \quad \xi_{ij}, \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2(\phi_i)), \end{cases} \tag{3.5}$$

$i = 1, \ldots, N$, with initial condition $X_i(t_0) = x_{i0} \in \mathbb{R}^d$. ϕ_i is a q-dimensional random effects parameter (subject specific) called individual parameter, with distribution depending on a set of parameters θ:

$$\phi_i \sim \pi(\cdot, \theta).$$

The $W_i(t)_{t \geq 0}$ are N standard independent Wiener processes such that $(\phi_1, \ldots, \phi_N)$ and $W_1, \ldots, W_N$ are independent. The ξ_{ij}'s are independent Gaussian random variables with variance σ^2 representing the measurement errors. $g(\cdot)$, and the drift and the diffusion functions $b(\cdot)$ and $\gamma(\cdot)$ are assumed known-up to the parameters and common to the N subjects. In the following, $X_i = (X_i(t_{i0}), \ldots, X_i(t_{i,n_i}))$ will denote the vector of the i^{th} latent process realization at observation times (t_{ij}).

The likelihood in mixed-effects diffusion models has a nontrivial form. In any mixed-effects model, the marginal density of i^{th} data vector $\mathbf{y}_i$ is obtained by integrating the conditional density of the data given the non-observable random effects ϕ_i with respect to the marginal density of the individual parameters:

$$p(\mathbf{y}_i; \theta) = \int p(\mathbf{y}_i | \phi_i) \pi(\phi_i; \theta) d\phi_i. \tag{3.6}$$

Here, the observations depend on a Markovian latent process, and for all $i = 1, \ldots, N$, we have

$$p(\mathbf{y}_i | \phi_i) = p(y_{i0} | \phi_i) \prod_{j=1}^{n_i} p(y_{ij} | y_{i0}, \ldots, y_{i,j-1}, \phi_i).$$

(3.6) has generally no closed form expression, due to the integral with respect to ϕ_i and to the intricate form of $p(\mathbf{y}_i | \phi_i)$.

Then, by independence of the N individuals, the likelihood function is given by:

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_N; \theta) = \prod_{i=1}^{N} p(\mathbf{y}_i; \theta). \tag{3.7}$$

Remark 7. (3.6) gives general expression of the marginal distribution of $\mathbf{y}_i$, applicable to any distribution for $\mathbf{y}_i|\phi_i$ and any distribution $\pi(\cdot,\theta)$ for the random parameters. For example, the ϕ_i's can be assumed to be Gaussian random variables, with mean μ and covariance matrix Ω, as in the illustrative section of the present paper. Thus, $\theta = (\mu, \Omega)$. This can also be easily extended to models including covariates, such as

$$\phi_i = \beta C_i + \eta_i \quad , \quad \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega),$$

where β is a K $\times$ p fixed-effects matrix, and C_i is a p-vector of covariates for individual i, $i = 1, \ldots, N$.

3.2.3 Estimation

In the present section, we suggest some inference methodology specific to mixed-effects diffusion models, mainly based on the (continuous-discrete) EKF and the SAEM algorithm. The present methodology tackles population parameter estimation, individual parameter estimation and estimation of the individual latent process. We also give computational method for the likelihood and the Fisher information matrix.

3.2.3.1 Maximum likelihood estimation

Main issue in mixed-effects models is to assess both the variability and the average trend in the population. Thus, main inference problem is not as in (3.3) to infer the individual parameters ϕ_i from the observations of subject i but the distribution of the ϕ_i's in the population from the observations of the N subjects.

Our aim is to propose maximum likelihood estimation of θ in mixed-effects diffusion models based on a specific version of the SAEM algorithm. Recall that as the likelihood function is not explicit, maximizing (3.7) with respect to θ is intractable. The estimation is complex because the N random parameters ϕ_i and the N latent processes X_i are not directly observed. In a general manner, linear and nonlinear mixed-effects models, including mixed-effects models based on stochastic differential equations, could be seen as incomplete data models in which the random effects $\phi = (\phi_1, \ldots, \phi_N)$ are the non-observed data and the population parameters are the parameters of the model that need to be estimated from the N individual observations vectors $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. The EM algorithm [6] iteratively performs parameter estimation in such models. The algorithm requires to compute at each iteration the conditional expectation $E\left(\log p(\mathbf{y}, \phi; \theta)|\mathbf{y}, \theta^{(k-1)}\right)$, where $\theta^{(k-1)}$ represents the current estimation of θ. In many situations, especially when dealing with nonlinear mixed-effects models, this conditional expectation has no closed form. Some variants of the algorithm get around this difficulty. In the SAEM algorithm [5], the E-step is evaluated by a stochastic approximation procedure.

a) General description of the SAEM algorithm

Let $\theta^{(k-1)}$ denote the current estimate for the population parameters. Iteration k of the SAEM algorithm involves three steps [5]:

– In the simulation step, $\theta^{(k-1)}$ is used to simulate the missing data $\phi_i^{(k)}$ under the conditional distribution $p(\phi_i|\mathbf{y}_i, \theta^{(k-1)})$, $i = 1, \ldots, N$.

– In the stochastic approximation step, the simulated data $\phi^{(k)}$ and the observations $\mathbf{y}$ are used together to update the stochastic approximation $Q_k(\theta)$ of the conditional expectation $E\left(\log p(\mathbf{y}, \phi; \theta)|\mathbf{y}, \theta^{(k-1)}\right)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log p(\mathbf{y}, \phi^{(k)}; \theta) - Q_{k-1}(\theta)\right], \tag{3.8}$$

where $(\gamma_k)_{k>0}$ is a sequence of positive step sizes decreasing to 0 and starting with $\gamma_1 = 1$.

– In the maximization step, an updated value of the estimate $\theta^{(k)}$ is obtained by maximization of $Q_k(\theta)$ with respect to θ:

$$\theta^{(k)} = \underset{\theta}{\operatorname{argmax}}\, Q_k(\theta).$$

This procedure is iterated until numerical convergence of the sequence $\left(\theta^{(k)}\right)_{k>0}$ to some estimate $\hat{\theta}$ is achieved. Convergence results can be found in [5].

b) Adaptation of SAEM to mixed-effects diffusion models

Although the unobserved data in these models consist of both the sequences of hidden states X_i, $i = 1, \ldots, N$ and the individual parameters ϕ_i, $i = 1, \ldots, N$, it is not required to simulate the hidden states at each iteration of the algorithm. Thus, at simulation step of SAEM, a simulation under $p(\phi_i|\mathbf{y}_i; \theta^{(k-1)})$ is required, but direct simulation of such distribution is impossible. We implement a Metropolis-Hastings algorithm to perform simulation of $\phi_i^{(k)}$, $i = 1, \ldots, N$, at simulation step of SAEM. Refer to [13] for additional details about the MCMC-SAEM algorithm and to [13, 1] for theoretical convergence results about the MCMC-SAEM algorithm. Computation of the acceptance probabilities in the Metropolis-Hastings algorithm however requires knowledge of the expression of $p(\mathbf{y}_i, \phi_i; \theta)$, and the ability to explicitely calculate it. Expression of $p(\mathbf{y}_i, \phi_i; \theta)$ is also necessary to update function $Q_k(\theta)$ at each iteration of the algorithm. The key of the SAEM algorithm is therefore a (quick) computation of $p(\mathbf{y}_i, \phi_i; \theta)$ for any ϕ_i and any θ for all $i = 1, \ldots, N$. Recall that

$$p(\mathbf{y}_i, \phi_i; \theta) = p(\mathbf{y}_i|\phi_i)p(\phi_i; \theta).$$

Computing $p(\phi_i; \theta)$ is straightforward since the ϕ_i's are Gaussian variables, but computing $p(\mathbf{y}_i|\phi_i)$ is generally impossible in an exact form. The (continuous-discrete) EKF provides a way to numerically approximate $p(\mathbf{y}_i|\phi_i)$ with Gaussian densities. Therefore, we suggest coupling the continuous-discrete EKF with the MCMC-SAEM algorithm to estimate population parameters in stochastic differential mixed-effects models.

The continuous-discrete EKF is described in next paragraph. To ease the reading, we focus on a single individual. Therefore, we omit the index i.

i) The continuous-discrete extended Kalman filter

The continuous-discrete EKF is intended to performing state estimation from continuous time models with discrete time measurements as in (3.3) by linearization of $b(\cdot)$, $\sigma(\cdot)$ and $g(\cdot)$, and thus Gaussian approximation of the densities. It consists in a recursive method divided in two steps repeated n times ($j = 1, \ldots, n$), after initialization:

– **Prediction**
The predictive cycle approximates the probability density function $p(X(t_j)|y(t_0),\dots,y(t_{j-1}),\varphi)$, ie its mean $\hat{X}_{j|j-1}$ and its covariance matrix $\hat{P}_{j|j-1}$, by solving

$$\begin{cases} \dot{x} &= b(x(t),\varphi) \\ \dot{P} &= B(x(t),\varphi)P(t) + P(t)B(x(t),\varphi)^T + \Omega(x(t),\varphi) \end{cases}$$

on time interval $[t_{j-1},t_j]$, with initial conditions $x(t_{j-1}) = \hat{X}_{j-1|j-1}$ and $P(t_{j-1}) = \hat{P}(j|j-1)$, and setting:

$$\begin{cases} \hat{X}_{j|j-1} &= x(t_j), \\ \hat{P}_{j|j-1} &= P(t_j). \end{cases}$$

Here, $\Omega(x(t),\varphi) = \gamma(x(t),\varphi)\gamma(x(t),\varphi)^T$ and B stands for the Jacobian of $b(\cdot)$ with respect to x.

– **Updtate**
The filtering cycle approximates the conditional probability density function $p(X_{t_j}|y(t_0),\dots,y(t_j),\varphi)$, ie its mean $\hat{X}_{j|j}$ and its covariance matrix $\hat{P}_{j|j}$, by computing the Kalman gain:

$$K_j = \hat{P}_{j|j-1}G(X_{j|j-1})^T(G(X_{j|j-1})\hat{P}_{j|j-1}G(X_{j|j-1})^T + \sigma^2)^{-1},$$

and incorporating it into the expressions of $\hat{X}$ and $\hat{P}$ updates:

$$\begin{cases} \hat{X}_{j|j} &= \hat{X}_{j|j-1} + K_j(y_j - g(\hat{X}_{j|j-1})), \\ \hat{P}_{j|j} &= (I - K_jG(X_{j|j-1}))\hat{P}_{j|j-1}, \end{cases}$$

where G is the Jacobian of $g(\cdot)$ with respect to x and I denotes the identity matrix.

Once the EKF is calculated, $p(y(t_j)|y(t_0),\dots,y(t_{j-1}),\varphi)$ is approximated with a Gaussian distribution with mean $g(X_{j|j-1})$ and covariance $G(X_{j|j-1})P_{j|j-1}G(X_{j|j-1})^T + \sigma^2$ and $p(\mathbf{y})$ directly follows, where $\mathbf{y}$ denotes the data vector.

ii) Resolution of the moment ODEs

However, the moments ordinary differential equations in the prediction step generally have no closed form solution. Thus, their solution $\hat{X}_{j|j-1}$ and $\hat{P}_{j|j-1}$ are approximated. Here, we suggest using a simplified version of method described in article of Mazzoni [17] based on higher order Taylor approximations. As described in [17], we can provide an approximation of $\hat{X}_{j|j-1}$ and $\hat{P}_{j|j-1}$ by setting:

$$\hat{X}_{j|j-1} = \hat{X}_{j-1|j-1} + \left(I - B(\hat{X}_{j-1|j-1})\frac{\Delta t_j}{2}\right)^{-1} b(\hat{X}_{j-1|j-1})\Delta t_j,$$

$$\hat{P}_{j|j-1} = \hat{P}_{j-1|j-1} + M_{\tau_j}\left(B(X_{\tau_j})\hat{P}_{j-1|j-1} + \hat{P}_{j-1|j-1}B(X_{\tau_j})^T + \Omega(X_{\tau_j},\varphi)\right)M_{\tau_j}^T\Delta t_j,$$

where

$$\Delta t_j = t_j - t_{j-1}, \quad \tau_j = t_{j-1} + \frac{\Delta t_j}{2},$$

$$M_{\tau_j} = \left(I - X_{\tau_j}\frac{\Delta t_j}{2}\right)^{-1},$$

$$X_{\tau_j} = \frac{1}{2}\left(\hat{X}_{j-1|j-1} + \hat{X}_{j|j-1} - B(X_{\tau_j})b(X_{\tau_j},\varphi)\frac{\Delta t_j^2}{4}\right).$$

See [17] for more details about these approximations.

Remark 8. As this method only requires knowledge of the jacobian function of the drift function of the dynamical process, it is easily implementable in many stochastic differential models.

Remark 9. This can be easily extended to models involving multidimensionnal observations.

3.2.3.2 Estimation of the Fisher Information matrix

When an estimate $\hat{\theta}$ of θ has been obtained with the SAEM algorithm, the standard errors of its components can be derived by computing the Fisher information matrix $I(\hat{\theta}) = -\frac{\partial^2 \log(p(\mathbf{y};\theta))}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}}$. Once again, due to the complex expression of the likelihood, the Fisher information matrix is not known in a closed form. We therefore estimate $I(\hat{\theta})$ with a stochastic approximation procedure as suggested in [14]. This procedure is based on Louis's formula, and approximates $\frac{\partial^2 \log(p(\mathbf{y};\theta))}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}}$ by the sequence (H_k), defined as

$$\begin{aligned}
\Delta_k &= \Delta_{k-1} + \gamma_k \left[\frac{\partial \log(p(\mathbf{y}, \phi^{(k)}; \theta^{(k)}))}{\partial\theta} - \Delta_{k-1} \right], \\
D_k &= D_{k-1} + \gamma_k \left[\frac{\partial^2 \log(p(\mathbf{y}, \phi^{(k)}; \theta^{(k)}))}{\partial\theta\partial\theta'} - D_{k-1} \right], \\
G_k &= G_{k-1} + \gamma_k \left[\frac{\partial \log(p(\mathbf{y}, \phi^{(k)}; \theta^{(k)}))}{\partial\theta} \frac{\partial \log(p(\mathbf{y}, \phi^{(k)}; \theta^{(k)}))}{\partial\theta}' - G_{k-1} \right], \\
H_k &= D_k + G_k - \Delta_k \Delta_k',
\end{aligned}$$

where the $\phi^{(k)}$'s are simulated under $p(\cdot|\mathbf{y}, \theta^{(k-1)})$ at each iteration via a Metropolis-Hastings algorithm. As above, we mainly need to be able to compute $p(\mathbf{y}_i|\phi_i)$ for any subject $i = 1, \ldots, N$ and any individual parameter ϕ_i, for computation of the acceptance rate of the Metropolis-Hastings algorithm. This is treated using the extended Kalman filter.

3.2.3.3 Estimation of the likelihood

Standard model selection criteria such as the Bayesian Information Criteria (BIC) require computation of the observed log-likelihood $\log(p(\mathbf{y};\hat{\theta}))$. As the log-likelihood can not be computed in a closed form here, it is approximated using an Importance Sampling integration procedure as initially suggested in [14]. This consists in drawing $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(M)}$ under a given sampling distribution $\tilde{\pi}(\cdot, \theta)$, and approximating the likelihood with:

$$p(\mathbf{y};\theta) \approx \frac{1}{M} \sum_{k=1}^{M} p(\mathbf{y}|\phi^{(k)}, \theta) \frac{\pi(\phi^{(k)}, \theta)}{\tilde{\pi}(\phi^{(k)}, \theta)},$$

and $\theta = \hat{\theta}$.This procedure therefore requires $p(\mathbf{y}_i|\phi_i)$ for all ϕ_i, $i = 1, \ldots, N$, which is approximated using the extended Kalman filter.

3.2.3.4 Estimation of the individual parameters

When an estimate $\widehat{\theta}$ of the population parameters has been obtained, the distribution of the ϕ_i's is fully defined. Estimation of $\phi_1, \ldots, \phi_N$ can be performed several ways:

1. The first consists in estimating the individual parameters using the Maximum A Posteriori (MAP) approach. The estimate $\hat{\phi}_i$ of ϕ_i, $i = 1, \ldots, N$ is then given by:

$$\begin{aligned} \hat{\phi}_i &= \underset{\phi_i}{\operatorname{argmax}}\, p(\phi_i|\mathbf{y}_i, \hat{\theta}), \\ &= \underset{\phi_i}{\operatorname{argmax}}\, p(\mathbf{y}_i|\phi_i)p(\phi_i, \hat{\theta}). \end{aligned} \tag{3.9}$$

   Maximization of the right-hand term in (3.9) is intractable in mixed-effects models based on SDEs, and requires a numerical optimization procedure.

2. The second consists in estimating ϕ_i with the mean of the conditional distribution of ϕ_i given the observations $\mathbf{y}_i$ and $\hat{\theta}$:

$$\hat{\phi}_i = \mathbb{E}(\phi_i|\mathbf{y}_i, \hat{\theta}). \tag{3.10}$$

   Once again, the expression of $\mathbb{E}(\phi_i|\mathbf{y}_i, \hat{\theta})$ is not explicit in our models, and the conditional mean of ϕ_i is estimated with a Metropolis-Hastings algorithm.

3.2.3.5 Estimation of the latent process

Using the individual parameter estimate $\hat{\phi}_i$ obtained either as the conditional mode or as the conditional mean of $p(\phi_i|\mathbf{y}_i, \hat{\theta})$, we look for the most likely values for the latent process $X_i(t)_{t\geq 0}$ at measurement times $(t_{i0}, \ldots, t_{i,n_i})$. The extended Kalman filter does already provide estimation of the $X_i(t_{ij})$'s by their conditional expectations given the observations of subject i up to time t_{ij}: $\mathbb{E}(X_i(t_{ij})|y_{i0}, \ldots, y_{ij}, \hat{\phi}_i, \theta)$, which only takes into account the "past" information relative to $X_i(t_{ij})$. By incorporating the "future" observations relative to $X_i(t_{ij})$, we can obtain a more refined state estimate. Therefore, we rather estimate the $X_i(t_{ij})$'s using the MAP:

$$\hat{X}_i(t_{ij}) = \underset{X_i(t_{ij})}{\operatorname{argmax}}\, p(X_i(t_{ij})|y_{i0}, \ldots, y_{i,n_i}, \hat{\phi}_i, \theta).$$

This can be performed using some fixed-interval Kalman smoother [11].

3.2.4 Illustration

3.2.4.1 Model examples: dynamical systems for linear transfers

Let us first consider models based on ordinary differential equations (ODEs) representing linear transfers between different entities:

$$dX(t) = KX(t)dt, \tag{3.11}$$

where $X(t)$ is a vector whose d^{th} component represents condition of d^{th} entity at time t and $K = (K_{l,l'})$ is a deterministic matrix defined as:

$$\begin{cases} K_{l,l'} = k_{l,l'} & \text{if} \quad l \neq l', \\ K_{l,l} = -k_{l0} - \sum_{l'} k_{l,l'} & \text{otherwise,} \end{cases} \tag{3.12}$$

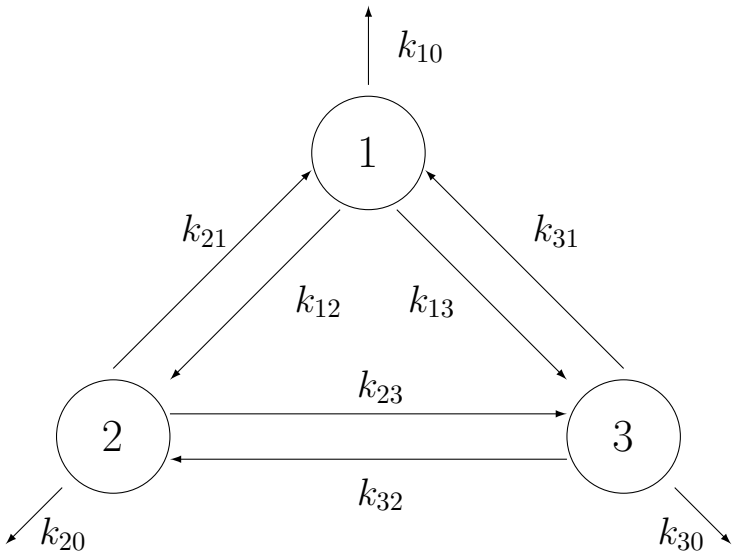

Figure 3.6: Example of dynamical system with 3 components and linear transfers between components.

with $k_{l,l'}$ representing the transfer rate from entity l to entity l', and k_{l0} the elimination rate from entity l. An example of such dynamical system with 3 components is represented in Figure **??**. In this particular model, matrix K is defined as

$$K = \begin{pmatrix} -k_{10} - k_{12} - k_{13} & k_{21} & k_{31} \\ k_{12} & -k_{20} - k_{21} - k_{23} & k_{32} \\ k_{13} & k_{23} & -k_{30} - k_{31} - k_{32} \end{pmatrix}.$$

Applications of such dynamical models are numerous: description of viral dynamics, pharmacokinetics (PK) compartmental models, description of population flows, description of interactions between cells could be a few of them.

Model defined by equations (3.11) and (3.12) is a deterministic model which assumes the transfers to take place at the same rate at all times. This is however often restrictive assumption. In the reality, dynamical systems usually experience some more or less small perturbations. Thus, it is reasonnable to consider that some or all transfer rates have volatility. This new assumption leads to the following dynamical system:

$$dX(t) = K(t)X(t)dt, \tag{3.13}$$

with K having the same structure as in (3.12) with

$$\begin{cases} dk_{l,l'}(t) = b_{l,l'}(k_{l,l'}(t))dt + \gamma_{l,l'}(k_{l,l'}(t))dW_{l,l'}(t) & \text{if} \quad k_{l,l'} \quad \text{has volatility,} \\ dk_{l,l'}(t) = 0 & \text{elsewhere.} \end{cases} \tag{3.14}$$

In situations evoked above, the transfer rates are not expected to move too far from a threshold value, and keep positive values over time. The drift and the volatility functions have to be chosen accordingly. Let us now illustrate construction of such diffusion-based models on

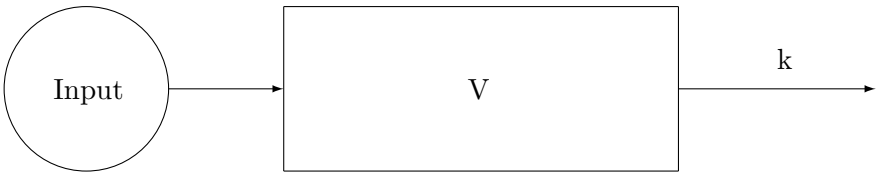

Figure 3.7: Bolus with linear elimination.

some specific examples in pharmacokinetics.

Example* 3.1. *Bolus with linear elimination
The following ordinary differential equation

$$dQ(t) = -kQ(t)dt,$$

is usually used in PK for bolus, *ie* a drug administered by rapid injection in plasma. In bolus-specific compartmental models, plasma is assimilated to a single compartment of the human body. $Q(t)$ represents the amount of the drug substance in plasma at time t after injection, and k is the elimination constant rate (Figure 3.7). Now assume that the drug's dynamics is perturbed and define k as a stochastic process. Many diffusion models ensure k to oscillate around a threshold value and to keep positive values. For example, we can define the logarithm of the transfer rate as an Ornstein-Uhlenbeck diffusion process:

$$d(\log k(t)) = -\alpha \left[\log k(t) - \log k^*\right] dt + \gamma dW_1(t),$$

where W_1 is a standard one dimensional Wiener process. This results in the following diffusion system:

$$dX(t) = b(X(t))dt + \gamma(X(t))dW(t),$$

where $X(t) = (Q(t), \log k(t))'$, $b(x) = (-x_1 \exp x_2, -\alpha(x_2 - \log k^*))'$,
$\gamma(x) = \begin{pmatrix} 0 & 0 \\ 0 & \gamma \end{pmatrix}$, and W represents a bidimensional Wiener process. Note that in this specific example, the Jacobian matrix of the drift function has a simple form:

$$B = \begin{pmatrix} -\exp x_2 & -x_1 \exp x_2 \\ 0 & -\alpha \end{pmatrix},$$

allowing easy implementation of previously described resolution method for the moment equations of the extended Kalman filter.

Example* 3.2. *Monocompartmental model with first-order absorption and linear elimination
Such system is well-known in PK modelisation where it is used to describe the fate of a drug

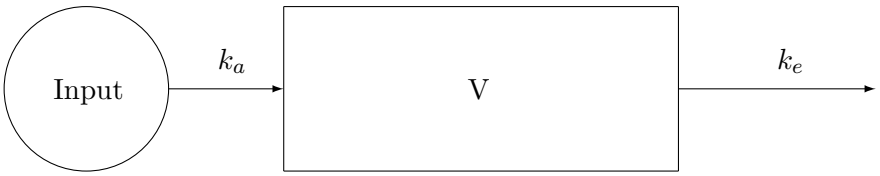

Figure 3.8: Monocompartmental model with first-order absorption and linear elimination.

oraly administered through a unique compartment of the human body. The drug is absorbed by the compartment with absorption rate k_a and removed with elimination rate k_e (Figure 3.8). Related models are based on the following system of ODEs:

$$\frac{d}{dt}\begin{pmatrix} Q_a(t) \\ Q(t) \end{pmatrix} = \begin{pmatrix} -k_a & 0 \\ k_a & -k_e \end{pmatrix}\begin{pmatrix} Q_a(t) \\ Q(t) \end{pmatrix}, \tag{3.15}$$

where $Q(t)$ represents the amount of drug at time t in the compartment. Now assume that the constant of elimination is driven by a stochastic differential equation, for example

$$dk_e(t) = -\alpha(k_e - k_e^*)dt + \gamma\sqrt{k_e(t)}dW_1(t),$$

where W_1 is a standard one dimensional Wiener process, then (3.15) becomes:

$$dX(t) = b(X(t))dt + \gamma(X(t))dW(t),$$

where $X(t) = (Q_a(t), Q(t), k_e(t))'$, $W(t)$ is a three-dimensional Wiener process, and the drift and volatility functions are defined as $b(x) = (-k_a x_1, x_1 - x_3 x_2, -\alpha(x_3 - k_e^*))'$ and $\gamma(x) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \gamma\sqrt{x_3} \end{pmatrix}$ for $x = (x_1, x_2, x_3)'$. Like in Example 3.1, the Jacobian of $b(\cdot)$ has an easy expression:

$$B(x) = \begin{pmatrix} -k_a & 0 & 0 \\ k_a & -x_3 & -x_2 \\ 0 & 0 & -\alpha \end{pmatrix}.$$

3.2.4.2 Simulation study

In the present section, we investigate the properties of the maximum likelihood estimates obtained with the SAEM algorithm combined with the extended Kalman filter through a short simulation study.

a) Design of the simulations

The model used for the simulations is a mixed-effects bolus model with an elimination constant rate defined as a stochastic process as in Example 3.1, with discrete-time observations consisting of the log-concentration of drug, up to a Gaussian noise.

For $i = 1, \ldots, N$,

$$\begin{aligned} &dQ_i(t) = -k_i(t)Q_i(t)dt, \\ &d\log k_i(t) = -\alpha_i(\log k_i(t) - \log k_i^*)dt + \gamma_i dW_i(t), \\ &y_{ij} = \log(Q_i(t_{ij})/V_i) + \xi_{ij} \quad ; \quad \xi_{ij}, \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \end{aligned}$$

with initial conditions $Q_i(0) = D_i$ and $k_i(0) = k_i^*$. D_i denotes the dose of drug injected at time $t = 0$ and V_i stands for the apparent volume of distribution of the medication. In the present simulation study, we set $\alpha_i = \alpha$ and $\gamma_i = \gamma$ and $\log k_i^* = \log k^* + \eta_{1i}$, $\log V_i = \log V + \eta_{2i}$ for all $i = 1, \ldots, N$. We consider that random variables η_{1i} and η_{2i} are independant centered Gaussian variables with variance ω_k^2 and ω_V^2 respectively. We simulate equally spaced observation times between 0 and 24 hours. For reasons of identifiability of the model, we assume that parameter α is known. We set $\alpha = 1$. The population parameter vector is $\theta = (\gamma, k^*, V, \omega_k, \omega_V)$.

100 datasets are simulated with different numbers of subjects ($N = 30$ and $N = 100$) and different numbers of observations per subject ($n = 25$ and $n = 100$). The values of the fixed-effects and the variance of the random effects used in simulations are given in tables 1 and 2. The population parameters and standard errors of the estimated parameters are estimated with SAEM for each simulated dataset, according to the procedures described in Section 3.2.3. The algorithm is implemented in Matlab and integrated in a working version of MONOLIX. The algorithm is initialized with values randomly chosen in a neighborhood of the true parameter values. For $m = 1, 2, \ldots, 100$, let $\hat{\theta}_m$ be the estimated vector of population parameters obtained with the m^{th} simulated dataset and let $\widehat{se}_m$ be their respective estimated standard errors. For each design, we have computed the mean estimated parameter:

$$\bar{\theta} = \frac{1}{100}\sum_{m=1}^{100} \hat{\theta}_m,$$

the parameter standard deviation

$$sd = \frac{1}{100}\sum_{m=1}^{100} (\hat{\theta}_m - \bar{\theta})^2,$$

and the mean estimated standard error:

$$\bar{se} = \frac{1}{100}\sum_{m=1}^{100} \widehat{se}_m.$$

The distribution of the relative estimation errors $100 \times \dfrac{\hat{\theta}_m - \theta_0}{\theta_0}$, where θ_0 denotes the true value for the population parameters, is depicted in Figure 3.9.

b) Results

According to Tables 3.1 and 3.2, the population parameter estimates show very little bias except for parameters σ and γ, since in general, $\bar{\theta}$ takes values very close to those of θ_0. When the number of simulated observations per subject is $n = 25$, parameters γ and σ are under-estimated, whatever the number of subjects in the simulated datasets ($N = 30$ or $N = 100$). The order of the median relative error for $\hat{\gamma}$ is about 10%, whereas it is about 20%

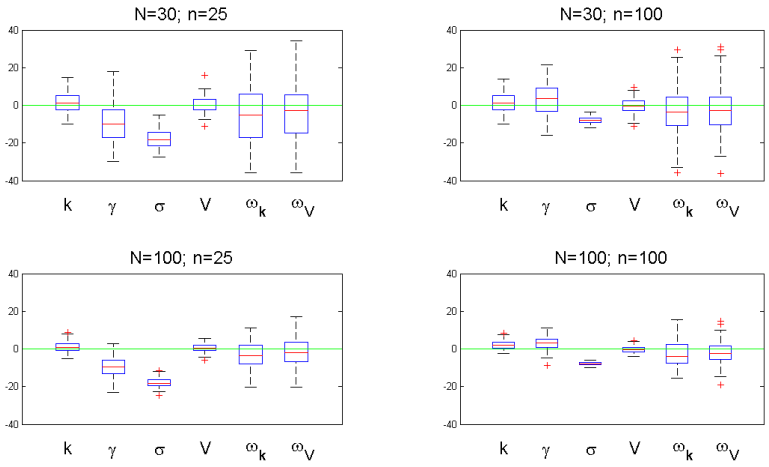

Figure 3.9: Results of the Monte Carlo simulation study: distribution of the relative estimation errors for the different designs for (N, n).

for $\hat{\sigma}$. Due to the intricate form of the dynamical model, strong identifiability issues, thus estimation difficulties, are expected. Estimation results reveal $\hat{\gamma}$ and $\hat{\sigma}$ to be correlated. The coefficient of correlation between the two estimates is estimated up to -0.6 with a stochastic approximation procedure. Nevertheless, Figure 3.9 shows that increasing the number of simulated observations per subject decreases bias in the estimation of both parameters σ and γ. The standard errors of the estimates are quite small, whatever the values for N and n (Tables 3.1 and 3.2). We also note that increasing the number of simulated subjects reduces variance of the estimates. Else, we notice from Tables 3.1 and 3.2 that $\bar{se}$ is close to the standard deviation of the estimates. During calibration of the algorithm, we have however noticed high sensitivity of the algorithm to the choice of the initial values, especially for parameters γ and σ. Considering one single simulated dataset,convergence is obtained in about 200 iterations, and requires 880 seconds CPU time (processors: 8 cores Intel i7-2600 at 3.4 GHz) for $N = 30$ subjects and $n = 25$ observations per subject, and 5950 seconds CPU time in largest simulated datasets ($N = 100$, $n = 100$). Simulations with other parameter values give similar results. Figure 3.10 depicts some individual simulated profiles from mixed-effects bolus model and illustrates reconstruction of the drug's dynamics with the Kalman smoother. The graphics shows that the kinetics estimated with the Kalman smoother are close to the simulated ones.

3.2.5 Discussion

This paper proposes a maximum likelihood estimation method for mixed-effects diffusion models, observed at discrete time points up to a Gaussian noise. These models have direct applications, especially for dynamical systems with linear transfers, like in pharmacokinetics. More precisely, our approach is to define the transfer rates as diffusion processes to model the perturbations observed on the sytems rather than constants as it is often done in practice. We suggest a specific extension of the SAEM algorithm combined with the extended Kalman filter for estimating the population parameters in these models. The performances of

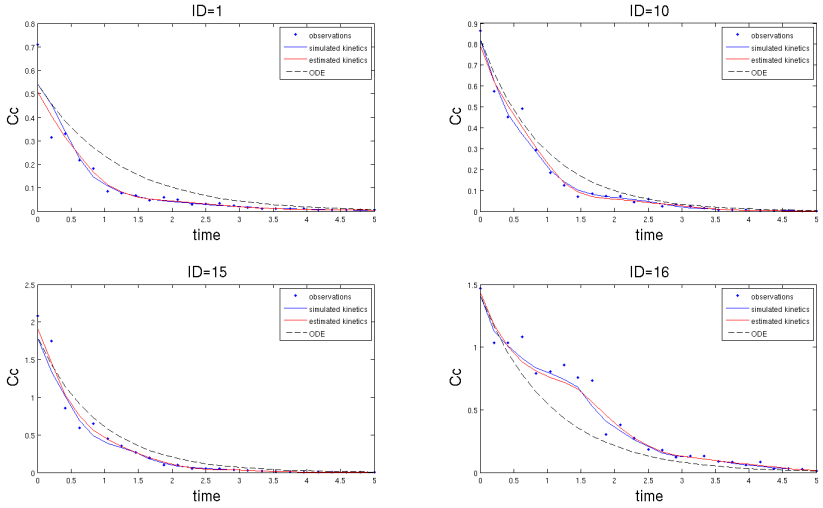

Figure 3.10: Reconstruction of the drug's dynamics for four simulated profiles using the Kalman smoother.

the new estimation method are illustrated with a brief Monte-Carlo simulation study based on a stochastic extension of PK bolus model. The simulation globally show satisfying performances of the new algorithm, except for the estimation of the volatility parameter and the estimation of the variance of the measurement errors. At present, the reason for this is unclear. The estimates for these two parameters are biased, but the bias reduces or even vanishes when the size of the simulated dataset is very high. The implemented version of the continuous-discrete extended Kalman filter does perhaps introduce errors when the time intervals between consecutive observations are large (corresponding to small values for n in the present simulation study), but such problem could be addressed by incorporating an adaptive controlled step size into the algorithm, as it has been suggested in [17]. On the other hand, mixed-effects diffusion models pose strong identifiability issues, leading to non-optimal properties of the maximum-likelihood estimates of some parameters of the model. In particular here, the simulation results could suggest that the MLE for parameters γ and σ are biased when the sample size is finite, but are asymptotically unbiased. This should be theoretically checked in future works.

3.2.6 Tables

		$N = 30, n = 25$			$N = 30, n = 100$		
Parameter	θ_0	$\bar{\theta}$	sd	$\bar{se}$	$\bar{\theta}$	sd	$\bar{se}$
k^*	0.2	0.203	0.011	0.011	0.203	0.011	0.011
V	0.5	0.501	0.0203	0.0184	0.498	0.019	0.018
γ	0.3	0.271	0.033	0.027	0.309	0.024	0.020
σ	0.1	0.082	0.005	0.004	0.092	0.002	0.001
ω_k	0.3	0.284	0.045	0.038	0.290	0.043	0.039
ω_V	0.2	0.194	0.027	0.027	0.194	0.025	0.026

Table 3.1: Results of the Monte Carlo simulation study for $N = 30$: for each parameter of the model, $n = 25$ and $n = 100$, the table displays the mean estimate, the standard deviation of the estimates and the mean estimated standard error.

		$N = 100, n = 25$			$N = 100, n = 100$		
Parameter	θ_0	$\bar{\theta}$	sd	$\bar{se}$	$\bar{\theta}$	sd	$\bar{se}$
k^*	0.2	0.202	0.006	0.006	0.204	0.005	0.006
V	0.5	0.502	0.011	0.010	0.498	0.011	0.010
γ	0.3	0.270	0.029	0.014	0.309	0.012	0.011
σ	0.1	0.081	0.007	0.002	0.0922	0.001	0.001
ω_k	0.3	0.291	0.025	0.021	0.292	0.020	0.022
ω_V	0.2	0.197	0.015	0.015	0.195	0.014	0.014

Table 3.2: Results of the Monte Carlo simulation study for $N = 100$: for each parameter of the model, $n = 25$ and $n = 100$, the table displays the mean estimate, the standard deviation of the estimates and the mean estimated standard error.

3.3 Deuxième article : Maximum likelihood estimation for stochastic differential equations with random effects

Cet article, co-écrit avec Adeline Samson et Valentine Genon-Catalot, est actuellement soumis à Scandinavian Journal of Statistics et disponible sur le HAL [4].

Abstract

We consider N independent stochastic processes $(X_i(t), t \in [0, T_i])$, $i = 1, \ldots, N$, defined by a stochastic differential equation with drift term depending on a random variable ϕ_i. The distribution of the random effect ϕ_i depends on unknown parameters which are to be estimated from the continuous observation of the processes X_i. We give the expression of the exact likelihood. When the drift term depends linearly on the random effect ϕ_i and ϕ_i has Gaussian distribution, an explicit formula for the likelihood is obtained. We prove that the maximum likelihood estimator is consistent and asymptotically Gaussian, when $T_i = T$ for all i and N tends to infinity. We discuss the case of discrete observations. Estimators are computed on simulated data for several models and show good performances even when the length time interval of observations is not very large.

Keywords: Asymptotic normality, Consistency, Maximum likelihood estimator, Mixed-effects models, Stochastic differential equations.

3.3.1 Introduction

Statistical analysis of data collected over time on a series of subjects requires to account for both the intra-individual variability, *i.e.* the variability occurring within the dynamics of each individual over time, and the variability existing between subjects. Modeling of such data goes through mixed-effects models which are very popular in the biomedical field [3, 25]. In mixed-effects stochastic differential equations (SDEs), the model for each individual set of data is given by a SDE, thus modeling the intra-individual variability in the data, and the parameters of each individual SDE are random variables, thus handling the variability between subjects. A major area of application for mixed effects SDEs is in pharmacokinetic/pharmacodynamic modeling, where they have been introduced as an alternative to the classical ODE-based models [7, 22, 8]. SDEs with random effects have also been proposed for neuronal data [24].
Maximum likelihood estimation (MLE) of the parameters of the random effects, also called population parameters, is generally not straightforward as the likelihood function can rarely be expressed in a closed-form. Approximations of the likelihood have been proposed, based on linearization [2] or Laplace's approximation [29]. Alternative methods have also been developed such as the SAEM algorithm [13]. Maximum likelihood estimation in SDEs with random effects has been tackled in a few papers. [7] show that in the specific case of a mixed-effects Brownian motion with drift, the likelihood function can be explicitly derived, leading to explicit parameters estimators. For general mixed SDEs, approximations of the likelihood have been proposed [24, 23].

For theoretical properties of the MLE in the context of mixed effects models, the main contribution to our knowledge is due to [21, 19, 20] and covers the asymptotic properties of the MLE for the population parameters under several asymptotic frameworks, depending on whether the number of subjects and/or the number of observations per subject goes to infinity. Nie's results are nevertheless based on a series of technical assumptions, which may be uneasy to check.

In the present work, we focus on mixed-effects SDEs with drift term depending on random effects and diffusion term without random effects. More precisely, we consider N real valued stochastic processes $(X_i(t), t \geq 0)$, $i = 1, \ldots, N$, with dynamics ruled by the following SDEs:

$$d\,X_i(t) = b(X_i(t), \phi_i)dt + \sigma(X_i(t))\,dW_i(t), \quad X_i(0) = x^i,\ i = 1, \ldots, N, \tag{3.16}$$

where $(W_1, \ldots, W_N)$ are N independent Wiener processes, $\phi_1, \ldots, \phi_N$ are N *i.i.d.* $\mathbb{R}^d$-valued random variables, $(\phi_1, \ldots, \phi_N)$ and $(W_1, \ldots, W_N)$ are independent and $x^i, i = 1, \ldots, N$ are known real values. The diffusion coefficient $\sigma : \mathbb{R} \to \mathbb{R}$ is a known real-valued function. The drift function $b(x, \varphi)$ is a known function defined on $\mathbb{R} \times \mathbb{R}^d$ and real-valued. Each process $(X_i(t))$ represents an individual and the random vector ϕ_i represents the random effect of individual i. We assume that the random variables $\phi_1, \ldots, \phi_N$ have a common distribution $g(\varphi, \theta)d\nu(\varphi)$ on $\mathbb{R}^d$, where θ is an unknown parameter belonging to a set $\Theta \subset \mathbb{R}^p$ and, for all θ, $g(\varphi, \theta)$ is a density w.r.t. a dominating measure ν on $\mathbb{R}^d$. Below, we denote by θ_0 the true value of the parameter. The process $(X_i(t))$ is continuously observed on a time interval $[0, T_i]$ with $T_i > 0$ given. Our aim is to estimate the parameters θ of the density of the random effects from the observations $\{X_i(t), 0 \leq t \leq T_i, i = 1, \ldots, N\}$. We introduce assumptions ensuring that the models (3.16) are well-defined together with the exact likelihood function. Then, we focus on the special case of one-dimensional linear Gaussian random effects, *i.e.* $b(x, \phi_i) = \phi_i b(x)$, where b is a known real function and ϕ_i is Gaussian. It turns out in this case that the likelihood has a simple and explicit expression depending on θ and the sufficient statistics:

$$U_i = \int_0^{T_i} \frac{b(X_i(s))}{\sigma^2(X_i(s))} dX_i(s), \quad V_i = \int_0^{T_i} \frac{b^2(X_i(s))}{\sigma^2(X_i(s))} ds, \quad i = 1, \ldots, N.$$

For the asymptotic study, the main difficulties are encountered to obtain specific moment properties of the random variables (U_i, V_i), and to prove identifiability. We prove the consistency and the asymptotic normality of the exact MLE as N tends to infinity and give the expression of the Fisher information matrix. The results are extended to Gaussian multidimensional linear random effects. The present likelihood theory is derived from continuous observations of the $X_i's$. In practice, one rather disposes of discrete observations on the time interval $[0, T_i]$. Thus we suggest to discretize the r.v.'s U_i, V_i in the expression of estimators and we show that under conditions on the discretization step, and thus on the number of observations per subject, the asymptotic properties of the estimates based on continuous observations are preserved. Our simulations are presented within the framework of discretely observed stochastic processes.

The paper is organized as follows. Section 3.3.2 introduces the notations and assumptions. In Section 3.3.3, we make the likelihood function explicit. In Sections 3.3.4 and 3.3.5, we show that the strong consistency and the asymptotic normality of the MLE when the model includes a Gaussian one-dimensional and a Gaussian multi-dimensional random effect respectively. The impact of discretization on the estimators is detailed in Section 3.3.6. A simulation

study is presented in Section 3.3.7. Concluding remarks are given in Section 3.3.8. Proofs are gathered in Appendix.

3.3.2 Model, assumptions and notations

Consider N real valued stochastic processes $(X_i(t), t \geq 0)$, $i = 1, \ldots, N$, with dynamics ruled by (3.16). The processes $(W_1, \ldots, W_N)$ and the r.v.'s $\phi_1, \ldots, \phi_N$ are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We introduce assumptions ensuring that the processes (3.16) are well defined and allowing to compute the exact likelihood of our observations. Consider the filtration $(\mathcal{F}_t, t \geq 0)$ defined by $\mathcal{F}_t = \sigma(\phi_i, W_i(s), s \leq t, i = 1, \ldots, N)$. As $\mathcal{F}_t = \sigma(W_i(s), s \leq t) \vee \mathcal{F}_t^i$, with $\mathcal{F}_t^i = \sigma(\phi_i, \phi_j, W_j(s), s \leq t, j \neq i)$ independent of W_i, each process W_i is a $(\mathcal{F}_t, t \geq 0)$-Brownian motion. Moreover, the random variables ϕ_i are $\mathcal{F}_0$-measurable.

Assumption 3.1. (i) The function $(x, \varphi) \to b(x, \varphi)$ is C^1 on $\mathbb{R} \times \mathbb{R}^d$, and such that:

$$\exists K > 0, \forall (x, \varphi) \in \mathbb{R} \times \mathbb{R}^d, \quad b^2(x, \varphi) \leq K(1 + x^2 + |\varphi|^2),$$

(ii) The function $\sigma(.)$ is C^1 on $\mathbb{R}$ and

$$\forall x \in \mathbb{R}, \quad \sigma^2(x) \leq K(1 + x^2).$$

Under Assumption 3.1, for all φ, the stochastic differential equation

$$d\, X_i^\varphi(t) = b(X_i^\varphi(t), \varphi)dt + \sigma(X_i^\varphi(t))\, dW_i(t), \quad X_i^\varphi(0) = x^i, \tag{3.17}$$

admits a unique strong solution process $(X_i^\varphi(t), t \geq 0)$ adapted to the filtration $(\mathcal{F}_t, t \geq 0)$. Let $C(\mathbb{R}^+, \mathbb{R})$ be the space of continuous functions on $\mathbb{R}^+$, endowed with the Borel σ-field associated with the topology of uniform convergence on compact sets. The distribution of $X_i^\varphi(.)$ is uniquely defined on this space. Moreover, as x^i is deterministic, for all integer k, all φ and all $t \geq 0$,

$$\sup_{s \leq t} \mathbb{E}[X_i^\varphi(s)]^{2k} < +\infty. \tag{3.18}$$

For the observed processes, we have the following result.

Proposition 3.3.1. *Under Assumption 3.1, for $i = 1, \ldots, N$, equation (3.16) admits a unique solution process $(X_i(t), t \geq 0)$, adapted to the filtration $(\mathcal{F}_t, t \geq 0)$. Given that $\phi_i = \varphi$, the conditional distribution of $(X_i(t), t \geq 0)$ is identical to the distribution of the process $(X_i^\varphi(t), t \geq 0)$. The processes $(X_i(t), t \geq 0), i = 1, \ldots, N$ are independent. If for $k \geq 1$, $\mathbb{E}|\phi_i|^{2k} < \infty$, then for all $T > 0$, $\sup_{t \in [0,T]} \mathbb{E}[X_i(t)]^{2k} < \infty$.*

3.3.3 Likelihood

We introduce the canonical model associated with the observations. Let C_{T_i} denote the space of real continuous functions $(x(t), t \in [0, T_i])$ defined on $[0, T_i]$, endowed with the σ-field $\mathcal{C}_{T_i}$ associated with the topology of uniform convergence on $[0, T_i]$. Under Assumption 3.1, we introduce the distribution $Q_\varphi^{x^i, T_i}$ on $(C_{T_i}, \mathcal{C}_{T_i})$ of $(X_i^\varphi(t), t \in [0, T_i])$ given by (3.17). On $\mathbb{R}^d \times C_{T_i}$, let $P_\theta^i = g(\varphi, \theta) d\nu(\varphi) \otimes Q_\varphi^{x^i, T_i}$ denote the joint distribution of $(\phi_i, X_i(.))$ and let Q_θ^i denote the marginal distribution of $(X_i(t), t \in [0, T_i])$ on $(C_{T_i}, \mathcal{C}_{T_i})$. From now on, we denote by $(\phi_i, X_i(.))$ the canonical process of $\mathbb{R}^d \times C_{T_i}$. Let us consider the following assumptions.

Assumption 3.2. For $i = 1, \dots, N$, and for all φ, φ'

$$Q_{\varphi}^{x^i,T_i}\left(\int_0^{T_i} \frac{b^2(X_i^{\varphi}(t),\varphi')}{\sigma^2(X_i^{\varphi}(t))}dt < +\infty\right) = 1.$$

Assumption 3.3. For $f = \frac{\partial b}{\partial \varphi_j}, j = 1, \dots, d$, there exist $c > 0$ and some $\gamma \geq 0$ such that

$$\sup_{\varphi \in \mathbb{R}^d} \frac{|f(x,\varphi)|}{\sigma^2(x)} \leq c(1 + |x|^{\gamma}).$$

Proposition 3.3.2. *Suppose that Assumptions 3.1 to 3.3 are verified, and let $\varphi_0 \in \mathbb{R}^d$.*

- *The distributions $Q_{\varphi}^{x^i,T_i}$ are absolutely continuous w.r.t. $Q^i := Q_{\varphi_0}^{x^i,T_i}$ with density:*

$$\frac{dQ_{\varphi}^{x^i,T_i}}{dQ^i}(X_i) = L_{T_i}(X_i,\varphi) = e^{\ell_{T_i}(X_i,\varphi)} \quad with \quad \ell_{T_i}(X_i,\varphi) =$$

$$\int_0^{T_i} \frac{b(X_i(s),\varphi) - b(X_i(s),\varphi_0)}{\sigma^2(X_i(s))}dX_i(s) - \int_0^{T_i} \frac{b^2(X_i(s),\varphi) - b^2(X_i(s),\varphi_0)}{2\sigma^2(X_i(s))}ds,$$

  where $(X_i = X_i(s), s \leq T_i)$ denotes the canonical process of C_{T_i} given by $(X_i(s)(x) = x(s), s \leq T_i)$.
- *The function $\varphi \to L_{T_i}(X_i,\varphi)$ admits a continuous version Q^i-a.s. and $(X_i,\varphi) \to L_{T_i}(X_i,\varphi)$ is measurable on $(C_{T_i} \times \mathbb{R}^d, \mathcal{C}_{T_i} \otimes \mathcal{B}(\mathbb{R}^d))$.*

Remark 10. For a given drift function $b(x,\varphi)$, it is often possible to check directly that $\varphi \to L_{T_i}(X,\varphi)$ is continuous even if Assumption 3.3 is not fulfilled. Then, the joint measurability follows.

To simplify notations, we assume that there is one value φ_0 such that $b(x,\varphi_0) \equiv 0$. Thus, we can choose the dominating measure $Q^i = Q_{\varphi_0}^{x^i,T_i}$ which is the distribution of (3.17) with nul drift. Formula $L_{T_i}(X_i,\varphi)$ simplifies into:

$$L_{T_i}(X_i,\varphi) = \exp\left(\int_0^{T_i} \frac{b(X_i(s),\varphi)}{\sigma^2(X_i(s))}dX_i(s) - \frac{1}{2}\int_0^{T_i} \frac{b^2(X_i(s),\varphi)}{\sigma^2(X_i(s))}ds\right). \tag{3.19}$$

By independence of the individuals, $P_{\theta} = \otimes_{i=1}^N P_{\theta}^i$ is the distribution of $(\phi_i, X_i(.))$, $i = 1, \dots, N$ on the product space $\prod_{i=1}^N \mathbb{R}^d \times C_{T_i}$ and $Q_{\theta} = \otimes_{i=1}^N Q_{\theta}^i$ is the distribution of the whole sample $(X_i(t), t \in [0,T_i], i = 1, \dots, N)$ on $C = \prod_{i=1}^N C_{T_i}$. We now compute the density of Q_{θ} w.r.t. $Q = \otimes_{i=1}^N Q^i$. We denote by E_{θ} the expectation w.r.t. P_{θ}.

Proposition 3.3.3. *Suppose that Assumptions 3.1 to 3.3 are verified.*

- *The probability measure Q_{θ}^i admits a density w.r.t. Q^i equal to:*

$$\frac{dQ_{\theta}^i}{dQ^i}(X_i) = \int_{\mathbb{R}^d} L_{T_i}(X_i,\varphi)g(\varphi,\theta)d\nu(\varphi) := \lambda_i(X_i,\theta).$$

– *The distribution* Q_θ *on* $C = \prod_{i=1}^{N} C_{T_i}$ *admits a density given by*

$$\frac{dQ_\theta}{dQ}(X_1, \ldots, X_N) = \prod_{i=1}^{N} \lambda_i(X_i, \theta).$$

– *The exact likelihood of the whole sample* $(X_i(t), t \in [0, T_i], i = 1, \ldots, N)$ *is*

$$\Lambda_N(\theta) = \prod_{i=1}^{N} \lambda_i(X_i, \theta). \tag{3.20}$$

On this general expression, if we can check [19]'s assumptions, then weak consistency and asymptotic normality of the MLE will follow. However, these assumptions, even when the random effects are Gaussian, are uneasy.

3.3.4 Gaussian one-dimensional linear random effects

In this section, we consider model (3.16) with drift $b(x, \varphi) = \varphi b(x)$ where $\varphi \in \mathbb{R}$, $b(.), \sigma(.)$ are known functions. In this case, we simplify Assumptions 3.1-3.2 and assume that b, σ are C^1 and have linear growth, which implies Proposition 3.3.1. And, we assume that

$$\int_0^{T_i} b^2(X_i(s))/\sigma^2(X_i(s))ds < \infty,$$

$Q_\varphi^{x^i, T_i}$-a.s. for all φ. As $\varphi \to L_{T_i}(X_i, \varphi)$ is obviously continuous, Assumption 3.3 is not required. We also assume that, for $i = 1, \ldots, N$, $T_i = T, x^i = x$ so that the observed processes $(X_i(t), t \in [0, T]), i = 1, \ldots, N$ are *i.i.d.*. Let us introduce:

$$U_i = \int_0^T \frac{b(X_i(s))}{\sigma^2(X_i(s))} dX_i(s), \quad V_i = \int_0^T \frac{b^2(X_i(s))}{\sigma^2(X_i(s))} ds, \tag{3.21}$$

which are well defined under Assumptions 3.1-3.2. Hence,

$$\lambda_i(X_i, \theta) = \int_{\mathbb{R}} g(\varphi, \theta) \exp\left(\varphi U_i - \frac{\varphi^2}{2} V_i\right) d\nu(\varphi). \tag{3.22}$$

3.3.4.1 Exact likelihood

We propose here to model the random effects distribution by a Gaussian distribution $\mathcal{N}(\mu, \omega^2)$, and set $\theta = (\mu, \omega^2) \in \mathbb{R} \times (0, +\infty)$ for the unknown parameters to be estimated. This choice leads to an explicit exact likelihood.

Proposition 3.3.4. *Assume that* $g(\varphi, \theta) d\nu(\varphi) = \mathcal{N}(\mu, \omega^2)$. *Then,*

$$\lambda_i(X_i, \theta) = \frac{1}{(1 + \omega^2 V_i)^{1/2}} \exp\left[-\frac{V_i}{2(1 + \omega^2 V_i)} \left(\mu - \frac{U_i}{V_i}\right)^2\right] \exp\left(\frac{U_i^2}{2V_i}\right).$$

The conditional distribution, under P_θ^i, *of* ϕ_i *given* X_i *is the distribution*

$$\mathcal{N}\left(\frac{\mu + \omega^2 U_i}{1 + \omega^2 V_i}, \frac{\omega^2}{1 + \omega^2 V_i}\right).$$

Therefore, the logarithm of the likelihood function (3.20) is explicitely given by

$$\mathcal{L}_N(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\log(1+\omega^2 V_i) - \frac{1}{2}\sum_{i=1}^{N}\frac{V_i}{1+\omega^2 V_i}\left(\mu - \frac{U_i}{V_i}\right)^2 + \sum_{i=1}^{N}\frac{U_i^2}{2V_i}. \tag{3.23}$$

The derivatives of the log-likelihood (3.23) are

$$\begin{aligned}
\frac{\partial}{\partial\mu}\mathcal{L}_N(\theta) &= \sum_{i=1}^{N}\left(\frac{U_i}{1+\omega^2 V_i} - \mu\frac{V_i}{1+\omega^2 V_i}\right), \\
\frac{\partial}{\partial\omega^2}\mathcal{L}_N(\theta) &= \frac{1}{2}\sum_{i=1}^{N}\left[\left(\frac{U_i}{1+\omega^2 V_i} - \mu\frac{V_i}{1+\omega^2 V_i}\right)^2 - \frac{V_i}{1+\omega^2 V_i}\right].
\end{aligned}$$

When ω_0^2 is known, we obtain the explicit estimator for μ_0:

$$\widehat{\mu}_N = \frac{\displaystyle\sum_{i=1}^{N}\frac{U_i}{1+\omega_0^2 V_i}}{\displaystyle\sum_{i=1}^{N}\frac{V_i}{1+\omega_0^2 V_i}}. \tag{3.24}$$

When both parameters are unknown, the maximum likelihood estimators of $\theta_0 = (\mu_0, \omega_0^2)$ are given by the system:

$$\widehat{\mu}_N = \left(\sum_{i=1}^{N}\frac{V_i}{1+\widehat{\omega}_N^2 V_i}\right)^{-1}\left(\sum_{i=1}^{N}\frac{U_i}{1+\widehat{\omega}_N^2 V_i}\right),$$

$$\sum_{i=1}^{N}\left(\widehat{\mu}_N - \frac{U_i}{V_i}\right)^2\frac{V_i^2}{(1+\widehat{\omega}_N^2 V_i)^2} = \sum_{i=1}^{N}\frac{V_i}{1+\widehat{\omega}_N^2 V_i}.$$

Remark 11. Note that, when the effect ϕ_i is non random and $\phi_i \equiv \mu_0$, the estimator of μ_0 is standardly given by:

$$\widetilde{\mu}_N = \frac{\displaystyle\sum_{i=1}^{N} U_i}{\displaystyle\sum_{i=1}^{N} V_i}, \tag{3.25}$$

which corresponds to $\omega_0^2 = 0$ in $\widehat{\mu}_N$.

3.3.4.2 Preliminary moments properties

For studying the maximum likelihood estimators of $\theta = (\mu, \omega^2)$, we need investigate properties of the following random variables:

$$\gamma_i(\theta) = \frac{U_i - \mu V_i}{1+\omega^2 V_i}, \quad I_i(\omega^2) = \frac{V_i}{1+\omega^2 V_i}. \tag{3.26}$$

Indeed under Q_θ, $(\gamma_i(\theta), I_i(\omega^2))_{i=1,\dots,N}$ are *i.i.d.* and the score function is

$$\frac{\partial}{\partial\mu}\mathcal{L}_N(\theta) = \sum_{i=1}^{N}\gamma_i(\theta), \qquad \frac{\partial}{\partial\omega^2}\mathcal{L}_N(\theta) = \frac{1}{2}\sum_{i=1}^{N}(\gamma_i^2(\theta) - I_i(\omega^2)). \tag{3.27}$$

Evidently, $0 < I_i(\omega^2) \leq 1/\omega^2$ is bounded. By the following lemma, which is crucial for the statistical study, we prove that $\gamma_i(\theta)$ admits a finite Laplace transform, hence moments of any order.

Lemma 3.3.5. *For all $\theta = (\mu, \omega^2) \in \mathbb{R} \times (0, +\infty)$, and all $u \in \mathbb{R}$,*

$$E_\theta(\exp(u\frac{U_1}{1+\omega^2 V_1})) < +\infty.$$

We can now compute some useful moments of functions of $\gamma_1(\theta), I_1(\omega^2)$.

Proposition 3.3.6. *For all $\theta \in \mathbb{R} \times (0, +\infty)$, the following relations hold:*

$$E_\theta(\gamma_1(\theta)) = 0, \quad E_\theta(\gamma_1^2(\theta)) = E_\theta(I_1(\omega^2)), \quad E_\theta(\gamma_1^3(\theta)) = 3E_\theta(\gamma_1(\theta)\, I_1(\omega^2)),$$

$$E_\theta\left(\gamma_1^2(\theta) - I_1(\omega^2)\right)^2 = 4E_\theta(\gamma_1^2(\theta)I_1(\omega^2)) - 2E_\theta(I_1^2(\omega^2)).$$

3.3.4.3 Convergence in distribution of the normalized score function

Based on Lemma 3.3.5 and Proposition 3.3.6, we can state:

Proposition 3.3.7. *For all θ, under Q_θ, as N tends to infinity, the random vector*

$$\frac{1}{\sqrt{N}}\begin{pmatrix} \frac{\partial}{\partial\mu}\mathcal{L}_N(\theta) \\ \frac{\partial}{\partial\omega^2}\mathcal{L}_N(\theta) \end{pmatrix} = \frac{1}{\sqrt{N}}\begin{pmatrix} \sum_{i=1}^{N}\gamma_i(\theta) \\ \frac{1}{2}\sum_{i=1}^{N}(\gamma_i^2(\theta) - I_i(\omega^2)) \end{pmatrix}$$

converges in distribution to $\mathcal{N}_2(0, \mathcal{I}(\theta))$ and the matrix

$$-\frac{1}{N}\begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\mathcal{L}_N(\theta) & \frac{\partial^2}{\partial\mu\partial\omega^2}\mathcal{L}_N(\theta) \\ \frac{\partial^2}{\partial\mu\partial\omega^2}\mathcal{L}_N(\theta) & \frac{\partial^2}{\partial\omega^2\partial\omega^2}\mathcal{L}_N(\theta) \end{pmatrix}$$

converges in probability to $\mathcal{I}(\theta)$ where

$$\mathcal{I}(\theta) = \begin{pmatrix} E_\theta(I_1(\omega^2)) & E_\theta(\gamma_1(\theta)I_1(\omega^2)) \\ E_\theta(\gamma_1(\theta)I_1(\omega^2)) & E_\theta(\gamma_1^2(\theta)I_1(\omega^2)) - \frac{1}{2}E_\theta(I_1^2(\omega^2)) \end{pmatrix} \tag{3.28}$$

is the covariance matrix of the vector

$$\begin{pmatrix} \gamma_1(\theta) \\ \frac{1}{2}(\gamma_1^2(\theta) - I_1(\omega^2)) \end{pmatrix}.$$

The following corollary holds immediately.

Corollary 3.3.8. *When $\omega^2 = \omega_0^2$ is known, the explicit estimator $\widehat{\mu}_N$ (3.24) is consistent and $\sqrt{N}(\widehat{\mu}_N - \mu_0)$ converges in distribution under Q_{μ_0} to $\mathcal{N}(0, 1/E_{\mu_0}(V_1/(1+\omega_0^2 V_1)))$ where $1/E_{\mu_0}(V_1/(1+\omega_0^2 V_1)) \geq \omega_0^2$.*

If $\phi_1, \ldots, \phi_N$ were observed, the MLE of μ_0 would be $\bar{\phi} = \frac{1}{N}(\phi_1 + \ldots + \phi_N)$ which satisfies $\sqrt{N}(\bar{\phi} - \mu_0) \sim \mathcal{N}(0, \omega_0^2)$. As $\phi_1, \ldots, \phi_N$ are not observed, we obtain that the MLE $\widehat{\mu}_N$ has a larger asymptotic variance.

3.3.4.4 Consistency and asymptotic normality

When both parameters μ_0, ω_0^2 are unknown, we need to introduce additional assumptions to prove identifiability and consistency. Recall that Q is the distribution on C such that the canonical processes $(X_i(s), s \leq T, i = 1, \ldots, N)$ are *i.i.d.* and X_i satisfies the SDE with nul drift:

$$dX_i(t) = \sigma(X_i(t))dW_i(t), \quad X_i(0) = x.$$

We assume that

Assumption 3.4. The function $b(.)/\sigma(.)$ is not constant. Under Q, the random variable (U_1, V_1) admits a density $f(u, v)$ w.r.t. the Lebesgue measure on $\mathbb{R} \times (0, +\infty)$ which is jointly continuous and positive on an open ball of $\mathbb{R} \times (0, +\infty)$.

Assumption 3.5. The parameter set Θ is a compact subset of $\mathbb{R} \times (0, +\infty)$.

Assumption 3.6. The true value θ_0 belongs to $\mathring{\Theta}$.

Assumption 3.7. The matrix $\mathcal{I}(\theta_0)$ is invertible (see (3.28)).

Under smoothness assumptions on functions b, σ, Assumption 3.4 will be fulfilled by application of Malliavin calculus tools [1]. The case where $b(.)/\sigma(.)$ is constant is rather simple and is treated separately in Section 3.3.7. Assumptions 3.5 to 3.7 are classical. We first state an identifiability result.

Proposition 3.3.9. *Set $K(Q^1_{\theta_0}, Q^1_\theta)$ the Kullback information of $Q^1_{\theta_0}$ w.r.t. Q^1_θ.*

(i) Under Assumptions 3.1, 3.2 and 3.4, $Q^1_\theta = Q^1_{\theta_0}$ implies that $\theta = \theta_0$. Hence, $\theta \to K(Q^1_{\theta_0}, Q^1_\theta)$ admits a unique minimum at $\theta = \theta_0$.

(ii) Under Assumptions 3.1-3.2, the function $\theta \to K(Q^1_{\theta_0}, Q^1_\theta)$ is continuous on $\mathbb{R} \times (0, +, \infty)$.

We are now able to prove the consistency and asymptotic normality of $\hat{\theta}_N$.

Proposition 3.3.10. *1. Assume Assumptions 3.1-3.2 and 3.4-3.5 are verified. Let $\hat{\theta}_N$ be a maximum likelihood estimator defined as any solution of $\mathcal{L}_N(\hat{\theta}_N) = \sup_{\theta \in \Theta} \mathcal{L}_N(\theta)$. Under Q_{θ_0}, $\hat{\theta}_N$ converges in probability to θ_0.*

2. Assume Assumptions 3.1-3.2 and 3.4 to 3.7 are verified. The maximum likelihood estimator satisfies, as N tends to infinity,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to_{\mathcal{D}} \mathcal{N}_2(0, \mathcal{I}^{-1}(\theta_0)).$$

Note that the consistency obtained here is a strong consistency in the sense that any solution of the likelihood equation is consistent.

1. If σ and $f = b/\sigma$ are C^∞, if their derivatives of any order greater than 1 are bounded, if σ is bounded below, and if the Lebesgue measure of the set of values x such that $f(x)f'(x) = 0$ is zero, then assumption (H4) holds.

3.3.5 Gaussian multidimensional linear random effects

In this section, we extend the previous results to multidimensional linear random effects. Let $\phi_i = (\phi_i^1, \dots, \phi_i^d)'$ be a d-dimensional random vector and $b(x) = (b^1(x), \dots, b^d(x))'$ be a function $\mathbb{R} \to \mathbb{R}^d$. Consider the SDE

$$d\,X_i(t) = \phi_i'\, b(X_i(t))dt + \sigma(X_i(t))\, dW_i(t), \quad X_i(0) = x. \tag{3.29}$$

We assume that $b^1(x), \dots, b^d(x)$ are such that $b(x, \varphi) = \sum_{j=1}^{d} \varphi^j b^j(x)$ satisfies Assumptions 3.1-3.2 and that $(\phi_i, i = 1, \dots, N)$ are i.i.d. Gaussian vectors, with expectation vector μ and covariance matrix $\Omega \in \mathcal{S}_d(\mathbb{R})$ where $\mathcal{S}_d(\mathbb{R})$ is the set of positive definite symetric matrices. The parameter to be estimated is $\theta = (\mu, \Omega) \in \mathbb{R}^d \times \mathcal{S}_d(\mathbb{R})$. To compute the likelihood, we introduce the random vectors

$$U_i \;=\; \int_0^T \frac{b(X_i(s))}{\sigma^2(X_i(s))} dX_i(s),$$

and the $d \times d$ random matrices

$$V_i \;=\; \int_0^T \frac{b(X_i(s))b'(X_i(s))}{\sigma^2(X_i(s))} ds.$$

The following assumption is now required.

Assumption 3.8. For $i = 1, \dots, N$ the matrix V_i is positive definite Q^i-a.s. and Q_θ^i-a.s. for all θ.

If the functions (b^j/σ^2) are not linearly independent, Assumption 3.8 is not true. Thus, Assumption 3.8 can be interpreted as ensuring a well-defined dimension of the vector ϕ_i.

We deduce the invertibility of matrices involved in the likelihood computation.

Lemma 3.3.11. *Under Assumption 3.8, the matrices $V_i + \Omega^{-1}$, $I_d + V_i\Omega$, $I_d + \Omega V_i$ are invertible Q^i-a.s. and Q_θ^i-a.s. for all θ.*

Then we can compute the likelihood.

Proposition 3.3.12. *Under Assumption 3.8, set $R_i^{-1} = (I_d + V_i\Omega)^{-1}V_i$, we have*

$$\lambda_i(X_i, \theta) = \frac{1}{\sqrt{\det(I_d + V_i\Omega)}} \exp\left[-\frac{1}{2}(\mu - V_i^{-1}U_i)'R_i^{-1}(\mu - V_i^{-1}U_i)\right] \exp\left(\frac{1}{2}U_i'V_i^{-1}U_i\right).$$

The conditional distribution of ϕ_i given X_i is the Gaussian distribution

$$\mathcal{N}_d\left((I_d + \Omega V_i)^{-1}\mu + (\Omega^{-1} + V_i)^{-1}V_i, (I_d + \Omega V_i)^{-1}\Omega\right)$$

The likelihood is $\Lambda_N(\theta) = \prod_{i=1}^{N} \lambda_i(X_i, \theta)$.

The score function (respectively a d-vector and a $d \times d$ matrix) is given by:

$$\frac{\partial}{\partial \mu}\mathcal{L}_N(\theta) = \sum_{i=1}^{N} \gamma_i(\theta), \qquad \frac{\partial}{\partial \Omega}\mathcal{L}_N(\theta) = \frac{1}{2}\sum_{i=1}^{N}(\gamma_i(\theta)\gamma_i'(\theta) - I_i(\Omega))$$

where $\theta = (\mu, \Omega)$ and :

$$\gamma_i(\theta) = (I_d + \Omega V_i)^{-1}(U_i - V_i\,\mu), \quad I_i(\Omega) = (I_d + \Omega V_i)^{-1}V_i. \tag{3.30}$$

When Ω_0 is known, the estimator for μ_0 is explicit.

Lemma 3.3.5 and Propositions 3.3.6 and 3.3.7 can be readily extended to the multidimensional case.

Proposition 3.3.13. *1. For all $\theta = (\mu, \Omega) \in \mathbb{R} \times (0, +\infty)$, and all $u \in \mathbb{R}$,*

$$E_\theta(\exp\,(u'(I_d + \Omega V_i)^{-1}U_1)) < +\infty.$$

2. For all $\theta \in \mathbb{R} \times (0, +\infty)$, the following relations hold:

$$E_\theta(\gamma_1(\theta)\gamma_1'(\theta)) = E_\theta(I_1(\Omega)), \quad E_\theta(\gamma_1(\theta)\gamma_1'(\theta)\gamma_1(\theta)) = 3E_\theta(I_1(\Omega)\,\gamma_1(\theta)),$$

$$E_\theta(\gamma_1(\theta)) = 0, \quad E_\theta\left(\gamma_1(\theta)\gamma_1'(\theta) - I_1(\Omega)\right)^2 = 4E_\theta(I_1(\Omega)\gamma_1(\theta)\gamma_1'(\theta)) - 2E_\theta(I_1^2(\Omega)).$$

3. For all θ, under Q_θ, as N tends to infinity, the random vector

$$\frac{1}{\sqrt{N}}\begin{pmatrix} \frac{\partial}{\partial \mu}\mathcal{L}_N(\theta) \\ \frac{\partial}{\partial \Omega}\mathcal{L}_N(\theta) \end{pmatrix} = \frac{1}{\sqrt{N}}\begin{pmatrix} \sum_{i=1}^{N}\gamma_i(\theta) \\ \frac{1}{2}\sum_{i=1}^{N}(\gamma_i(\theta)\gamma_i'(\theta) - I_i(\Omega)) \end{pmatrix}$$

converges in distribution to $\mathcal{N}(0, \mathcal{I}(\theta))$ where $\mathcal{I}(\theta)$ is the covariance matrix of the vector

$$\begin{pmatrix} \gamma_1(\theta) \\ \frac{1}{2}(\gamma_1(\theta)\gamma_1'(\theta) - I_1(\Omega)) \end{pmatrix}$$

which is also the limit of the observed Fisher information matrix.

The study of $\widehat{\theta}_N = (\widehat{\mu}_N, \widehat{\Omega}_N)$ can be done as above.

3.3.6 Discrete data

In this section, we briefly discuss the case of discrete data. Let us assume that we observe synchronously the processes $X_i(t)$ at times $t_k^n = t_k = k\frac{T}{n}$, $k = 0, 1, \ldots, n$. To build estimators $\widehat{\theta}_N^{(n)}$ based on these data, we simply replace the r.v.'s U_i, V_i, $i = 1, \ldots, N$ by their discretized versions:

$$U_i^n = \sum_{k=0}^{n-1} \frac{b(X_i(t_k))}{\sigma^2(X_i(t_k))}(X_i(t_{k+1}) - X_i(t_k)), \tag{3.31}$$

$$V_i^n = \sum_{k=0}^{n-1} \frac{b^2(X_i(t_k))}{\sigma^2(X_i(t_k))}(t_{k+1} - t_k). \tag{3.32}$$

Looking at the expressions of (3.23) and its derivatives w.r.t. θ, it is enough to study the differences $U_i - U_i^n$, $V_i - V_i^n$. We can prove

Lemma 3.3.14. *Assume that b/σ is bounded and Lipschitz, $\sigma(.) \geq \epsilon > 0$, b and σ Lipschitz, then for all $p \geq 1$ and all $i = 1, \ldots, N$, there exists a constant C such that*

$$\mathbb{E}_{\theta_0}(|V_i - V_i^n|^p + |U_i - U_i^n|^p) \leq \frac{C}{n^{p/2}}.$$

We deduce

Proposition 3.3.15. *If $n \to +\infty$, then $\widehat{\theta}_N - \widehat{\theta}_N^{(n)} = o_{P_{\theta_0}}(1)$. If $n = n(N) \to +\infty$ in such a way that $\frac{n}{N} \to +\infty$, then $\sqrt{N}(\widehat{\theta}_N - \widehat{\theta}_N^{(n)}) = o_{P_{\theta_0}}(1)$.*

3.3.7 Simulation study

Several models are simulated. For each SDE model, 100 datasets are generated with N subjects on the same time interval $[0, T]$ and three experimental designs: $(N = 20, T = 5)$, $(N = 50, T = 5)$ and $(N = 50, T = 10)$. The empirical mean and variance of the MLE are computed from the 100 datasets. When possible, $N^{-1}\mathcal{I}_N^{-1}(\theta_0)$ is also computed and compared with the empirical variance.

3.3.7.1 When $b(x) = c\,\sigma(x)$

Let us consider the case where $b(x) = c\,\sigma(x)$, with $c \neq 0$ known. Then we have

$$V_i = c^2 T, \quad U_i = c \int_0^T \frac{dX_i(s)}{\sigma(X_i(s))}. \tag{3.33}$$

The estimators of μ_0, ω_0^2 are simple and explicit:

$$\widehat{\mu}_N = \frac{1}{c^2 TN} \sum_{i=1}^N U_i = \frac{1}{c^2 T} \bar{U}_N, \quad \widehat{\omega}_N^2 = \frac{1}{(c^2 T)^2} \left(\frac{1}{N} \sum_{i=1}^N (U_i - \bar{U}_N)^2 - c^2 T \right).$$

Using that $U_i = c^2 T \phi_i + c W_i(T)$, an elementary study shows that $\widehat{\mu}_N$ and $\widehat{\omega}_N^2$ are strongly consistent, that $\sqrt{N}(\widehat{\mu}_N - \mu_0)$ has distribution $\mathcal{N}(0, \omega_0^2 + 1/(c^2 T))$ and that $\sqrt{N}(\widehat{\omega}_N^2 - \omega_0^2)$ converges in distribution to $\mathcal{N}(0, 2(\omega_0^2 + 1/(c^2 T))^2)$. The asymptotic variances of $\widehat{\mu}_N$ and $\widehat{\omega}_N^2$ are increased in comparison with the case of non random effects.
We stress the fact that, whatever the drift funtion $b(.)$, when $b(.)/\sigma(.)$ is constant, the estimators have the same distribution.

***Example* 3.3.** Consider a mixed-effects Brownian motion with drift

$$dX_i(t) \;=\; \phi_i dt + \sigma\, dW_i(t), \quad X_i(0) = 0,$$

with $\phi_i \sim \mathcal{N}(\mu, \omega^2)$. This model is considered by [7] and the estimators are the same. We use two sets of population parameters: $(\mu = -1,\ \omega^2 = 1)$ and $(\mu = 5,\ \omega^2 = 1)$. All simulations are performed with a discretization step-size $\delta = 0.001$ on $[0, T]$ and $\sigma = 1$ (σ known). Results, presented in Table 3.3, are satisfactory overall. Increasing N improves the accuracy of both estimates $\widehat{\mu}_N$ and $\widehat{\omega}_N^2$. For $T = 5$, both estimates are less biased when the number of subjects is 50 instead of 20. The variance of the estimates is also decreased with larger values of N. In general the empirical variances coincide with the values of the asymptotic variances. Apparently, increasing T does not have any significant impact on the properties

of the estimates. Additional simulations with other values for μ_0 and ω_0^2 have basically shown that the properties of the estimates degrade when ω_0^2 takes bigger values. We have also considered simulations of the mixed-effects geometric Brownian motion where $b(x) = x$, $\sigma(x) = \sigma\, x$ and the model where $b(x) = \sqrt{1+x^2}$, $\sigma(x) = \sigma\sqrt{1+x^2}$. On our simulated data, the true parameter values were correctly estimated.

3.3.7.2 General case

***Example* 3.4.** Consider an Ornstein-Uhlenbeck process with one random effect

$$dX_i(t) \quad = \quad \phi_i X_i(t)dt + \sigma\, dW_i(t), \quad X_i(0) = 0,$$

with $\phi_i \sim \mathcal{N}(\mu, \omega^2)$. We separate three situations: *i)* ω^2 is known ($\theta = \mu$), *ii)* μ is known ($\theta = \omega^2$), *iii)* both parameters are unknown ($\theta = (\mu, \omega^2)$). When ω_0^2 is known, we use the explicit expression of $\widehat{\mu}_N$. Otherwise, numerical optimization procedures are required for maximizing the log-likelihood with respect to μ and ω^2. In situations *i)* and *ii)*, the asymptotic variance of the estimate has an explicit expression and is computed. Each individual diffusion is simulated with a discretization step-size $\delta = 0.001$ on $[0, T]$ and $\sigma = 1$. Several sets of parameter values are used: ($\mu = -5,\ \omega^2 = 1$) and ($\mu = 10,\ \omega^2 = 1$). Table 3.4 displays the results of the three inferences *i)*, *ii)* and *iii)*. Results highlight the accuracy of the estimates of both parameters μ and ω^2 whatever the design and the parameter values. In the three considered inference situations, increasing N leads to smaller bias of the parameter estimates and smaller variances. Moreover, we notice great similarities between $N^{-1}\mathcal{I}_N^{-1}(\theta_0)$ and the empirical variance of $\widehat{\theta}_N$, especially when N is large. Finally, we don't observe any significant impact of T neither on the bias nor on the variance of the parameter estimates.

***Example* 3.5.** Consider an Ornstein-Uhlenbeck process with two random effects

$$dX_i(t) \quad = \quad (-\phi_i^1 X_i(t) + \phi_i^2)dt + \sigma\, dW_i(t), \quad X_i(0) = 0,$$

with $\phi_i = (\phi_i^1, \phi_i^2)' \sim \mathcal{N}_2(\mu, \Omega)$, $\mu = (\mu_1, \mu_2)'$ and a diagonal matrix Ω with components (ω_1^2, ω_2^2). For this model, Assumption 3.8 is satisfied. Indeed

$$V_i = \frac{1}{\sigma^2}\begin{pmatrix} \int_0^T X_i^2(s)ds & \int_0^T X_i(s)ds \\ \int_0^T X_i(s)ds & T \end{pmatrix}.$$

Using the equality case in the Cauchy-Schwarz inequality, we see that $\det(V_i) = 0$ if and only if $X_i(t) \equiv cste$ on $[0, T]$ which is impossible. The estimation of $\theta = (\mu_1, \mu_2, \omega_1^2, \omega_2^2)$ is obtained by optimizing numerically the log-likelihood. Each individual diffusion is simulated with a discretization step-size $\delta = 0.001$ on $[0, T]$ and $\sigma = 1$. Several sets of parameter values are used: ($\mu_1 = 0.1, \mu_2 = 1, \omega_1^2 = 0.01, \omega_2^2 = 1$) and ($\mu_1 = 0.1, \mu_2 = 1, \omega_1^2 = 0.001, \omega_2^2 = 1$). Table 3.5 displays the results of estimation. Parameters are well estimated although μ_2 has a larger bias when ω_2^2 is greater. Biases decrease when N increases. Again, the influence of T is small.

***Example* 3.6.** Consider the process with single random effect

$$dX_i(t) \quad = \quad \phi_i X_i(t)dt + \sigma\sqrt{1 + X_i(t)^2}dW_i(t), \quad X_i(0) = 0,$$

with $\phi_i \sim \mathcal{N}(\mu, \omega^2)$. We obtain the estimate for $\theta = (\mu, \omega^2)$ by numerical optimization of the log-likelihood with respect to μ and ω^2. We simulate N individual diffusions with a discretization step-size $\delta = 0.001$ on $[0, T]$, and several sets of parameter values are used: $(\mu = -1, \omega^2 = 1)$ and $(\mu = 5, \omega^2 = 1)$. Table 3.6 displays the results of estimation. The results are satisfactory overall, although parameter ω^2 is estimated with larger bias than parameter μ. Bias and empirical variances of the estimates decrease when N becomes larger. $T = 10$ leads to better results (smaller bias) for the estimation of ω^2 than $T = 5$.

3.3.8 Concluding remarks

In this paper, we have studied maximum likelihood estimation for *i.i.d.* observations of stochastic differential equations including a random effect in the drift term. When the drift term depends linearly on the random effect, we prove that the likelihood is given by a closed-form formula and that the exact MLE is strongly consistent and asymptotically Gaussian as the number of observed processes tends to infinity.

For the clarity of exposure, we have considered only one-dimensional SDEs, but the theory can be done in the same way for multidimensional SDEs. For a drift term depending linearly on the random effects, the likelihood is still explicit although its formulae may be much more cumbersome.

One could also include non random effects in the drift without much changes.

Acknowledgements

We thank Arnaud Gloter for his helpful contribution on Assumption 3.4.

3.3.9 Appendix: proofs

3.3.9.1 Proof of Proposition 3.3.1

Consider the two-dimensional SDE:

$$\begin{aligned} dX_i(t) &= b(X_i(t),\phi_i(t))dt + \sigma(X_i(t))dW_i(t), \quad X_i(0) = x^i, \\ d\phi_i(t) &= 0, \quad \phi_i(0) = \phi_i. \end{aligned}$$

Under Assumption 3.1, the above system admits a unique strong solution and there exists a functional F such that $X_i(.) = F_.(x^i, \phi_i, W_i(.))$ where $F_. : \mathbb{R} \times \mathbb{R}^d \times C(\mathbb{R}^+, \mathbb{R}) \to C(\mathbb{R}^+, \mathbb{R})$ is measurable [see *e.g.* 12, p.310].
Moreover, $X_i^\varphi(.) = F_.(x^i, \varphi, W_i(.))$. By the Markov property of the joint process $((X_i(t), \phi_i(t) \equiv \phi_i), t \geq 0)$, the conditional distribution of $X_i(.)$ given $\phi_i = \varphi$ is identical to the distribution of $X_i^\varphi(.)$. As $(\phi_i, W_i(.))$ are independent, the processes $(X_i(.))$ are independent. As (x^i, ϕ_i) is the initial condition, the moment result follows.

3.3.9.2 Proof of Proposition 3.3.2

Under Assumptions 3.1-3.2, the first part is classical [see *e.g.* 16].
To prove the continuity in φ, two kinds of terms are to be studied. The first is, for a given X_i, the ordinary integral

$$\varphi \to \int_0^{T_i} \frac{b^2(X_i(s),\varphi)}{\sigma^2(X_i(s))} ds. \tag{3.34}$$

Using Assumptions 3.1 to 3.3 and the continuity theorem for ordinary integrals, we obtain easily the continuity of (3.34).
Second, the stochastic integral

$$\varphi \to \int_0^{T_i} \frac{b(X_i(s),\varphi)}{\sigma^2(X_i(s))} dX_i(s) := I_1(\varphi) + I_2(\varphi), \tag{3.35}$$

with

$$I_1(\varphi) = \int_0^{T_i} \frac{b(X_i(s),\varphi)}{\sigma^2(X_i(s))} b(X_i(s),\varphi_0) ds,$$

$$I_2(\varphi) = \int_0^{T_i} \frac{b(X_i(s),\varphi)}{\sigma^2(X_i(s))} \left(dX_i(s) - b(X_i(s),\varphi_0)ds\right).$$

The function $\varphi \to I_1(\varphi)$ is studied using Assumptions 3.1 to 3.3 and the continuity theorem for ordinary integrals again. For $I_2(\varphi)$, using the Burkholder-Davis-Gundy inequality, we get, using Assumptions 3.1 to 3.3:

$$\mathbb{E}_{Q^i}\left(I_2(\varphi) - I_2(\varphi')\right)^{2k} \leq C_k \mathbb{E}_{Q^i}\left[\int_0^{T_i} \frac{\left(b(X_i(s),\varphi) - b(X_i(s),\varphi')\right)^2}{\sigma^2(X_i(s))} ds\right]^k$$

$$\leq C_k |\varphi - \varphi'|^{2k} \mathbb{E}_{Q^i}\left[\int_0^{T_i} c^2 K (1 + |X_i(s)|^{2\gamma})(1 + |X_i(s)|)^2 ds\right]^k$$

$$\leq C(k,\gamma)|\varphi - \varphi'|^{2k} \int_0^{T_i} (1 + \mathbb{E}_{Q^i}(|X_i(s)|^{2(\gamma+1)}) ds.$$

Using (3.18) and choosing $2k > d$, the Kolmogorov criterion [see *e.g.* 26] yields that $\varphi \to I_2(\varphi)$ admits a continuous version Q^i-a.s..
As $X_i \to L_{T_i}(X_i, \varphi)$ is measurable for all φ and $\varphi \to L_{T_i}(X_i, \varphi)$ is continuous for all X_i, the joint measurability can be proved as follows. For $m \in \mathbb{N}$ and $k = (k_1, \ldots, k_d) \in \mathbb{Z}^d$, set $B_{k,m} = \prod_{i=1}^{d} [k_i/2^m, (k_i+1)/2^m[$. These sets are disjoint and for all m, $\mathbb{R}^d = \cup_{k\in\mathbb{Z}^d} B_{k,m}$. Let $\varphi_{k,m} = (k_i/2^m, i = 1, \ldots, d)$ and set:

$$L_m(X_i, \varphi) = \sum_{k\in\mathbb{Z}^d} L_{T_i}(X_i, \varphi_{k,m}) 1_{B_{k,m}}(\varphi).$$

Thus, $L_m(X_i, \varphi)$ is jointly measurable. As $L_{T_i}(X_i, \varphi)$ is continuous w.r.t. φ,

$$L_m(X_i, \varphi) \to_{m\to+\infty} L_{T_i}(X_i, \varphi).$$

Hence, the result.

3.3.9.3 Proof of Proposition 3.3.3

For H a positive measurable function on C_{T_i}, we have:

$$E_{Q^i_\theta}(H(X_i)) = E_{P^i_\theta}(H(X_i)) = E_{P^i_\theta}[E_{P^i_\theta}(H(X_i)|\phi_i)].$$

By Propositions 3.3.1 and 3.3.2, as $L_{T_i}(X_i, \varphi)$ is the density of Q^{x_i,T_i}_φ w.r.t. Q^i, we get:

$$E_{P^i_\theta}(H(X_i)|\phi_i = \varphi) = E_{Q^{x_i,T_i}_\varphi}(H(X_i)) = \mathbb{E}_{Q^i}\left(H(X_i) L_{T_i}(X_i, \varphi)\right).$$

Using the joint measurability of $L_{T_i}(X_i, \varphi)$ w.r.t. (X_i, φ), the Fubini theorem yields:

$$\begin{aligned}\mathbb{E}_{Q^i_\theta}(H(X_i)) &= \int_{\mathbb{R}^d} g(\varphi, \theta) d\nu(\varphi) \mathbb{E}_{Q^i}\left(H(X_i) L_{T_i}(X_i, \varphi)\right) \\ &= \mathbb{E}_{Q^i} H(X_i) \int_{\mathbb{R}^d} g(\varphi, \theta) L_{T_i}(X_i, \varphi) d\nu(\varphi).\end{aligned}$$

Thus, the density of Q^i_θ w.r.t. Q^i is computed as:

$$\frac{dQ^i_\theta}{dQ^i}(X_i) = \int_{\mathbb{R}^d} L_{T_i}(X_i, \varphi) g(\varphi, \theta) d\nu(\varphi) := \lambda_i(X_i, \theta).$$

The formula for the exact likelihood follows.

3.3.9.4 Proof of Proposition 3.3.4

We need compute first the joint density of (ϕ_i, X_i) w.r.t. $d\varphi \otimes dQ^i$:

$$\exp\left(\varphi U_i - \frac{\varphi^2}{2} V_i\right) \times \frac{1}{\omega\sqrt{2\pi}} \exp\left[-\frac{1}{2\omega^2}(\varphi - \mu)^2\right].$$

Developping the exponent yields:

$$E_i = -\frac{1}{2}\left[\varphi^2(V_i + \omega^{-2}) - 2\varphi(U_i + \omega^{-2}\mu)\right] - \frac{1}{2}\omega^{-2}\mu^2. \tag{3.36}$$

Let us set:

$$m_i = \frac{U_i + \omega^{-2}\mu}{V_i + \omega^{-2}} = \frac{\mu + \omega^2 U_i}{1 + \omega^2 V_i}, \quad \sigma_i^2 = (V_i + \omega^{-2})^{-1} = \frac{\omega^2}{1 + \omega^2 V_i}. \tag{3.37}$$

Thus, the conditional distribution of ϕ_i given X_i is the Gaussian law $\mathcal{N}(m_i, \sigma_i^2)$. After some elementary algebra, we get:

$$E_i = -\frac{1}{2\sigma_i^2}(\varphi - m_i)^2 - \frac{1}{2}V_i(1 + \omega^2 V_i)^{-1}(\mu - V_i^{-1}U_i)^2 + \frac{1}{2}V_i^{-1}U_i^2.$$

Thus, the result.

3.3.9.5 Proof of Lemma 3.3.5

For the proof, we set $\gamma_1(\theta) = \gamma_1$ and $I_1(\omega^2) = I_1$ (see (3.26)). Let $l(X_1, \theta) = \log \lambda_1(X_1, \theta)$ and set $\theta(u) = (\mu + u, \omega^2)$. Developping $(U_1 - (\mu + u)V_1)^2$, we get:

$$l(X_1, \theta(u)) = l(X_1, \theta) + u\gamma_1 - \frac{u^2}{2}I_1.$$

Here, $\dfrac{\partial}{\partial\mu}l(X_1, \theta) = \gamma_1$ and $\dfrac{\partial^2}{\partial\mu^2}l(X_1, \theta) = -I_1$. Taking exponential yields:

$$\lambda_1(X_1, \theta)\exp(u\gamma_1) = \lambda_1(X_1, \theta(u))\exp(\frac{u^2}{2}I_1).$$

Integrating w.r.t. the dominating measure Q^1, we obtain, as $I_1 \leq 1/\omega^2$,

$$E_\theta \exp(u\gamma_1) = E_{\theta(u)} \exp(\frac{u^2}{2}I_1) \leq \exp(\frac{u^2}{2\omega^2}) < +\infty.$$

Now, as $u\mu \leq (u + \mu)^2/4$,

$$E_\theta(\exp(u\frac{U_1}{1 + \omega^2 V_1})) \leq E_\theta \exp(u\gamma_1)\exp(\frac{(u + \mu)^2}{4\omega^2}) < +\infty.$$

3.3.9.6 Proof of Proposition 3.3.6

We set $\gamma_1(\theta) = \gamma_1$ and $I_1(\omega^2) = I_1$. Let $\theta = (\mu, \omega^2)$ and $\tau = (0, \omega^2)$ and set

$$p_1(\theta) = \frac{\lambda_1(X_1, \theta)}{\lambda_1(X_1, \tau)} = \frac{dQ_\theta^1}{dQ_\tau^1} = \exp\left(\mu\frac{U_1}{1 + \omega^2 V_1} - \frac{\mu^2}{2}\frac{V_1}{1 + \omega^2 V_1}\right),$$

so that $\int_{C_T} p_1(\theta)dQ_\tau^1 = 1$. Provided that we can interchange derivation w.r.t. μ and integration w.r.t. Q_τ^1, we have for $j \geq 1$,

$$\int_{C_T} \frac{\partial^j p_1}{\partial\mu^j}(\theta)dQ_\tau^1 = 0. \tag{3.38}$$

Before justifying the interchange of derivation and integration, let us compute the successive derivatives of $p_1(\theta)$. We have:

$$\frac{\partial p_1}{\partial\mu}(\theta) = \gamma_1 p_1(\theta), \quad \frac{\partial^2 p_1}{\partial\mu^2}(\theta) = \left(\gamma_1^2 - I_1\right)p_1(\theta), \quad \frac{\partial^3 p_1}{\partial\mu^3}(\theta) = \left(\gamma_1^3 - 3\gamma_1\, I_1\right)p_1(\theta),$$

$$\frac{\partial^4 p_1}{\partial \mu^4}(\theta) = \left(\gamma_1^4 - 6\gamma_1^2\, I_1 + 3I_1^2\right) p_1(\theta).$$

Therefore, (3.38) for $j = 1, 2, 3, 4$ imply the announced moments relations.
It remains to justify the interchange of derivation and integration. Let us fix $\bar{\mu}$ and $\varepsilon > 0$. For $\mu \in [\bar{\mu} - \varepsilon, \bar{\mu} + \varepsilon]$, we have the bound

$$|\frac{\partial p_1}{\partial \mu}(\theta)| \leq \left(|\frac{U_1}{1+\omega^2 V_1}| + \frac{C}{\omega^2}\right)\left(\exp\left((\bar{\mu}-\varepsilon)\frac{U_1}{1+\omega^2 V_1}\right) + \exp\left((\bar{\mu}+\varepsilon)\frac{U_1}{1+\omega^2 V_1}\right)\right),$$

where $C = |\bar{\mu} + \varepsilon| + |\bar{\mu} - \varepsilon|$. The upper bound is integrable w.r.t. Q_τ^1 by Lemma 3.3.5 and independent of μ. Therefore, the interchange is justified. We proceed analogously for the other derivatives.

3.3.9.7 Proof of Proposition 3.3.7

We set $\gamma_1(\theta) = \gamma_1$ and $I_1(\omega^2) = I_1$. Let us compute the second derivatives of the loglikelihood. Using that

$$\partial\gamma_i/\partial\mu = -I_i, \qquad \partial\gamma_i/\partial\omega^2 = -\gamma_i I_i, \qquad \partial I_i/\partial\omega^2 = -I_i^2(\omega^2),$$

we obtain

$$\frac{\partial^2}{\partial\mu^2}\mathcal{L}_N(\theta) = -\sum_{i=1}^N I_i(\omega^2), \qquad \frac{\partial^2}{\partial\mu\partial\omega^2}\mathcal{L}_N(\theta) = -\sum_{i=1}^N \gamma_i(\theta) I_i(\omega^2), \tag{3.39}$$

$$\frac{\partial^2}{\partial\omega^2\partial\omega^2}\mathcal{L}_N(\theta) = -\frac{1}{2}\sum_{i=1}^N \left(2\gamma_i^2(\theta) I_i(\omega^2) - I_i^2(\omega^2)\right). \tag{3.40}$$

We use the simple law of large numbers, the standard central limit theorem and Proposition 3.3.6 to conclude.

3.3.9.8 Proof of Proposition 3.3.9

First note that $\lambda_1(X_1, \theta) = \lambda_1(U_1, V_1, \theta)$ depends on X_1 only through (U_1, V_1). Hence, under Q_θ^1, by Assumption 3.4, the couple (U_1, V_1) admits a density w.r.t. the Lebesgue measure equal to $f_\theta(u, v) = \lambda_1(u, v, \theta) f(u, v)$. Assuming that $Q_\theta^1 = Q_{\theta_0}^1$ implies that $f_\theta(u, v) = f_{\theta_0}(u, v)$ *a.e.*, by the continuity of the functions $f_\theta(u, v), f_{\theta_0}(u, v)$, the equality holds everywhere. As $f(u, v)$ is positive on an open ball B of $\mathbb{R} \times (0, +\infty)$, we deduce that, on the ball B, the following equality holds:

$$\left(\frac{1+\omega_0^2 v}{1+\omega^2 v}\right)^{1/2} = \exp\left[-\frac{v}{2(1+\omega_0^2 v)}\left(\mu_0 - \frac{u}{v}\right)^2 + \frac{v}{2(1+\omega^2 v)}\left(\mu - \frac{u}{v}\right)^2\right].$$

The left-hand side is a function of v only while the right-hand side is a function of (u, v). This is only possible if $\omega = \omega_0$ and $\mu = \mu_0$. As $K(Q_{\theta_0}^1, Q_\theta^1) \geq 0$ and $= 0$ if and only if $Q_{\theta_0}^1 = Q_\theta^1$, we get (i).
Let $\mathcal{L}_1(\theta) = \log \lambda_1(X_1, \theta)$ (see (3.3.4)-(3.23)). We have

$$K(Q_{\theta_0}^1, Q_\theta^1) = E_{\theta_0}(\mathcal{L}_1(\theta_0) - \mathcal{L}_1(\theta)).$$

Rearranging terms, we get:

$$\begin{aligned}\mathcal{L}_1(\theta_0) - \mathcal{L}_1(\theta) &= \frac{1}{2}\log\left(\frac{1+\omega^2 V_1}{1+\omega_0^2 V_1}\right) + \frac{1}{2}\frac{(\omega_0^2-\omega^2)U_1^2}{(1+\omega^2 V_1)(1+\omega_0^2 V_1)} \\ &\quad + \frac{\mu^2 V_1}{2(1+\omega^2 V_1)} - \frac{\mu U_1}{1+\omega^2 V_1} - \left(\frac{\mu_0^2 V_1}{2(1+\omega_0^2 V_1)} - \frac{\mu_0 U_1}{1+\omega_0^2 V_1}\right).\end{aligned}$$

Let us prove that this r.v. has finite expectation under E_{θ_0}. We have the upper bound:

$$0 < \frac{1+\omega^2 V_1}{1+\omega_0^2 V_1} < 1 + \frac{\omega^2}{\omega_0^2}.$$

Introducing the function $h(x) = x - 1 - \log x$, which is defined on $(0, +\infty)$ and non-negative, we get the lower bound:

$$\log\left(\frac{1+\omega^2 V_1}{1+\omega_0^2 V_1}\right) = h\left(\frac{1+\omega_0^2 V_1}{1+\omega^2 V_1}\right) + (\omega^2-\omega_0^2)\frac{V_1}{1+\omega^2 V_1} \geq (\omega^2-\omega_0^2)\frac{V_1}{1+\omega^2 V_1}.$$

Thus,

$$|\log\left(\frac{1+\omega^2 V_1}{1+\omega_0^2 V_1}\right)| \leq \log(1+\frac{\omega^2}{\omega_0^2}) + \frac{|\omega^2-\omega_0^2|}{\omega^2}. \tag{3.41}$$

For the second term, we write:

$$0 < \frac{U_1^2}{(1+\omega^2 V_1)(1+\omega_0^2 V_1)} = \left(\frac{U_1}{1+\omega_0^2 V_1}\right)^2 \frac{1+\omega_0^2 V_1}{1+\omega^2 V_1} \leq \left(\frac{U_1}{1+\omega_0^2 V_1}\right)^2 (1+\frac{\omega_0^2}{\omega^2}) \tag{3.42}$$

which has finite E_{θ_0}-expectation by Lemma 3.3.5. For the last terms, we only need to check that $E_{\theta_0}|U_1/(1+\omega^2 V_1)| < +\infty$. For this, we remark that:

$$\frac{U_1}{1+\omega^2 V_1} = \frac{U_1}{1+\omega_0^2 V_1}\left(1 + (\omega_0^2-\omega^2)\frac{V_1}{1+\omega^2 V_1}\right).$$

Therefore,

$$|\frac{U_1}{1+\omega^2 V_1}| \leq |\frac{U_1}{1+\omega_0^2 V_1}|\left(1 + \frac{|\omega_0^2-\omega^2|}{\omega^2}\right). \tag{3.43}$$

By Lemma 3.3.5, the right-hand side has finite E_{θ_0}-expectation. The function $\theta \to \mathcal{L}_1(\theta_0) - \mathcal{L}_1(\theta)$ is continuous. For all $\theta = (\mu, \omega^2) \in [\underline{\mu}, \overline{\mu}] \times [\underline{\omega}^2, \overline{\omega}^2] \subset \mathbb{R} \times (0, +\infty)$, using inequalities (3.41)-(3.42)-(3.43), we can easily obtain an upper bound for $|\mathcal{L}_1(\theta_0) - \mathcal{L}_1(\theta)|$ which has finite E_{θ_0}-expectation and is uniform on the interval $[\underline{\mu}, \overline{\mu}] \times [\underline{\omega}^2, \overline{\omega}^2]$. The continuity of the Kullback information follows. This gives (ii).

Remark 12. It is worth noting that, although we have an explicit expression of $K(Q^1_{\theta_0}, Q^1_\theta)$, we cannot prove directly, using this expression, that $K(Q^1_{\theta_0}, Q^1_\theta) = 0$ implies $\theta = \theta_0$.

3.3.9.9 Proof of Proposition 3.3.10

We first prove 1. As $(1/N)(\mathcal{L}_N(\theta_0) - \mathcal{L}_N(\theta)$ converges to $K(Q^1_{\theta_0}, Q^1_\theta)$ in Q_{θ_0}-probability, the loglikelihood $-(1/N)\mathcal{L}_N(\theta)$ is a contrast process with contrast function $\theta \to K(Q^1_{\theta_0}, Q^1_\theta)$. Following the usual standard proof of consistency of minimum contrats estimators [see *e.g.* 28], it remains to study the continuity modulus of $-(1/N)\mathcal{L}_N(\theta)$ defined by:

$$w_N(\eta) = \sup_{\|\theta-\theta'\|\leq\eta,\theta,\theta'\in\Theta} |\mathcal{L}_N(\theta) - \mathcal{L}_N(\theta')|/N.$$

We simply use $w_N(\eta) \leq \eta \ \sup_{\theta\in\Theta} \|\nabla\mathcal{L}_N(\theta)/N\|$ and bound the score function (3.27). By Assumption 3.5, we have $\Theta \subset [\underline{\mu}, \overline{\mu}] \times [\underline{\omega}^2, \overline{\omega}^2]$ with $\underline{\mu} < \overline{\mu}, 0 < \underline{\omega}^2 < \overline{\omega}^2$. We have:

$$\gamma_i(\theta) = \frac{U_i}{1+\omega_0^2 V_i}\left(1 + \frac{(\omega_0^2-\omega^2)V_i}{1+\omega^2 V_i}\right) - \mu\frac{V_i}{1+\omega^2 V_i}.$$

Thus,

$$\sup_{\theta\in\Theta} |\gamma_i(\theta)| \leq |\frac{U_i}{1+\omega_0^2 V_i}| \left(2 + \frac{\omega_0^2}{\underline{\omega}^2}\right) + \frac{|\overline{\mu}|}{\underline{\omega}^2}. \tag{3.44}$$

Therefore, there is a constant C such that

$$E_{\theta_0} w_N(\eta) \leq C\ \eta\ E_{\theta_0}\left(|\frac{U_1}{1+\omega_0^2 V_1}| + (\frac{U_1}{1+\omega_0^2 V_1})^2\right).$$

This leads to the consistency of $\hat{\theta}_N$.

For the second point, the proof follows the standard scheme. By the consistency and Assumption 3.5, $Q_{\theta_0}(\hat{\theta}_N \in \mathring{\Theta}) \to 1$. For the proof, let $\hat{\theta}_{N,i}, \theta_{0,i}$ be the components of $\hat{\theta}_N, \theta_0$. Set $U_N(\theta) = -(1/N)\mathcal{L}_N(\theta)$ and denote by $U'_{N,i}, U''_{N,ij}$ the derivatives of U_N w.r.t. θ_i or $\theta_i\theta_j$. The Taylor formula writes:

$$0 = U'_{N,i}(\hat{\theta}_N) = U'_{N,i}(\theta_0) + \sum_{j=1,2} (\hat{\theta}_{N,j} - \theta_{0,j})(U''_{N,ij}(\theta_0) + R_N),$$

with

$$R_N = \int_0^1 \left(U''_{N,ij}(\theta_0 + s(\hat{\theta}_N - \theta_0)) - U''_{N,ij}(\theta_0)\right) ds.$$

Using Propositions 3.3.10 and 3.3.7, it remains to check that R_N tends in Q_{θ_0}-probability to 0. For this, we compute the third derivatives of U_N using (3.39)-(3.40):

$$\frac{1}{N}\frac{\partial^3}{\partial\mu^3}\mathcal{L}_N(\theta) = 0, \quad \frac{1}{N}\frac{\partial^3}{\partial\omega^2\partial\omega^2\partial\omega^2}\mathcal{L}_N(\theta) = \frac{1}{N}\sum_{i=1}^N \left[3\gamma_i^2(\theta) I_i^2(\omega^2) - I_i^3(\omega^2)\right],$$

$$\frac{1}{N}\frac{\partial^3}{\partial^2\mu\partial\omega^2}\mathcal{L}_N(\theta) = \frac{1}{N}\sum_{i=1}^N I_i^2(\omega^2), \quad \frac{1}{N}\frac{\partial^3}{\partial\mu\partial\omega^2\partial\omega^2}\mathcal{L}_N(\theta) = \frac{2}{N}\sum_{i=1}^N \gamma_i(\theta) I_i^2(\omega^2).$$

Using (3.44), we obtain, for C a constant depending on Θ

$$|R_N| \leq C\ |\hat{\theta}_N - \theta_0| \frac{1}{N}\sum_{i=1}^N (1 + \left(|\frac{U_i}{1+\omega_0^2 V_i}| + (\frac{U_i}{1+\omega_0^2 V_i})^2\right).$$

As we have proved the consistency, R_N tends to 0. Hence the result.

3.3.9.10 Proof of Lemma 3.3.11

The matrix $V_i + \Omega^{-1}$ is symmetric and satisfies, for all $x \in \mathbb{R}^d$, $x \neq 0$, $x'(V_i + \Omega^{-1})x = x'V_i x + x'\Omega^{-1}x > 0$ as V_i and Ω^{-1} are positive definite. Thus $V_i + \Omega^{-1}$ is positive definite hence invertible. Noting that $I_d + V_i\Omega = (\Omega^{-1} + V_i)\Omega$ and $I_d + \Omega V_i = \Omega(\Omega^{-1} + V_i)$ yields that both matrices are invertible.

3.3.9.11 Proof of Proposition 3.3.12

We compute the joint density of (ϕ_i, X_i) w.r.t. $d\varphi \otimes dQ^i$:

$$\exp\left(\varphi' U_i - \frac{1}{2}\varphi' V_i \varphi\right) \times \exp\left[-\frac{1}{2}(\varphi - \mu)'\Omega^{-1}(\varphi - \mu)\right].$$

Let

$$E_i = -\frac{1}{2}\left[\varphi'(V_i + \Omega^{-1})\varphi - 2\varphi'(U_i + \Omega^{-1}\mu)\right] - \frac{1}{2}\mu'\Omega^{-1}\mu, \tag{3.45}$$

and set:

$$m_i = \Sigma_i(U_i + \Omega^{-1}\mu), \quad \Sigma_i^2 = (V_i + \Omega^{-1})^{-1}. \tag{3.46}$$

Thus, the conditional distribution of ϕ_i given X_i is the Gaussian law $\mathcal{N}(m_i, \Sigma_i^2)$. Hence, we have

$$E_i = -\frac{1}{2}(\varphi - m_i)'\Sigma_i^{-1}(\varphi - m_i) - \frac{1}{2}(\mu - V_i^{-1}U_i)'R_i^{-1}(\mu - V_i^{-1}U_i) + \frac{1}{2}U_i'V_i^{-1}U_i.$$

Thus, the result.

3.3.9.12 Proof of Proposition 3.3.13

For the first point, we proceed as in Lemma 3.3.5. Let $\gamma_1(\theta) = \gamma_1$, $I_1(\Omega) = I_1$, $l(X_1, \theta) = \log \lambda_1(X_1, \theta)$ and $\theta(u) = (\mu + u, \Omega)$. Developping $((\mu + u) - V_1^{-1}U_1)'R_1^{-1}((\mu + u) - V_1^{-1}U_1)$ yields:

$$l(X_1, \theta(u)) = l(X_1, \theta) + u'\gamma_1 - \frac{1}{2}u'I_1 u.$$

Thus

$$\lambda_1(X_1, \theta)\exp(u'\gamma_1) = \lambda_1(X_1, \theta(u))\exp(\frac{1}{2}u'I_1 u).$$

Remark that $u'I_1 u \leq u'\Omega^{-1}u$ as $I_1 = \Omega^{-1} - (\Omega V_1 \Omega + \Omega)^{-1}$. Integrating w.r.t. Q^1, we obtain,

$$E_\theta \exp(u'\gamma_1) = E_{\theta(u)} \exp(\frac{1}{2}u'I_1 u) \leq \exp(\frac{1}{2}u'\Omega^{-1}u) < +\infty.$$

Now, we can write, as $u'I_1\mu = \frac{1}{4}((u+\mu)'I_1(u+\mu) - (u-\mu)'I_1(u-\mu)) \leq \frac{1}{4}(u+\mu)'I_1(u+\mu)$:

$$E_\theta(\exp(u'(I_d + \Omega V_i)^{-1}U_1)) \leq E_\theta \exp(u'\gamma_1)\exp\left(\frac{1}{4}(u+\mu)'\Omega^{-1}(u+\mu)\right) < +\infty.$$

This gives 1. The proof of the second point and third points is analogous to Proposition 3.3.6 and Proposition 3.3.7.

3.3.9.13 Proof of Lemma 3.3.14

We only treat the case $p \geq 2$. Assumptions imply that b^2/σ^2 is Lipschitz say with constant L. Using the Hölder inequality twice, we get

$$\mathbb{E}_{\theta_0}|V_i - V_i^{(n)}|^p \leq L^p\, T^{p-1} \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{\theta_0}|X_i(s) - X_i(t_k)|^p ds.$$

Now, for $0 \leq t \leq t+h \leq T$, we have

$$X_i(t+h) - X_i(t) = \int_t^{t+h} \phi_i b(X_i(s))ds + \int_t^{t+h} \sigma(X_i(s))dW_i(s).$$

Using $\phi_i b(X_i(s)) \leq \phi_i^2/2 + b^2(X_i(s))/2$ and the Hölder inequality, we get

$$|X_i(t+h) - X_i(t)|^p \leq C\left(\phi_i^{2p} h^p + \int_t^{t+h} b^{2p}(X_i(s))ds\; h^{p-1} + |\int_t^{t+h} \sigma(X_i(s))dW_i(s)|^p\right).$$

We use that b and σ have linear growth, the Burkholder-Davis-Gundy inequality and $M_k = \sup_{s\in[0,T]} \mathbb{E}_{\theta_0}|X_i(s)|^k < \infty$ for all k, then

$$\mathbb{E}_{\theta_0}|X_i(t+h) - X_i(t)|^p \leq C\left(h^p \mathbb{E}_{\theta_0}\phi_i^{2p} + h^p\,(1 + M_{2p}) + h^{p/2}(1 + M_p)\right) \leq C h^{p/2}.$$

Thus we obtain

$$\mathbb{E}_{\theta_0}|V_i - V_i^{(n)}|^p \leq \frac{C}{n^{p/2}}.$$

For the difference $U_i - U_i^n$, we have

$$U_i - U_i^n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \left(\frac{b}{\sigma^2}(X_i(s)) - \frac{b}{\sigma^2}(X_i(t_k))\right) dX_i(s) = A_1 + A_2$$

where A_1 is a term analogous to the one already studied above and

$$A_2 = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \left(\frac{b}{\sigma^2}(X_i(s)) - \frac{b}{\sigma^2}(X_i(t_k))\right) \sigma(X_i(s))dW_i(s).$$

We introduce the process

$$H_s^{(n)} = \sum_{k=0}^{n-1} 1_{]t_k, t_{k+1}]}(s) \left(\frac{b}{\sigma^2}(X_i(s)) - \frac{b}{\sigma^2}(X_i(t_k))\right) \sigma(X_i(s))$$

so that $A_2 = \int_0^T H_s^{(n)} dW_i(s)$. We treat $\mathbb{E}_{\theta_0}|A_2|^p$ using the Burkholder-Davis-Gundy inequality and similar tools as above.

3.3.10 Tables

True value	$N = 20$, $T = 5$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$	$N = 50$, $T = 5$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$	$N = 50$, $T = 10$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$
$\mu = -1$	-0.97 (0.06)	0.06	-0.99 (0.02)	0.02	-0.99 (0.03)	0.02
$\omega^2 = 1$	0.91 (0.16)	0.14	0.99 (0.06)	0.06	0.97 (0.06)	0.05
$\mu = 5$	5.01 (0.06)	0.06	5.00 (0.03)	0.02	5.01 (0.02)	0.02
$\omega^2 = 1$	0.99 (0.16)	0.14	0.97 (0.05)	0.06	0.97 (0.05)	0.05

Table 3.3: Example 3.3: Mixed Brownian motion with drift. Empirical mean and variance (in brackets) of $\widehat{\mu}_N$ and $\widehat{\omega}_N^2$ computed from 100 datasets for three designs and two sets of parameters (μ, ω^2). The exact asymptotic variance diag $\left(N^{-1}\mathcal{I}_N^{-1}(\theta_0)\right)$ is also computed.

True value		$N = 20$, $T = 5$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$	$N = 50$, $T = 5$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$	$N = 50$, $T = 10$ Mean (Var)	$\frac{1}{N}\mathcal{I}_N^{-1}$
μ estimated, ω^2 fixed							
$\mu = -5$, $\omega^2 = 1$	$\widehat{\mu}_N$	-5.03 (0.14)	0.11	-5.01 (0.06)	0.05	-4.99 (0.04)	0.02
$\mu = 10$, $\omega^2 = 1$	$\widehat{\mu}_N$	9.90 (0.05)	0.05	9.91 (0.01)	0.02	9.90 (0.02)	0.02
μ fixed, ω^2 estimated							
$\mu = -5$, $\omega^2 = 1$	$\widehat{\omega}_N^2$	0.96 (0.31)	0.62	0.96 (0.14)	1.49	0.93 (0.16)	0.27
$\mu = 10$, $\omega^2 = 1$	$\widehat{\omega}_N^2$	0.99 (0.08)	0.20	1.00 (0.04)	0.04	0.97(0.03)	0.05
μ and ω^2 estimated							
$\mu = -5$, $\omega^2 = 1$	$\widehat{\mu}_N$	-4.95 (0.22)	-	-4.99 (0.07)	-	-4.96 (0.03)	-
	$\widehat{\omega}_N^2$	0.74 (1.00)	-	0.91 (0.34)	-	0.99 (0.21)	-
$\mu = 10$, $\omega^2 = 1$	$\widehat{\mu}_N$	9.85 (0.05)	-	9.85 (0.02)	-	9.84 (0.01)	-
	$\widehat{\omega}_N^2$	0.94 (0.08)	-	0.95 (0.04)	-	0.94 (0.05)	-

Table 3.4: Example 3.4: Mixed Ornstein-Uhlenbeck model with one random effect. Estimation of μ when ω_0^2 is known, of ω^2 when μ_0 is known and simultaneous estimation of μ and ω^2, for different values of (N, T) and different parameter values. Empirical mean, variance (in brackets) and estimated value of the asymptotic variance diag $\left(N^{-1}\mathcal{I}_N^{-1}(\theta_0)\right)$ are computed from 100 repeated simulated datasets.

True parameter values	$N = 20$, $T = 5$ Mean (Var)	$N = 50$, $T = 5$ Mean (Var)	$N = 50$, $T = 10$ Mean (Var)
$\mu_1 = 0.1$	0.095 (0.082)	0.102 (0.059)	0.102 (0.028)
$\mu_2 = 1$	1.007 (0.275)	1.020 (0.193)	0.984 (0.175)
$\omega_1^2 = 0.01$	0.010 (0.014)	0.010 (0.009)	0.010 (0.005)
$\omega_2^2 = 1$	0.919 (0.526)	0.956 (0.342)	1.029 (0.278)
$\mu_1 = 0.1$	0.123 (0.073)	0.104 (0.047)	0.106 (0.019)
$\mu_2 = 1$	1.085 (0.287)	1.010 (0.166)	1.022 (0.171)
$\omega_1^2 = 0.001$	0.002 (0.004)	0.002 (0.004)	0.001 (0.002)
$\omega_2^2 = 1$	1.095 (0.488)	1.005 (0.353)	1.024 (0.253)

Table 3.5: Example 3.5: Mixed Ornstein-Uhlenbeck model with two random effects. Empirical mean, variance (in brackets) are computed from 100 repeated simulated datasets for different values of (N, T) and different parameter values.

True value	Est	$N = 20$, $T = 5$ Mean (Var)	$N = 50$, $T = 5$ Mean (Var)	$N = 50$, $T = 10$ Mean (Var)
$\mu = -1$	$\widehat{\mu}_N$	-1.03 (0.12)	-0.98 (0.04)	-1.02 (0.03)
$\omega^2 = 1$	$\widehat{\omega}_N^2$	0.95 (0.42)	0.94 (0.11)	0.98 (0.09)
$\mu = 5$	$\widehat{\mu}_N$	4.99 (0.05)	5.00 (0.02)	5.04 (0.03)
$\omega^2 = 1$	$\widehat{\omega}_N^2$	0.96 (0.17)	0.97 (0.04)	1.00 (0.04)

Table 3.6: Example 3.6. Empirical mean and variance (in brackets) of the estimates $\widehat{\mu}_N$ and $\widehat{\omega}_N^2$ computed from 100 datasets for different values of (N, T) and different parameter values.

3.4 Résultats complémentaires

Cette section présente des résultats complémentaires à l'étude par simulations proposée dans l'article *Maximum likelihood estimation for stochastic differential equations with random effects* (Section 3.3). Nous nous restreignons dans cette étude annexe à des processus dont le drift est une fonction linéaire d'effets aléatoires Gaussiens unidimensionnels.

3.4.1 Propriétés asymptotiques de $\widehat{\theta}_N$ lorsque $b(x) = c\sigma(x)$: compléments

Pour illustrer les propriétés du maximum de vraisemblance dans les modèles vérifiant $b(x) = c\sigma(x)$, nous nous sommes limités au mouvement Brownien avec drift à effets aléatoires. Nous complétons l'étude par simulations proposée dans l'article en considérant d'autres modèles.

***Exemple* 3.7.** Considérons un mouvement Brownien géométrique à effets aléatoires ($b(x) = x$,$\sigma(x) = \sigma x$) :

$$dX_i(t) \quad = \quad \phi_i X_i(t)dt + \sigma X_i(t)dW_i(t), \quad X_i(0) = 0,$$

où $\phi_i \sim \mathcal{N}(\mu, \omega^2)$.
L'inférence de μ et ω^2 peut se traiter de la même manière que dans le mouvement Brownien avec drift à effets aléatoires, en utilisant les formules explicites des extimateurs :

$$\widehat{\mu}_N = \frac{1}{c^2TN}\sum_{i=1}^{N} U_i = \frac{1}{c^2T}\bar{U}_N, \quad \widehat{\omega}^2_N = \frac{1}{(c^2T)^2}\left(\frac{1}{N}\sum_{i=1}^{N}(U_i - \bar{U}_N)^2 - c^2T\right). \tag{3.47}$$

Dans ce modèle, les statistiques U_i et V_i s'écrivent :

$$U_i = \frac{1}{\sigma^2}\int_0^T \frac{dX_i(s)}{X_i(s)}, \quad V_i = \frac{1}{\sigma^2}T.$$

On peut également se ramener à un mouvement Brownien avec drift à un effet aléatoire en posant $Y_i(t) = \log X_i(t)$:

$$dY_i(t) \quad = \quad \varphi_i dt + \sigma\, dW_i(t),$$

où $\varphi_i \sim \mathcal{N}(\mu - \frac{1}{2}\sigma^2, \omega^2)$. Le mouvement Brownien avec drift à effets aléatoires ayant déjà fait l'objet de simulations, nous ne présentons pas de simulations pour le mouvement Brownien géométrique à effets aléatoires.

***Exemple* 3.8.** Considérons le modèle suivant :

$$dX_i(t) \quad = \quad \phi_i\sqrt{1 + X_i(t)^2}dt + \sqrt{1 + X_i(t)^2}dW_i(t), \quad X_i(0) = 0,$$

où $\phi_i \sim \mathcal{N}(\mu, \omega^2)$. Dans ce modèle, $b(x) = \sigma(x) = \sqrt{1 + x^2}$. Les paramètres à estimer sont $\theta = (\mu, \omega^2)$. Contrairement au mouvement Brownien géométrique à effets aléatoires, il n'est pas possible ici de se ramener à un problème d'estimation de paramètres dans un mouvement Brownien à effets aléatoires. Nous utilisons trois designs expérimentaux : ($N = 20, T = 5$), ($N = 50, T = 5$) et ($N = 50, T = 10$), et différents jeux de paramètres : ($\mu = -1, \omega^2 = 1$) et

$(\mu = 5, \omega^2 = 1)$ pour les simulations. Pour chaque design et chaque jeu de paramètres, nous simulons 100 jeux de données constitués chacun de N sujets sur le même intervalle de temps $[0, T]$ ($T_i = T$ pour $i = 1, \ldots, N$), avec un pas de discrétisation $\delta = 0.001$. Les conditions de la Proposition 3.3.15 sont alors vérifiées. Nous calculons pour chaque jeu de données simulé l'estimateur du maximum de vraisemblance de θ selon les formules rappelées en (3.47), et en discrétisant les statistiques U_i. La moyenne et la variance empiriques des 100 estimateurs sont calculés ; les résultats sont présentés dans le Tableau 3.7. Pour l'ensemble des designs simulés, les paramètres sont correctement estimés. On remarque néanmoins un biais plus important pour $\hat{\omega}_N^2$ que pour $\hat{\mu}_N$, essentiellement lorsque la taille de l'échantillon est petite ($N = 20$). On remarque également que la variance des estimateurs diminue lorsque la taille de l'échantillon augmente. Augmenter le temps d'observation T réduit sensiblement la variance de $\hat{\omega}_N^2$.

3.4.2 Influence des paramètres μ_0, ω_0^2 et T sur les propriétés des estimateurs

Nous nous sommes intéressés à l'influence des "vraies valeurs" des paramètres μ et ω^2 sur les propriétés asymptotiques de leurs estimateurs respectifs $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$, ainsi qu'à l'impact du temps d'observation T sur les performances des estimateurs du maximum de vraisemblance dans les modèles étudiés dans l'article.

S'il paraît raisonnable de penser que la valeur de μ_0 a peu d'impact sur les propriétés asymptotiques de $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$, la variance des effets aléatoires $\phi_1, \ldots, \phi_N$ se répercute sur la variance asymptotique des deux estimateurs. Selon l'intuition, la variance asymptotique des estimateurs du maximum de vraisemblance est d'autant plus grande que la valeur de ω_0^2 est grande. De même, lorsque la durée d'observation T des trajectoires des processus X_i, $i = 1, \ldots, N$, est grande, $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$ devraient fournir une estimation plus précise de μ_0 et ω_0^2.

Dans les modèles tels que $b(x) = c\sigma(x)$, la loi des estimateurs est explicite :

$$\begin{aligned} \sqrt{N}(\hat{\mu}_N - \mu_0) &\to_{\mathcal{D}} \mathcal{N}(0, \omega_0^2 + 1/(c^2 T)), \\ \sqrt{N}(\hat{\omega}_N^2 - \omega_0^2) &\to_{\mathcal{D}} \mathcal{N}(0, 2(\omega_0^2 + 1/(c^2 T))^2), \end{aligned}$$

montrant que conformément à l'intuition, plus la variance des effets aléatoires ω_0^2 est grande, plus la variance des estimateurs est grande, alors qu'au contraire, plus le temps d'observation T est important, plus la variance des estimateurs du maximum de vraisemblance est faible, μ_0 n'ayant aucun impact.

La loi asymptotique de $(\widehat{\mu}_N, \widehat{\omega}_N^2)$ fait l'objet de la Proposition 3.3.10, mais du fait de l'expression compliquée de la matrice d'information de Fisher en fonction des espérances des statistiques U_i et V_i, $i = 1, \ldots, N$, inconnues dans la plupart des modèles, nous ne pouvons pas clairement établir les rôles respectifs de μ_0, ω_0^2 et T dans des modèles plus généraux ($b(x) \neq c\sigma(x)$). Nous proposons d'apporter une première réponse à travers une étude par simulations Monte-Carlo. Nous considérons donc des modèles de la forme :

$$dX_i(t) = \phi_i b(X_i(t))dt + \sigma(X_i(t))dW_i(t)\ ,\ X_i(0) = x_i\ ,\ i = 1, \ldots, N,$$

$\phi_i \sim \mathcal{N}(\mu, \omega^2)$, avec μ et ω^2 inconnus. Différentes fonctions $b(\cdot)$ et $\sigma(\cdot)$ sont considérées dans cette étude : *i)* $b(x) = x$, $\sigma(x) = \sigma$ (Ornstein-Uhlenbeck) et *ii)* $b(x) = x$, $\sigma(x) = \sqrt{1 + x^2}$.

Dans chacun des modèles *i)* et *ii)*, nous faisons varier les valeurs de μ ($\mu = (-1, -0.5, 1)$) et ω^2 ($\omega^2 = (0.5, 1, 2)$) et simulons pour chaque jeu de paramètres 100 jeux de données de $N = 50$ sujets sur $[0, T]$, avec $T = 5$ puis $T = 10$, par un schéma d'Euler avec un pas de discrétisation $\delta = 0.001$. Les conditions de la Proposition 3.3.15 sont alors vérifiées, assurant que les estimateurs discrétisés pour μ et ω^2 reflètent les propriétés de $\widehat{\mu}_N$ et $\widehat{\omega}^2_N$. Dans le cas de l'Ornstein-Uhlenbeck à effets aléatoires, le paramètre σ est connu et tel que $\sigma = 1$. Les estimateurs $\widehat{\mu}_N$ et $\widehat{\omega}^2_N$ sont calculés dans chaque jeu de données simulé en résolvant numériquement les équations de vraisemblance rappelées dans la Section 3.3.4, et en discrétisant les statistiques U_i et V_i. Leurs moyenne et variance empiriques sont également calculés et reportés dans les Tableaux 3.8 (modèle *i)*) et 3.9 (modèle *ii)*).

Dans chacun des deux modèles examinés dans cette étude, nous remarquons d'abord que quelles que soient les valeurs de μ et ω^2, la variance des estimateurs $\widehat{\mu}_N$ et $\widehat{\omega}^2_N$ a globalement tendance à diminuer lorsque le temps d'observation passe de $T = 5$ à $T = 10$. Nous observons également une augmentation systématique de la variance des estimateurs $\widehat{\mu}_N$ et $\widehat{\omega}^2_N$ lorsque la valeur de ω^2 augmente. De plus, la valeur de ω^2 s'avère avoir un poids plus fort sur la variance de $\widehat{\omega}^2_N$ que sur la variance de $\widehat{\mu}_N$. Prenons l'exemple de l'Ornstein-Uhlenbeck à effets aléatoires. Pour $\mu = -0.5$ et $T = 5$, lorsque ω^2 passe de 0.5 à 2, la variance de $\widehat{\mu}_N$ est multipliée par 2.6, passant de 0.019 à 0.049, tandis que la variance de $\widehat{\omega}^2_N$ est multipliée par 7.4. Ces observations évoquent les résultats explicités dans les modèles tels que $b(x) = c\sigma(x)$ puisque plus généralement, en augmentant la valeur de ω^2 à μ et T fixés, les estimateurs se comportent comme si la variance de $\widehat{\mu}_N$ était une fonction linéaire de ω^2 et la variance de $\widehat{\omega}^2_N$ était une fonction quadratique de ω^2. Nous obtenons des résultats similaires lorsque $b(x) = x$ et $\sigma(x) = \sqrt{1 + x^2}$. Enfin, contrairement aux modèles simples, où $b(x) = c\sigma(x)$, la valeur de μ influe de manière évidente sur la variance des deux estimateurs, mais son rôle n'est pas très clairement identifié. Dans le modèle *i)* comme dans le modèle *ii)*, à ω^2 fixé, la tendance est plutôt à la diminuation des variances des estimateurs $\widehat{\mu}_N$ et $\widehat{\omega}^2_N$ lorsque la valeur de μ augmente.

3.4.3 Données discrètes

Dans l'ensemble des simulations réalisées jusqu'à présent, en simulant les jeux de données avec un pas de discrétisation très petit ($\delta = 0.001$), nous nous sommes placés dans un contexte de données discrètes très riche, puisque les données d'un même sujet consistaient en plus de 5000 observations, et ces observations étaient très rapprochées dans le temps. Cependant, les jeux de données réelles présentent rarement des designs aussi riches et il est important d'identifier les limites de la méthode d'estimation proposée dans ce travail.

Dans la Section 3.3.6, nous énonçons des conditions permettant d'obtenir des estimateurs satisfaisants pour μ et ω^2 à partir de données discrètes. Nous proposons ici d'illustrer les résultats de la Proposition 3.3.15 à travers une courte étude par simulations réalisée à partir du modèle d'Ornstein-Uhlenbeck à un effet aléatoire pour l'estimation de μ à ω^2 connu. Nous choisissons $\omega^2 = 1$ et $\sigma = 1$ connus, dans l'ensemble de l'étude présentée dans ce paragraphe. Nous choisissons deux valeurs de μ : $\mu = -5$ et $\mu = 1$. Pour chaque valeur de μ, nous simulons 100 jeux de données de $N = 50$ sujets sur $[0, T]$, $T = 5$, de façon exacte avec un même pas de discrétisation δ, puis calculons $\widehat{\mu}^{(n)}_N$. La moyenne et la variance empiriques des 100 estimateurs sont ensuite calculés. Nous répétons l'expérience pour différentes valeurs de δ :

$\delta = (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1)$. Les résultats figurent dans le Tableau 3.10.

Ils montrent que pour des pas de temps très petits, $\delta \leq 0.01$, les estimateurs de μ sont non biaisés et de même variance (la variance asymptotique de l'estimateur). Ces pas de temps correspondent à des designs à plus de 500 observations par sujet et $n/N \geq 10$. Pour $\delta \geq 0.05$ (soit $n/N \leq 2$), les estimateurs $\widehat{\mu}_N$ sont baisés, le biais étant beaucoup plus important pour le cas où $\mu = 1$, que pour le cas où $\mu = -5$, et les variances des estimateurs s'éloignent de la valeur de la variance asymptotique de $\widehat{\mu}_N$. Le biais de $\widehat{\mu}_N$ est d'autant plus important que δ devient grand.

Remarque 13. Dans les modèles à un effet aléatoire tels que $b(x) = c\sigma(x)$, le choix du pas de discrétisation δ n'a aucun impact sur la valeur des estimateurs $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$ et leurs propriétés asymptotiques. En effet, rappelons que dans ce modèle, les estimateurs pour μ et ω^2 sont explicites :

$$\widehat{\mu}_N = \frac{1}{c^2 TN} \sum_{i=1}^{N} U_i = \frac{1}{c^2 T} \bar{U}_N, \quad \widehat{\omega}_N^2 = \frac{1}{(c^2 T)^2} \left(\frac{1}{N} \sum_{i=1}^{N} (U_i - \bar{U}_N)^2 - c^2 T \right),$$

avec

$$U_i = c^2 T \phi_i + c W_i(T), \quad V_i = c^2 T,$$

ainsi que leurs lois. En particulier, nous remarquons que les statistiques U_i et V_i ne dépendent que du temps T. Cela signifie que l'observation des trajectoires complètes des N processus sur $[0, T]$ n'apporte aucune information sur θ qui ne soit pas contenue dans la dernière observation $X_i(T)$, $i = 1, \ldots, N$. Les seules observations en T des N processus suffisent donc à calculer $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$.

A l'exception des modèles concernés par la Remarque 13, il convient donc d'utiliser les estimateurs discrétisés avec beaucoup de prudence. En effet, ils n'auront de bonnes propriétés que si le nombre d'observations par sujet est suffisamment grand. Dans le cas contraire, les estimateurs discrétisés risquent de présenter un biais important, même si le nombre de sujets est grand.

Remarque 14. D'après la Proposition 3.3.15, les propriétés asymptotiques de $\widehat{\theta}_N$ sont préservées par la discrétisation dès lors que le nombre d'observations par sujet n tend vers l'infini, et ce plus rapidement que le nombre de sujets : $n \to +\infty$ et $\frac{n}{N} \to +\infty$. Nous retrouvons les mêmes conditions de vitesses sur n et N que Nie [20] dans son étude théorique de l'estimateur du maximum de vraisemblance dans les modèles à effets mixtes dans l'asymptotique conjointe en le nombre de sujets et le nombre d'observations par sujet.

3.4.4 Tableaux

True value	$N=20,\ T=5$ Mean	(Var) $N^{-1}\mathcal{I}_N^{-1}$	$N=50,\ T=5$ Mean	(Var) $N^{-1}\mathcal{I}_N^{-1}$	$N=50,\ T=10$ Mean	(Var) $N^{-1}\mathcal{I}_N^{-1}$
$\mu=-1$	-0.991	(0.070)	-1.009	(0.026)	-0.982	(0.029)
		0.060		0.024		0.022
$\omega^2=1$	0.950	(0.122)	0.982	(0.055)	0.989	(0.040)
		0.144		0.058		0.048
$\mu=5$	5.019	(0.061)	5.000	(0.023)	4.994	(0.020)
		0.060		0.024		0.022
$\omega^2=1$	0.884	(0.137)	1.025	(0.069)	0.958	(0.041)
		0.144		0.058		0.048

TABLE 3.7 – Exemple 3.8($b(x)=\sigma(x)=\sqrt{1+x^2}$) : Moyenne et (variance) empiriques de $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$ calculées à partir de 100 jeux de données simulés sous trois designs (N,T) et avec deux jeux de paramètres (μ_0,ω_0^2). La valeur exacte de la variance asymptotique de $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$, $\text{diag}\left(N^{-1}\mathcal{I}_N^{-1}(\theta_0)\right)$ est également calculée.

True Value		$N=50,\ T=5$ Mean	(Var)	$N=50,\ T=10$ Mean	(Var)
$(\mu=-1,\quad \omega^2=0.5)$	$\hat{\mu}_N$	-0.996	(0.025)	-1.001	(0.014)
	$\hat{\omega}_N^2$	0.463	(0.028)	0.493	(0.024)
$(\mu=-0.5,\quad \omega^2=0.5)$	$\hat{\mu}_N$	-0.501	(0.019)	-0.489	(0.012)
	$\hat{\omega}_N^2$	0.491	(0.028)	0.491	(0.018)
$(\mu=1,\quad \omega^2=0.5)$	$\hat{\mu}_N$	0.989	(0.010)	0.998	(0.013)
	$\hat{\omega}_N^2$	0.503	(0.014)	0.473	(0.011)
$(\mu=-1,\quad \omega^2=1)$	$\hat{\mu}_N$	-1.004	(0.032)	-1.000	(0.030)
	$\hat{\omega}_N^2$	0.973	(0.080)	0.977	(0.071)
$(\mu=-0.5,\quad \omega^2=1)$	$\hat{\mu}_N$	-0.506	(0.038)	-0.513	(0.024)
	$\hat{\omega}_N^2$	0.983	(0.105)	0.984	(0.050)
$(\mu=1,\quad \omega^2=1)$	$\hat{\mu}_N$	0.987	(0.026)	1.003	(0.019)
	$\hat{\omega}_N^2$	0.974	(0.057)	0.965	(0.040)
$(\mu=-1,\quad \omega^2=2)$	$\hat{\mu}_N$	-0.940	(0.053)	-0.974	(0.047)
	$\hat{\omega}_N^2$	1.886	(0.223)	1.994	(0.198)
$(\mu=-0.5,\quad \omega^2=2)$	$\hat{\mu}_N$	-0.518	(0.049)	-0.509	(0.043)
	$\hat{\omega}_N^2$	1.927	(0.206)	1.925	(0.228)
$(\mu=1,\quad \omega^2=2)$	$\hat{\mu}_N$	0.975	(0.043)	1.007	(0.042)
	$\hat{\omega}_N^2$	1.994	(0.147)	2.001	(0.203)

TABLE 3.8 – Ornstein-Uhlenbeck à un effet aléatoire : $b(x)=x$ et $\sigma(x)=1$. Moyenne et (variance) empiriques de $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$ calculées à partir de 100 jeux de données simulés sous trois designs (N,T) et avec neuf jeux de paramètres (μ_0,ω_0^2).

True Value		$N = 50, T = 5$		$N = 50, T = 10$	
		Mean	(Var)	Mean	(Var)
$(\mu = -1, \quad \omega^2 = 0.5)$	$\hat{\mu}_N$	-0.987	(0.031)	-1.002	(0.020)
	$\hat{\omega}_N^2$	0.497	(0.084)	0.484	(0.043)
$(\mu = -0.5, \quad \omega^2 = 0.5)$	$\hat{\mu}_N$	-0.507	(0.028)	-0.491	(0.016)
	$\hat{\omega}_N^2$	0.546	(0.068)	0.487	(0.030)
$(\mu = 1, \quad \omega^2 = 0.5)$	$\hat{\mu}_N$	0.993	(0.018)	0.987	(0.015)
	$\hat{\omega}_N^2$	0.493	(0.032)	0.483	(0.024)
$(\mu = -1, \quad \omega^2 = 1)$	$\hat{\mu}_N$	-0.975	(0.044)	-1.020	(0.030)
	$\hat{\omega}_N^2$	0.944	(0.111)	0.983	(0.093)
$(\mu = -0.5, \quad \omega^2 = 1)$	$\hat{\mu}_N$	-0.506	(0.033)	-0.500	(0.028)
	$\hat{\omega}_N^2$	0.959	(0.109)	0.996	(0.078)
$(\mu = 1, \quad \omega^2 = 1)$	$\hat{\mu}_N$	0.991	(0.029)	0.960	(0.022)
	$\hat{\omega}_N^2$	0.985	(0.093)	0.985	(0.058)
$(\mu = -1, \quad \omega^2 = 2)$	$\hat{\mu}_N$	-0.955	(0.066)	-0.999	(0.052)
	$\hat{\omega}_N^2$	2.026	(0.435)	1.927	(0.246)
$(\mu = -0.5, \quad \omega^2 = 2)$	$\hat{\mu}_N$	-0.533	(0.065)	-0.479	(0.046)
	$\hat{\omega}_N^2$	1.998	(0.383)	1.967	(0.199)
$(\mu = 1, \quad \omega^2 = 2)$	$\hat{\mu}_N$	0.990	(0.040)	1.029	(0.044)
	$\hat{\omega}_N^2$	1.917	(0.273)	1.993	(0.255)

TABLE 3.9 – $b(x) = x, \sigma(x) = \sqrt{1 + x^2}$. Moyenne et (variance) empiriques de $\widehat{\mu}_N$ et $\widehat{\omega}_N^2$ calculées à partir de 100 jeux de données simulés sous trois designs (N, T) et avec neuf jeux de paramètres (μ_0, ω_0^2).

δ	0.001	0.005	0.01	0.05	0.1	0.5	1
(n)	(5001)	(1001)	(501)	(101)	(51)	(11)	(6)
$\mu = -5$							
Mean	-5.022	-4.965	-5.006	-5.061	-5.061	-5.647	-6.481
(Var)	(0.067)	(0.056)	(0.054)	(0.068)	(0.078)	(0.144)	(0.324)
$\mu = 1$							
Mean	0.997	0.974	1.004	0.908	0.825	0.407	0.185
(Var)	(0.028)	(0.022)	(0.020)	(0.016)	(0.015)	(0.006)	(0.006)

TABLE 3.10 – Ornstein-Uhlenbeck à un effet aléatoire : $b(x) = x$ et $\sigma(x) = 1$. Moyenne et (variance) empiriques de $\widehat{\mu}_N^{(n)}$ calculés à partir de 100 jeux de données simulés, à $\omega^2 = 1$ connu, pour deux valeurs de μ et différents pas de discrétisation.

3.5 Bibliographie

[1] S. Allassonniere, E. Kuhn, and A. Trouvé. Construction of bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16(3) :641–678, 2010.

[2] S.L. Beal and L.B. Sheiner. Estimating population kinetics. *Critical Reviews in Biomedical Engineering*, 8 :195–222, 1982.

[3] M. Davidian and D.M. Giltinan. *Nonlinear models to repeated measurement data*. Chapman and Hall, 1995.

[4] M. Delattre, V. Genon-Catalot, and A. Samson. Maximum likelihood estimation for stochastic differential equations with random effects. URL `http://hal.archives-ouvertes.fr/hal-00650844`.

[5] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27 :94–128, 1999.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society - Series B*, 39 :1–38, 1977.

[7] S. Ditlevsen and A. De Gaetano. Mixed effects in stochastic differential equation models. *REVSTAT Statistical Journal*, 3 :137–153, 2005.

[8] S. Donnet and A. Samson. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM P& S*, 12 :196–218, 2008.

[9] S. Donnet and A. Samson. Em algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. 2010. submitted paper.

[10] J. Gabrielsson and D. Weiner. *Pharmacokinetic & Pharmacodynamic Data Analysis : Concepts and Applications*. 4 edition, 2006.

[11] M.S. Grewal and A.P. Andrews. *Kalman Filtering : Theory and Practice Using MATLAB*. Wiley, third edition, 2008.

[12] I. Karatsas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, 1997.

[13] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

[14] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49(4) :1020–1038, 2005.

[15] T. Kutoyants. *Parameter estimation for stochastic processes*. Helderman Verlag Berlin, 1984.

[16] R.S. Lipster and A.N. Shiryaev. *Statistics of random processes I : general theory*. Springer, 2001.

[17] T. Mazzoni. Computational aspects of continuous-discrete extended kalman-filtering. *Computational Statistics*, 23 :519–39, 2008.

[18] S.B. Mortensen, S. Klim, B. Dammann, N.R. Kristensen, H. Madsen, and R.V. Overgaard. A matlab framework for estimation of nlme models using stochastic differential equations-applications for estimation of insulin secretion rates. *Journal of Pharmacokinetics and Pharmacodynamics*, 34 :623–642, 2007.

[19] L. Nie. Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63(2) :123–143, 2006.

[20] L. Nie. Convergence rate of the mle in generalized linear and nonlinear mixed-effects models : Theory and applications. *Journal of Statistical Planning and Inference*, 137 : 1787–1804, 2007.

[21] L. Nie and M. Yang. Strong consistency of the mle in nonlinear mixed-effects models with large cluster size. *Sankhya : The Indian Journal of Statistics*, 67 :736–763, 2005.

[22] R.V. Overgaard, N. Jonsson, C.W. Tornøe, and H. Madsen. Non-linear mixed-effects models with stochastic differential equations : Implementation of an estimation algorithm. *Journal of Pharmacokinetics and Pharmacodynamics*, 32 :85–107, 2005.

[23] U. Picchini and S. Ditlevsen. Practical estimation of high dimensional stochastic differential mixed-effects models. *Computational Statistics and Data Analysis*, 55(3) :1426–1444, 2011.

[24] U. Picchini, A. De Gaetano, and S. Ditlevsen. Stochastic differential mixed-effects models. *Scandinavian Journal of Statistics*, 37 :67–90, 2010.

[25] J.C. Pinheiro and D.M. Bates. *Mixed-effect models in S and Splus*. Springer-Verlag, 2000.

[26] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer-Verlag, Berlin, third edition, 1999.

[27] C.W. Tornøe, R.V. Overgaard, H. Agersø, H.A. Nielsen, H. Madsen, and E.N. Jonsson. Stochastic differential equations in nonmem : Implementation, application, and comparison with ordinary differential equations. *Pharmaceutical Research*, 22 :1247–1258, 2005.

[28] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9 ; 0-521-78450-6.

[29] R. Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4) : 791–795, 1993. ISSN 0006-3444.

Chapitre 4

Sélection de modèles à effets mixtes

Contents

4.1 Introduction

Dans la dernière partie de cette thèse, nous nous intéressons au problème de la sélection de modèles dans une approche populationnelle par des procédures de type BIC (Bayesian Information Criterion). Les modèles considérés dans cette contribution sont des modèles à effets mixtes, linéaires ou non linéaires, et ne présentant pas nécessairement une structure markovienne. Avant de motiver nos résultats, nous rappelons l'intérêt de la pénalisation de la vraisemblance dans une démarche générale de choix de modèles.

Il n'existe pas de modèle statistique unique pour décrire un échantillon $\mathbf{y}$ donné, mais plusieurs modèles candidats n'ayant pas nécessairement la même complexité (*ie* le même nombre de paramètres). Dans une approche de population, ces modèles candidats peuvent différer par leur structure, le nombre de covariables ou encore le nombre d'effets aléatoires. Chacun d'entre eux est d'abord ajusté à l'échantillon en estimant ses paramètres, puis on cherche à déterminer le modèle le mieux ajusté aux données parmi les candidats. Contre l'intuition "naïve", le meilleur modèle n'est pas nécessairement celui pour lequel la vraisemblance des observations est la plus grande. Cette approche nous conduirait sans doute à choisir systématiquement le modèle de plus grande complexité. Plus le modèle est complexe et plus les écarts de ce modèle aux données sont susceptibles d'être faibles. En revanche, en estimer les paramètres avec précision n'est possible que si la taille de l'échantillon est suffisamment grande. Le meilleur modèle n'est pas non plus le modèle incluant le moins de paramètres. Bien que disposant d'une quantité suffisante d'observations on soit capable d'en estimer les paramètres avec précision, ce modèle risquerait de proposer une description trop sommaire voire fausse des données. Pour choisir le modèle le plus adéquat, il est donc primordial de réaliser un compromis entre le biais du modèle au données et la variance des estimateurs de ses paramètres. En d'autres termes, il s'agit de trouver un équilibre entre la dimension du modèle sélectionné et la taille de l'échantillon. Pénaliser la vraisemblance permet d'établir un tel équilibre.

Il existe deux grandes catégories de critères de type vraisemblance pénalisée : les critères non asymptotiques dont les propriétés reposent sur des inégalités Oracle (le lecteur pourra par exemple se référer à l'ouvrage de Massart [11]), et les critères asymptotiques, dont la forme est établie lorsque la taille de l'échantillon tend vers l'infini. Le BIC est l'un de ces critères asymptotiques, sans doute le plus utilisé dans la pratique. Il propose de pénaliser la log-vraisemblance par la dimension du modèle multipliée par le logarithme de la taille de l'échantillon [18, 16, 9]. Il existe à l'heure actuelle peu de références relatives à la sélection de modèles par des critères de vraisemblance pénalisée dans une approche de population. Nous pouvons essentiellement citer les travaux de Jiang and Sunil Rao [6] qui traitent de critères BIC et GIC (Generalized Information Criterion) pour le contexte précis des modèles linéaires à effets mixtes, restrictif en pratique. D'autre part, Jiang et al. [7] soulèvent la difficulté à définir la taille d'échantillon dans une appproche populationnelle. En effet, la définition même de ce qu'est la taille de l'échantillon ne s'impose pas clairement, tantôt choisie comme le nombre de sujets, tantôt comme le nombre d'observations par sujet, menant à un calcul arbitraire de la pénalité du BIC dans la pratique. Plusieurs versions du BIC sont d'ailleurs proposées et implémentées dans les logiciels destinés à l'analyse de modèles à effets mixtes. Par exemple, le logiciel MONOLIX ou encore la procédure `nlmixed` de SAS proposent de pénaliser la log-vraisemblance par le logarithme du nombre de sujets, alors que dans le logiciel NONMEM, la log-vraisemblance est pénalisée avec le nombre total d'observations.

Dans ce travail, nous utilisons les vitesses de convergence des différents paramètres et la forme de la matrice d'information de Fisher pour établir la forme de la pénalité du BIC attendue dans les modèles à effets mixtes. Nous nous plaçons dans le cadre non standard de double asymptotique où le nombre de sujets et le nombre d'observations par sujet tendent vers l'infini. Le critère obtenu forme un compromis entre les deux formulations du BIC utilisées dans la pratique en pénalisant certains paramètres par le logarithme du nombre de sujets et les autres par le logarithme du nombre total d'observations. Ce calcul de la pénalité du BIC nécessite au préalable l'étude théorique du maximum de vraisemblance dans les modèles à effets mixtes. Ce travail n'est pas standard, en lien avec la forme généralement non explicite de la vraisemblance d'un modèle à effets mixtes par des intégrales. Les seules contributions à l'heure actuelle sont celles de Nie and Yang [15], Nie [13, 14]. Elles montrent que les estimateurs des paramètres sont consistants et asymptotiquement gaussiens [13, 14], mais que lorsque le nombre de sujets et le nombre d'observations par sujet tendent conjointement vers l'infini, les vitesses de convergence des paramètres dépendent des niveaux de variabilité exprimés dans le modèle [14]. Ces propriétés sont illustrées sur des modèles pour lesquels les données d'un même sujet sont conditionnellement indépendantes sachant les paramètres individuels ϕ_i. Néanmoins, ce résultat repose sur des hypothèses techniques, difficiles à vérifier dans des modèles plus complexes incluant par exemple une structure de dépendance entre les observations individuelles. Nous vérifions que des propriétés similaires pour les paramètres dans les modèles de Markov cachés à effets mixtes s'établissent sous des conditions standard d'ergodicité des chaînes de Markov cachées.

Nous comparons ensuite les propriétés des trois pénalités du BIC (logarithme du nombre de sujets, logarithme du nombre total d'observations, pénalité hybride) sur des données simulées sous un modèle linéaire à effets mixtes pour la sélection de covariables. Nous montrons que la pénalité hybride se comporte comme la pénalité proportionnelle au logarithme du nombre de sujets et que la pénalité proportionnelle au logarithme du nombre total d'observations a tendance à sur-pénaliser les paramètres du modèle. À partir de ces premiers résultats, nous recommanderions de pénaliser la log-vraisemblance par le produit du nombre de paramètres et du logarithme du nombre de sujets. Des simulations sur des modèles plus généraux, incluant par exemple une structure de chaîne de Markov cachée sont encore nécessaires pour publier cette étude.

Ce travail, réalisé en collaboration avec Marie-Anne Poursat, et Marc Lavielle, est détaillé sous la forme d'un rapport de recherche INRIA. Les preuves techniques des propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans les modèles à effets mixtes sont disponibles auprès de l'auteur.

4.2 Rapport de recherche : BIC selection procedures in mixed effects models

Ce travail sera prochainement publié au format rapport de recherche INRIA.

Abstract

We consider the problem of variable selection in general nonlinear mixed-effets models, including mixed-effects hidden Markov models. These models are used extensively in the study of repeated measurements and longitudinal analysis. We propose a Bayesian Information Criterion (BIC) that is appropriate for nonstandard situations where both the number of subjects N and the number of measurements per subject n tend to infinity. In this case, the consistency rates of the maximum likelihood estimators (MLE) of the parameters depend on the level of variability designed in the model. We show that the MLE of the population parameters related to subject-specific parameters are $\sqrt{N}$-consistent whereas the MLE of the parameters related to fixed parameters are $\sqrt{Nn}$-consistent. We derive a BIC criterion with a penalty based on two terms proportional to $\log N$ and $\log Nn$. Finite-sample properties of the proposed selection procedure are investigated by simulation studies.

Keywords: Consistency rate, Nonlinear mixed model, Hidden Markov mixed-effects model, Variable selection.

4.2.1 Introduction

Nonlinear mixed-effects models are used in population studies where repeated measurements are observed from several independent subjects ([3]). Population studies occur in various fields such as pharmacokinetics or public health, for example to study disease evolution and to determine the effect of treatment or physiological covariates ([17], [4]). There is an extensive literature on parameter estimation in mixed models. However, the variable selection problem has been much less studied in these models. It is a standard practice to select the most relevant predictors using a Bayesian Information Criterion (BIC; [18]). The procedure consists in maximizing the observed log-likelihood penalized by the product of the dimension of the model and the logarithm of the sample size. Yet, the effective sample size is unclear in typical situations of mixed models. Therefore, the practice is to penalize the BIC either by the logarithm of the number of subjects or the logarithm of the total number of observations, without any guiding rule to choose between these two penalties. The purpose of this paper is to give a theoretical answer to the problem of choosing which penalty term is convenient in the practice to select the significant covariates.

To fix the notations, assume there are $i = 1, \ldots, N$ subjects and $j = 1, \ldots, n_i$ repeated observations nested within subject i. The ith observation consists in the vector of n_i observations $\mathrm{y}_i = (y_{i1}, \ldots, y_{in_i})$, where $y_{ij},\ j = 1, \ldots, n_i$ denotes the jth measure observed at time point t_{ij}. There are two specifications in a mixed model. First, the probability distribution of the y_i's is assumed to belong to a common parametric model $p(\mathrm{y}_i|\phi_i)$ specified by a vector of individual parameters ϕ_i of length K. Second, subject effects are added into the parametric

model by considering ϕ_i as a random vector. In this work, $\phi_i = (\phi_{i1}, \ldots, \phi_{iK})^T$ is defined as :

$$\phi_i = \beta X_i + \eta_i, \quad \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega), \quad i = 1, \ldots, N. \tag{4.1}$$

β is a $K \times p$ matrix of fixed-effects, X_i is the $p \times 1$ vector of known covariates for subject i. The vector of random effects $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^T$ represents the between subject variability that is not captured by the covariates. These are treated as random effects because the sampled subjects are thought to represent a population of subjects. They are assumed Gaussian and independent across subjects. The variance matrix Ω indicates the degree of heterogeneity of subjects. We denote by $\theta = (\beta, \Omega)$ the set of parameters of the global model that are to be estimated from the observations $\mathrm{y}_1, \ldots, \mathrm{y}_N$.

The idea of BIC is to penalize the log-likelihood by a term proportional to the logarithm of the sample size, which is N in the classic case. In mixed models, it can also be the total number of observations $N_{\text{tot}} = \sum_{i=1}^{N} n_i$. For example, consider the simple linear model $\mathrm{y}_{ij} = \phi_i + \xi_{ij}$, $\phi_i = \beta X_i + \eta_i$, $i = 1, \ldots, N$, $j = 1 \ldots, n_i$, where ξ_{ij} is a Gaussian residual error independent of the random effect η_i. The effective sample size to estimate the fixed effects $(\beta_{k1}, \ldots, \beta_{kp})$ is N if $\Omega_{kk} > 0$ whereas it is N_{tot} if $\Omega_{kk} = 0$ i.e. ϕ_{ik} is not random anymore but fixed. This motivates the asymptotic study of BIC when both the number of subjects N and the numbers of measurements per subject n_i tend to infinity.

As the penalty term in BIC relies on the asymptotic approximation of the distribution of the maximum likelihood estimators (MLE) of the model parameters, we first consider the consistency rates of the MLEs in the double-asymptotic situation where $N \to \infty$ and $n_i \to \infty$, $i = 1, \ldots, N$. In Section 4.2.2, we extend the results of [14] to general nonlinear mixed models such as mixed hidden Markov models where the repeated observations nested within each subject are dependent. In the double-asymptotic framework, the consistency rates of the maximum likelihood estimators (MLE) of the parameters depend on the level of variability designed in the model. We show that the MLE of the population parameters defining the subject-specific parameters are $\sqrt{N}$-consistent whereas the MLE of the parameters that are identical for all subjects are $\sqrt{N_{\text{tot}}}$-consistent. As a consequence, we obtain in Section 4.2.3 an appropriate BIC with a penalty based on two terms proportional to $\log N$ and $\log N_{\text{tot}}$. We illustrate the performance of this criterion with a simulation study. We conclude with a discussion in Section 4.2.4.

4.2.2 Asymptotic convergence of the MLE

For the sake of simplicity, we assume without loss of generality that the number of repeated observations n_i is the same across subjects : $n_i \equiv n$. The total number of observations is then $N_{\text{tot}} = \sum_i n_i = Nn$. We investigate the asymptotic behavior of the components of θ when $N \to \infty$ and $n \to \infty$.

Although $\sqrt{N}$-consistency of the MLE in standard parametric models are well-known results, some special features of the mixed-effects models may yield different convergence rates in the double-asymptotic framework.

We first consider two examples : the first one is a simple linear model that illustrates the special data structure of interest and the second one is a mixed hidden Markov model that motivated our study.

Example* 4.1. *Linear Mixed Model

$$\begin{aligned} y_{ij} &= \phi_i + \xi_{ij}, \\ \phi_i &= \mu + \eta_i, \end{aligned} \tag{4.2}$$

where $\eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2)$, $\xi_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, N$, $j = 1, \ldots, n$.
Here, the MLE of μ is

$$\begin{aligned} \widehat{\mu}_{\text{MLE}} &= \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} y_{ij}, \\ &= \mu + \frac{1}{N} \sum_{i=1}^{N} \eta_i + \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} \xi_{ij}, \end{aligned}$$

and

$$\text{Var}(\widehat{\mu}_{\text{MLE}}) = \frac{\omega^2}{N} + \frac{\sigma^2}{Nn}.$$

Thus, if $\omega > 0$, ϕ_i is an individual random parameter and the MLE of μ is $\sqrt{N}$-consistent, while if $\omega = 0$, $\phi_i = \mu$ is not a random parameter anymore and the MLE of μ is $\sqrt{Nn}$-consistent.

Example* 4.2. *Mixed Hidden Markov Model
We suppose that y_i is the realization of a S-state mixed hidden Markov model (MHMM) with Poisson emissions. For each $i = 1, \ldots, N$, the expression of $p(\text{y}_i|\phi_i)$ is given by

$$p(\text{y}_i|\phi_i) = p(y_{i1}) \prod_{j=2}^{n} p(y_{ij}|y_{i1}, \ldots, y_{i,j-1}, \phi_i),$$

and involves a sum over all possible sequences of n states in $\{1, \ldots, S\}$:

$$p(\text{y}_i|\phi_i) = \sum_{z_{i1}, \ldots, z_{in} \in \{1, \ldots, S\}^n} p(z_{i1}|\phi_i) \prod_{j=2}^{n} p(z_{ij}|z_{i,j-1}, \phi_i) \prod_{j=1}^{n} p(y_{ij}|z_{ij}, \phi_i). \tag{4.3}$$

We assume the initial state distribution to be uniform and known for all $i = 1, \ldots, N$. For $S = 2$, the specification of each individual hidden Markov model is reduced to the two Poisson parameters: λ_{1i} in state 1 and λ_{2i} in state 2, and the transition probabilities from state 1 to state 1 ($p_{11,i}$) and from state 2 to state 1 ($p_{21,i}$). This can be written as

$$\begin{aligned} \text{logit } p_{11,i} = \phi_{1i} \quad &, \quad \text{logit } p_{21,i} = \phi_{2i}, \\ \log \lambda_{1i} = \phi_{3i} \quad &, \quad \log \lambda_{2i} = \phi_{4i}, \end{aligned}$$

for all $i = 1, \ldots, N$, with $\phi_i = (\phi_{i1}, \ldots, \phi_{i4})^T$. See [4] for more details about this model and its use for epilepsy seizure count modelling.

4.2.2.1 The observed likelihood

In a mixed-effects model where the unobserved ϕ_i's are random variables, the probability distribution function (pdf) for subject i, $p(\mathbf{y}_i;\theta)$, $i = 1,\ldots,N$, is obtained by integrating the conditional distribution function of the data vector $\mathbf{y}_i$ with respect to ϕ_i's distribution:

$$p(\mathbf{y}_i;\theta) = \int p(\mathbf{y}_i,\phi_i;\theta)d\phi_i, \tag{4.4}$$

$$= \int p(\mathbf{y}_i|\phi_i)p(\phi_i;\theta)d\phi_i. \tag{4.5}$$

Except for linear mixed-effects models, the integral over the ϕ_i's do not have any explicit expression.

By independence of the N subjects, the joint pdf $p(\mathbf{y};\theta)$ is the product of the N individual pdf's $p(\mathbf{y}_i;\theta)$ and the MLE of θ is defined as

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\mathbf{y}_i;\theta).$$

To establish the theoretical properties of the MLE $\hat{\theta}$ of parameter θ when both the number of subjects N and the number of observations per subject n tend to infinity, we investigate the convergence rate of the components of the observed Fisher information matrix. Due to the special structure of the mixed effects models, the structure of the Fisher information matrix naturally divides the individual parameters ϕ_i into two groups : the individual parameters ϕ_{ik} which are not random (for which $\Omega_{kk} = 0$), and the individual parameters that randomly vary among subjects, corresponding to a random effect η_{ik} with non null variance. We denote by $\phi_{F,i}$ the components of ϕ_i that are not random and $\phi_{R,i}$ the components of ϕ_i that include a random component.

If the set of fixed parameters $\phi_{F,i}$ is not empty, then the decomposition of the pdf proposed in (4.5) does not hold any more. Indeed, to $\phi_i = (\phi_{F,i},\phi_{R,i})$ corresponds a natural partition of the model parameters $\theta = (\theta_F,\theta_R)$, and (4.5) should be replaced by

$$p(\mathbf{y}_i;\theta) = \int p(\mathbf{y}_i|\phi_{R,i},\phi_{F,i})p(\phi_{R,i};\theta_R)d\phi_{R,i}, \tag{4.6}$$

$$= \int p(\mathbf{y}_i|\phi_{R,i},\theta_F)p(\phi_{R,i};\theta_R)d\phi_{R,i}. \tag{4.7}$$

As the likelihood can not be expressed in a closed form, the study of the theoretical properties of the MLE in a mixed-effects model is not straightforward. There exists few results about the properties of the MLE in mixed-effects models. Some well known references to this topic are papers from Nie [15, 13, 14]. In these articles, the author suggests very specific tools for studying the MLE in mixed-effects models. In particular, the demonstrations proposed in [13, 14] are based on the individual complete likelihoods $p(\mathbf{y}_i,\phi_i;\theta)$, which generally have closed form expression, rather than the marginal likelihood of the observations. Thus, according to (4.7), the study of the asymptotic properties of the MLE is based on the following decomposition of the individual complete log-likelihoods:

$$l_i(\mathbf{y}_i,\phi_i;\theta) = l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F) + l_{i2}(\phi_{R,i},\theta_R).$$

Remark 15. The individual complete log-likelihood is decomposed into two terms which don't have the same order, since $l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F)$ is the conditional log-likelihood of the vector of observations $\mathbf{y}_i$ of size n, while $l_{i2}(\phi_{R,i}, \theta_R)$ is the log-likelihood of the random vector $\phi_{R,i}$ which size is fixed as given by the model.

4.2.2.2 Assumptions

For deriving the asymptotic distribution of the MLE of parameter θ, we need four classical assumptions, Assumptions 4.3 to 4.6 ensuring the regularity of the model and both continuity and invertibility of the Fisher Information Matrix in a neighborhood of the true parameter value $\theta^\star$. These assumptions can be found in [14] and are given in the Appendix.

When the number of observations per subject tends to infinity, an additional assumption, Assumption 4.1, is required, which ensures the regularity of $p(\mathbf{y}_i|\phi_i)$ for $i = 1, \ldots, N$:

Assumption 4.1. (i) $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left\{\left[{\phi_{R,i}}^T\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\phi_{R,i}\right]_{|\phi_i=\hat{\phi}_i}\right\} = o(n)$,

(ii) $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left\{\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\phi_{R,i}\right]_{|\phi_i=\hat{\phi}_i}\right\} = \mathcal{O}(n)$,

where Ω_R is the variance matrix of $\phi_{R,i}$, and $\hat{\phi}_i = \text{argmax}_{\phi_i}\ l_i(\mathbf{y}_i, \phi_i; \theta)$.

Assumption 4.1 is specific to the individual models and is necessary to evaluate the respective order of the components of the inverse Fisher information matrix. In Example 4.1, these conditions are easy to check. Nie [14] showed that they can be verified as well in mixed-effects regression models where the repeated observations nested within each subject are independent. In more general models such as Example 4.2, where the repeated observations are driven by a Markov structure dependence, we showed that Assumption 4.1 can be relaxed to classical ergodicity conditions:

Assumption 4.2. For all $j = 1, \ldots, n$, as n tends to infinity,

$$\left|\frac{\partial^2}{\partial\phi_i\partial\phi_i^T}\left(\log p(y_{ij}|y_{i,j-1},\ldots,y_{i1},\phi_i) - \log p(y_{ij}|y_{i,j-1},\ldots,\phi_i)\right)_{|\phi_i=\phi_i^\star}\right|$$

converges in probability to 0. $\phi_i^\star$ is the true individual parameter for subject i.

4.2.2.3 A Central Limit Theorem

We now give the asymptotic distribution of the MLE in mixed-effects models when $N, n \to +\infty$.

Lemma 4.2.1 (Asymptotic independence of $\hat{\theta}_R$ and $\hat{\theta}_F$)**.** *Under Assumptions 4.1 and 4.3 to 4.6, the MLEs for parameters θ_R and θ_F are asymptotically independent as N and n simultaneously tend to infinity, with $\frac{n}{N} \to +\infty$.*

Theorem 4.2.2 (Asymptotic distribution of the MLE when $N, n \to +\infty$)**.** *Assume that Assumptions 4.1 and 4.3 to 4.6 are verified. As N and n simultaneously tend to infinity, with*

$\frac{n}{N} \to +\infty$,

$$\begin{aligned}\sqrt{N}(\hat{\theta}_R - \theta_R^\star) &\xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_1(\theta^\star)),\\ \sqrt{Nn}(\hat{\theta}_F - \theta_F^\star) &\xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_2(\theta^\star)),\end{aligned}$$

where

$$\Gamma_1^{-1}(\theta^\star) = \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\phi_{R,i};\theta_R)}{\partial\theta_R \partial\theta_R{}^T}|_{\phi_i=\hat{\phi}_i}\right],$$

and

$$\begin{aligned}\Gamma_2^{-1}(\theta^\star) = \lim_{N,n\to+\infty} \frac{1}{Nn}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\Bigg[&\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F \partial\theta_F{}^T}\\ &-\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F \partial\phi_{R,i}{}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i} \partial\phi_{R,i}{}^T} + \Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i} \partial\theta_F{}^T}|_{\phi_i=\hat{\phi}_i}\Bigg],\end{aligned}$$

where $\theta^\star$ is the true parameter value.

Proof of Lemma 4.2.1 and Theorem 4.2.2. The starting point for getting the asymptotic distribution of the MLE in any parametric model is a Taylor series expansion of the log-likelihood around the true parameter value $\theta^\star$. From this, assuming that the model fulfills classical regularity conditions results in the convergence in distribution of the normalized score function, the convergence in probability of the normalized Hessian function of the log-likelihood, and the negligibility of the rest term of the Taylor series expansion. Invertibility of the Fisher information matrix is also required. When the number of subjects tends to infinity and the number of observations per subject is finite, provided that Assumptions 4.3 to 4.6 are fulfilled, we have:

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_R - \theta_R^\star\\ \hat{\theta}_F - \theta_F^\star\end{pmatrix} \underset{\mathrm{N}\to+\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta^\star)\right), \tag{4.8}$$

where

$$\mathcal{I}(\theta^\star) = -\lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\begin{pmatrix}\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R \partial\theta_R{}'} & \frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R \partial\theta_F{}'}\\ \frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial\theta_R{}'} & \frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial\theta_F{}'}\end{pmatrix} = \begin{pmatrix}\mathcal{I}_{11} & \mathcal{I}_{12}\\ \mathcal{I}_{21} & \mathcal{I}_{22}\end{pmatrix}.$$

See [1] for details. The distribution stated in (4.8) is degenerated as n tends to infinity, requiring adjustment of the rates of convergence of $\hat{\theta}_R$ and $\hat{\theta}_F$ as $n \to +\infty$. We show that some block components of $\mathcal{I}^{-1}(\theta^\star)$ tend to 0 as N, n tend to infinity, and adequately normalize the different components of the Fisher information matrix by evaluating the orders of the block components of $\mathcal{I}(\theta^\star)$ as suggested in [14]. Each component of the Fisher information matrix involves second derivatives of the marginal individual likelihoods with respect to θ_R and θ_F, but due to the expression of the likelihood as an integral over the ϕ_i's, evaluation of these derivatives is not straightforward and requires Laplace approximation of each second partial

derivative of $p(\mathbf{y}_i;\theta)$:

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta\partial\phi_i^T}\left(\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_i\partial\phi_i^T}\right)^{-1}\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta^T\partial\phi_i} + \mathcal{O}(n)\right]|_{\phi_i=\hat{\phi}_i}.$$

Using Assumption 4.1 and Laplace approximation of the partial derivatives of the individual log-likelihoods, we get the results of Lemma 4.2.1 and Theorem 4.2.2. More details are given in the Appendix.

□

4.2.3 Model selection

We will use the convergence results obtained in the previous section to derive the BIC penalty in a population approach context for selecting the covariate model.

4.2.3.1 BIC for mixed-effects models

Deriving the BIC penalty requires to take a finite collection of mixed-effects models, $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m)$, and to consider these models and their respective parameters as random variables. In model $\mathcal{M}_k$, the data is supposed to have distribution $p(\cdot;\theta_k)$ characterized by a set of population parameter θ_k.

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ denote the observations sample, composed of n independent sequences of longitudinal data: $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$, where the n observations for a given subject are not necessarily independent. The aim is to select model $\mathcal{M}_k$ among collection $\mathcal{M}$ presenting highest a posteriori distribution $p(\mathcal{M}_k|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_k)p(\mathcal{M}_k)$. Assuming uniform a priori distribution for the models of $\mathcal{M}$, the problem amounts to maximizing $p(\mathbf{y}|\cdot)$ among the models of the collection. For any given $k = 1, \ldots, m$,

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k,$$

is evaluated using a Laplace approximation, according to [9]:

$$\begin{aligned}\log p(\mathbf{y}|\mathcal{M}_k) &= \left(\sum_{i=1}^{N}\log p(\mathbf{y}_i|\theta_k^\star)\right) + \log p(\theta_k^\star|\mathcal{M}_k) + \frac{\dim(\theta_k)}{2}\log(2\pi) \\ &\quad -\frac{\dim(\theta_k)}{2}\log(N) - \frac{1}{2}\log\det\left(-\frac{1}{N}\frac{\partial^2 \log p(\mathbf{y}|\mathcal{M}_k)}{\partial\theta\partial\theta^T}|_{\theta_k=\theta_k^\star}\right) \\ &\quad +\mathcal{O}(N).\end{aligned}$$

The Hessian matrix $-\frac{1}{N}\frac{\partial^2 \log p(\mathbf{y}|\mathcal{M}_k)}{\partial\theta\partial\theta^T}|_{\theta_k=\theta_k^\star}$ is approximated with $\mathcal{I}(\hat{\theta}_k)$, but when both N and n tend to infinity, $\log\det\mathcal{I}(\hat{\theta}_k)$ can not be evaluated as a constant as in [9].

We assume here a linear model for the individual parameters as described in (4.1). We therefore decompose the vector of random effect η_i into $(\eta_{F,i}, \eta_{R,i})$ where $\eta_{F,i} = 0$ is associated

to the fixed individual parameters and $\eta_{R,i}$ to the random ones:

$$\phi_{F,i} = \beta_F X_i, \tag{4.9}$$
$$\phi_{R,i} = \beta_R X_i + \eta_{R,i}, \quad \eta_{R,i} \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega_R), \quad i = 1, \ldots, N, \tag{4.10}$$

where Ω_R is a positive-definite variance covariance matrix. We consider here that all the models of collection $\mathcal{M}$ have the same covariance structure in the sense that the covariance matrix Ω of the individual parameters ϕ_i's has the same structure in the m models. In other words, the decomposition $\phi_i = (\phi_{F,i}, \phi_{R,i})$ is the same for all the models. We focus on the covariate selection problem, *i.e.* the selection of the non zero elements of β_F and β_R, but not on the selection of the non zero elements of Ω. Here, θ is decomposed into (θ_R, θ_F), where $\theta_R = (\beta_R, \Omega_R)$ and θ_F is β_F, eventually augmented with model-specific fixed parameters such as the error variance σ^2 as in Example 4.1.

Using asymptotic independence of $\hat{\theta}_R$ and $\hat{\theta}_F$ (Lemma 4.2.1), and the convergence rates of $\hat{\theta}_R$ and $\hat{\theta}_F$ as $N, n \to +\infty$ (Theorem 4.2.2), we get

$$-\log\det\mathcal{I}(\hat{\theta}_k) \approx -\log\det(\Gamma_1(\hat{\theta}_k)) - \log\det(n\Gamma_2(\hat{\theta}_k)),$$

where, using the same notations as in previous section, $\Gamma_1(\hat{\theta}_k)$ and $n\Gamma_2(\hat{\theta}_k)$ represent the diagonal block components of $\mathcal{I}(\hat{\theta}_k)^{-1}$. According to Theorem 4.2.2, $\Gamma_1(\hat{\theta}_k)$ and $\Gamma_2(\hat{\theta}_k)$ are evaluated as constants when both $N, n \to +\infty$. Thus, correctly normalizing each term of Laplace approximation expressed above and neglecting all terms remaining bounded as $N, n \to +\infty$, we get

$$\log p(\mathbf{y}|\mathcal{M}_k) \approx \log p(\mathbf{y}|\hat{\theta}_k) - \frac{\dim(\theta_{R,k})}{2}\log N - \frac{\dim(\theta_{F,k})}{2}\log Nn,$$

as $N, n \to +\infty$ and $\dfrac{n}{N} \to +\infty$.

Theorem 4.2.3. *The BIC procedure consists in selecting the model that minimizes*

$$BIC(\mathcal{M}_k) = -2\log p(\mathbf{y}|\hat{\theta}_k, \mathcal{M}_k) + \dim(\theta_{R,k})\log N + \dim(\theta_{F,k})\log(Nn),$$

among the models $\mathcal{M}_k$ of collection $\mathcal{M}$.

Remark 16. Replace Nn by $N_{tot} = \sum_{i=1}^{N} n_i$ in BIC expression when the number of observations per subject differs from one subject to an other.

4.2.3.2 Simulation study

We now investigate the contributions of the hybrid BIC penalty in variable selection problems occuring in mixed-effects models. More precisely, we confront the new criteria $BIC_{N,Nn}$ to BIC penalized with either the logarithm of the number of subjects (BIC_N) or the logarithm of the total number of observations (BIC_{Nn}), in a variable selection problem via a short simulation study.

a) Model

We use the following linear mixed-effects model for the simulations, which generalizes the simple model of Example 4.1. For $i = 1, \ldots, N$ and $j = 1, \ldots, n$,

$$\begin{aligned} y_{ij} &= \mu + \phi_{i1} t_{ij} + \phi_{i2} t_{ij}^2 + \xi_{ij} \quad , \quad \xi_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \\ \phi_{i1} &= a_0 + a_1 x_i + \eta_i \quad , \quad \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2), \\ \phi_{i2} &= b_0 + b_1 x_i, \end{aligned} \tag{4.11}$$

where x_i is a covariate for subject i and t_{ij} the observation time of y_{ij}. Here, $\phi_{R,i} = \phi_{i1}$, $\phi_{F,i} = \phi_{i2}$, $X_i = (1, x_i)$ and $\Omega_R = \omega^2$.

In this simple model, the distribution of the observations is explicit: $y_{ij} \sim \mathcal{N}(m_{ij}, \gamma_{ij}^2)$, where $m_{ij} = \mu + (a_0 + a_1 x_i) t_{ij} + (b_0 + b_1 x_i) t_{ij}^2$ and $\gamma_{ij} = t_{ij}^2 \omega^2 + \sigma^2$. Thus, the likelihood can be easily derived for any vector of population parameter θ. Four sub-models, M_{11}, M_{10}, M_{01}, M_{00}, can be derived from model (4.11), depending on whether the coefficients a_1 and b_1 are null or not. The expression of the BIC hybrid penalty for each sub-model is given in Table 4.1. Note that in this specific model, the variable selection problem arises at two levels of the model: on the random parameter ϕ_{i1} on the one hand, on the non random parameter ϕ_{i2} on the other hand.

Model	Description	θ_R	θ_F	pen($BIC_{N,Nn}$)
M_{11}	($a_1 \neq 0$, $b_1 \neq 0$)	a_0, a_1, ω^2	μ, b_0, b_1, σ^2	$3 \log N + 4 \log Nn$
M_{10}	($a_1 \neq 0$, $b_1 = 0$)	a_0, a_1, ω^2	μ, b_0, σ^2	$3 \log N + 3 \log Nn$
M_{01}	($a_1 = 0$, $b_1 \neq 0$)	a_0, ω^2	μ, b_0, b_1, σ^2	$2 \log N + 4 \log Nn$
M_{00}	($a_1 = 0$, $b_1 = 0$)	a_0, ω^2	μ, b_0, σ^2	$2 \log N + 3 \log Nn$

Table 4.1: Definition of submodels M_{11}, M_{10}, M_{01} and M_{00}, and derivation of the corresponding penalties for hybrid BIC.

b) Design for the simulations

We will perform a Monte-Carlo experiment in order to evaluate the performance of our covariate model selection criteria.

Each of the $K = 500$ replicates of the Monte-Carlo experiment consists in simulating a new dataset $\mathcal{Y}_{N,n}^{k,k'}$ with N subjects and n observations per subject from model $M_{kk'}$, $k, k' \in \{0, 1\}^2$. The parameters θ_R and θ_F of models M_{11}, M_{10}, M_{01}, M_{00}, are estimated from the simulated observations using the EM algorithm. Using the estimated values $\hat{\theta}_R$ and $\hat{\theta}_F$, the observed likelihood is computed in each model of the collection, and the corresponding values for BIC_N, BIC_{Nn} and $BIC_{N,Nn}$ are derived. Then, minimization of each criteria leads to the selection of one of the four possible models. We can then obtain the number of times each model has been chosen according to each criteria.

Different designs (number N of subjects and number n of observations per subject) were investigated using $N = (20, 50, 100)$ and $n = (20, 100, 200, 500, 1000)$. The variance of the residual error was set to $\sigma^2 = 1$. The other model components of the model are randomly drawn at each replicate of the Monte Carlo experiment:

$$C_i \sim \mathcal{N}(0, 1), \mu, a_1, b_1 \sim \mathcal{N}(0.01, 1), b_0 \sim \mathcal{N}(0.005, 1), b_1 \sim \mathcal{N}(0.0025, 1) \text{ and } \omega^2 \sim \mathcal{U}_{[0.01, 1.01]}.$$

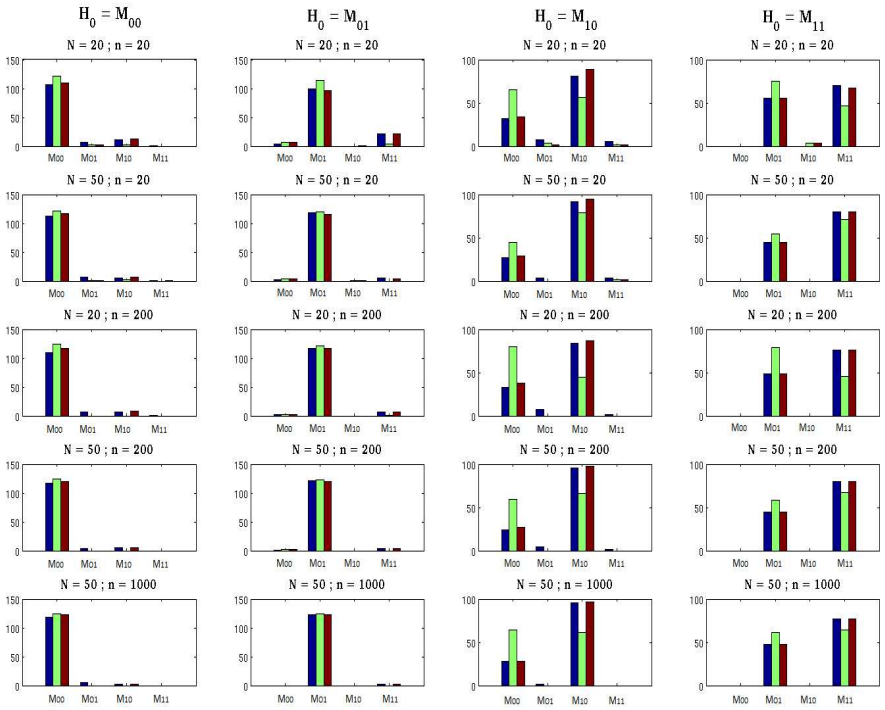

Figure 4.1: Results given by the three versions of BIC on 500 simulated datasets for the different sub-models of (4.11) (columns) and different values for (N, n) (raws). For each design, the histograms represent the numbers of times each model (from left to right: M_{00}, M_{01}, M_{10}, M_{11}) is selected by each version of BIC: BIC_N (blue), BIC_{Nn} (green), $BIC_{N,Nn}$ (red).

The t_{ij}'s are equally spaced in $[0, 10]$.

c) Results

The results are displayed in Figure 4.1. The Monte-Carlo study mainly shows that the new criteria has very similar properties than BIC penalized with the number of subjects, which is currently the most frequently used model selection criteria in a population approach in practice. Nevertheless, when the variable selection problem arises at random effect level $\phi_{R,i}$ of the model (models M_{10} and M_{11}), we notice most important differences between the three criteria. When there is a covariate effect on $\phi_{R,i}$ only (model M_{10}), the hybrid BIC is even slightly better than BIC_N - in the sense that $BIC_{N,Nn}$ most frequently selects the correct model than BIC_N - especially when the number of subjects is small ($N = 20$). We also show very bad performances of BIC_{Nn} in this context. When a covariate effect arises at least at random effect level of the model, it over-penalizes parameters θ_R of the model, resulting in most frequent selection of the model without covariate on $\phi_{R,i}$. On the other hand, BIC_{Nn} gives slightly better results when no covariate arises at random level of the model.

We have replicated the Monte-Carlo experiment with different values for the parameters and the t_{ij}'s. As expected, the performances of the three criteria are sensitive to the parameter scales. When greater weight is granted to the random effect term $\phi_{R,i}$ of the model, it is

intuitively easier to detect a covariate in $\phi_{R,i}$ than in $\phi_{F,i}$, and vice versa. For example, when the observation time interval is $[0, 1]$, the hybrid BIC show much better properties than above when the problem is to detect a covariate in $\phi_{R,i}$ and not in $\phi_{F,i}$ (model M_{10}), but mainly selects model M_{10} even when the true model is M_{11}. Similar results are recorded with the other criteria.

4.2.4 Discussion

The contribution of the present paper is twofold. First, we have extended the asymptotic results established by Nie [14] for nonlinear mixed-effects models to more general models involving a dependence structure within each subject, such as mixed-effects hidden Markov models. We consider the double-asymptotic framework where both the number of subjects N and the number of repeated measures n tend to infinity. We show that, provided some classical ergodicity conditions on the individual Markov chains, the rates of convergence of the MLEs depends on the levels of variability in the model : the parameters θ_R involved in the random components of the individual parameters $\phi_i,\ i = 1, \ldots, N$ are $\sqrt{N}$-consistent while the parameters θ_F related to the non-random individual parameters are $\sqrt{Nn}$-consistent. We have then derived from these results a specific version of BIC for covariate selection in mixed-effects models. The new BIC criterion penalizes the size of θ_R with the logarithm of the number of subjects and the size of θ_F with the logarithm of the total number of observations.

We have performed a simulation study for comparing the behavior of the proposed BIC with standard BIC criteria that are implemented in different softwares used for the analysis of mixed effects models. Using a simple linear mixed-effects model, we have found that the hybrid BIC mainly behaves as the classical BIC penalized with the logarithm of the number of subjects but outperforms in most cases the BIC penalized with the logarithm of the total number of observations. In this illustrative example, we have also noticed some slight superiority of the new criteria in some situations, even with moderate N and n.

Additional simulations involving more complex models such as mixed-effects hidden Markov models are required to investigate the empirical properties of the proposed criterion. Furthermore, the proposed BIC is designed only for covariate selection when the structure of the random effects of the model is given. Deriving a new criteria for selecting the covariance structure of the random effects as well remains an open problem. BIC-type procedures were studied by Jiang and Sunil Rao [6] to select both fixed and random effects but they are limited to linear mixed models. Few papers addressed the problem of variable selection in nonlinear mixed models. Jiang et al. [7] proposed a "fence" method to select predictors in generalized linear models, Kinney and Dunson [8] implemented a fully Bayesian selection approach in mixed logistic models, Ni et al. [12] proposed a double-penalized likelihood approach for simultaneous model selection and estimation in semiparametric mixed models. In the future we hope to develop an appropriate criterion that would be able to perform both covariate and covariance selection in a population approach.

4.2.5 Appendix

4.2.5.1 Assumption of Theorem 4.2.2

We formulate conditions providing asymptotic distribution of the MLE when $N, n \to +\infty$.

Let ϑ denote an open subset of Θ. We assume that for any given n:

Assumption 4.3. For all $i = 1, \ldots, N$, and for all $\theta \in \vartheta$, $p(\mathbf{y}_i; \theta)$ admits all first, second and third derivatives with respect to θ for almost all $\mathbf{y}_i$.

Assumption 4.4. (i) There exists $M > 0$ such that for all $i = 1, \ldots, N$, for all $\theta \in \vartheta$ and all $k = 1, \ldots, r$,

$$\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial \log p(\mathbf{y}_i;\theta)}{\partial \theta_k}|_{\theta=\theta^\star}\right]^2 < M \text{ and } \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial \theta_k \partial \theta_l}|_{\theta=\theta^\star}\right]^2 < M.$$

(ii) Moreover, there exists a sequence of functions $\{G_1(\mathbf{y}_1), \ldots, G_N(\mathbf{y}_N)\}$ such that for all $\theta \in \vartheta$, for all $i = 1, \ldots, N$ and for all $k, l, h = 1, \ldots, r$,

$$\left|\frac{\partial^3 \log p(\mathbf{y}_i;\theta)}{\partial \theta_k \partial \theta_l \partial \theta_h}\right| \leq G_i(\mathbf{y}_i) \text{ and } \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[G_i^2(\mathbf{y}_i)\right] \leq M.$$

Assumption 4.5. Let $V_N(\theta) = H_N(\theta^\star)^{-\frac{1}{2}} H_N(\theta) H_N(\theta^\star)^{-\frac{T}{2}}$, and $H_N(\theta) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T}$ for all $\theta \in \vartheta$. Matrix $H_N(\theta^\star)$ is positive definite and invertible, and

$$\max_{\theta\in\vartheta} ||V_N(\theta) - I_r|| \underset{N\to+\infty}{\longrightarrow} 0,$$

where I_r stands for the $r \times r$ identity matrix.

Assumption 4.6. $\liminf_{N\to+\infty} \lambda_N = \lambda > 0$ where λ_N is the smallest eigenvalue of matrix

$$F_N = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T}\right].$$

4.2.5.2 Evaluation of the Fisher Information Matrix when $N, n \to +\infty$

Here, we focus on the evaluation of the components of the inverse Fisher Information Matrix as $N, n \to +\infty$ and $n/N \to +\infty$. The upper and lower diagonal blocks of $\mathcal{I}^{-1}$ are respectively given by $\left(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}$ and $\left(\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right)^{-1}$ and the anti-diagonal blocs are given by $-(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21})\mathcal{I}_{12}\mathcal{I}_{22}^{-1}$ and its transpose. First of all, key issue is to get the orders of magnitude of $\mathcal{I}_{11}$, $\mathcal{I}_{21}$, $\mathcal{I}_{12}$ and $\mathcal{I}_{22}$ as $n \to +\infty$. Let us detail the approach on block $\mathcal{I}_{21}$. Getting an evaluation of $\mathcal{I}_{21}$ requires Laplace approximation of $-\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial {\theta_R}^T}$ for all $i = 1, \ldots, N$. It is given by:

$$-\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_R}^T} = \left[\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta_F\partial{\theta_R}^T} - \frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}\right)^{-1}\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_{R,i}\partial{\theta_R}^T}\right]|_{\phi_i=\hat{\phi}_i} + \mathcal{O}(n),$$

and can be simplified into

$$-\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R\partial{\theta_F}^T} = -\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial{\theta_R}^T}\right]|_{\phi_i=\hat{\phi}_i} + \mathcal{O}(n).$$

Linear Gaussian model for the ϕ_i's makes $\dfrac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial{\theta_R}^T}$ expressed as a first-order polynomial function of ϕ_i. Using results of Lemma 1, and correctly normalizing each component in Laplace approximation of $\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R\partial{\theta_F}^T}$ with n, we establish that as $N, n \to +\infty$ such that $n/N \to +\infty$:

$$-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_R}^T}\right] = \mathcal{O}(n).$$

In other words, $\mathcal{I}_{21}$ converges to a constant matrix as $n \to +\infty$. As $\mathcal{I}_{21} = \mathcal{I}_{12}^T$, we also evaluate $\mathcal{I}_{12}$ as a constant as $n \to +\infty$. Similarly, we use result of Lemma 1 as well and Laplace approximation of $-\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R\partial{\theta_R}^T}$, and we get

$$\lim_{N,n\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\theta_R\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial{\theta_R}^T}\right]_{|_{\phi_i=\hat{\phi}_i}} = 0,$$

thus convergence of $\mathcal{I}_{11}$ to a constant positive definite matrix as $n \to +\infty$. Indeed, $\mathcal{I}_{11} = \lim_{N\to+\infty}\dfrac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\dfrac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\theta_R\partial{\theta_R}^T}\right]$. Recall that the ϕ_i's are independent and identically distributed random variables. Thus, $\mathcal{I}_{11} = \dfrac{\partial^2 l_{12}(\phi_1^R,\theta_R)}{\partial\theta_R\partial{\theta_R}^T}$, which is a positive and definite matrix, and does not depend on the number of observations per subject.

Finally, using Laplace approximation of $\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}$, we get equivalence between

$$-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}\right] \text{ and}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\theta_F}^T}\right.$$
$$\left.-\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial\theta_F}\right]|_{\phi_i=\hat{\phi}_i}$$

as n tends to infinity.

Using Assumption 4.1 and previous Laplace approximation of the partial second derivatives of the individual log-likelihoods, we are able to evaluate each block component of matrix $\mathcal{I}$. Providing that $\hat{\phi}_i$ is a consistent estimate of ϕ_i^* for any value of θ, we show that $\left(\mathcal{I}_{11}-\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1}$ converges in probability to 0 as N tends to infinity for any value of n, giving result of Lemma 4.2.1. Similarly, we show that $\left(\mathcal{I}_{11}-\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}$ is of the order of a constant, and that the distribution of $\sqrt{N}(\hat{\theta}_F-\theta_F^\star)$ is degenerated since $\left(\mathcal{I}_{22}-\mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right)^{-1}$ converges in probability to 0 as N and n simultaneously tend to infinity. We take again the classical frame for demonstration of convergence in distribution of the maximum likelihood estimate, focusing on θ_F only, and normalizing the log-likelihood with Nn. To get $\sqrt{Nn}$ convergence in distribution of $\hat{\theta}_F$, we only need positive definiteness of $\lim_{N,n\to+\infty}-\frac{1}{Nn}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}\right]$. This results from Assumption 4.1 by evaluating N terms $\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}$ with its Laplace approximation.

4.3 Bibliographie

[1] R.A. Bradley and J.J. Gart. The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2) :205–214, 1962.

[2] O. Cappé, T. Rydén, and E. Moulines. *Inference in hidden Markov models*. Springer, 2005.

[3] M. Davidian and D.M. Giltinan. Nonlinear models for repeated measurement data : An overview and update. *Journal of Agricultural Biological and Environmental Statistics*, 8 :387–419, 2003.

[4] M. Delattre. Inference in mixed hidden markov models and applications to medical studies. *Journal de la Société française de statistique*, 151(1) :90–105, 2010.

[5] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1) :342–368, 1985.

[6] J. Jiang and J. Sunil Rao. Consistent procedures for mixed linear model selection. *Sankhyã : The Indian Journal of Statistics*, 65(1) :23–42, 2003.

[7] J. Jiang, J. Sunil Rao, Z. Gu, and T. Nguyen. Fence methods for mixed model selection. *The Annals of Statistics*, 36 :1669–1692, 2008.

[8] S.K. Kinney and D.B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63 :690–698, 2007.

[9] E. Lebarbier and T. Mary-Huard. Une introduction au critère bic : Fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1) :39–57, 2006.

[10] E.L. Lehmann and G. Casella. *Theory of point estimation*. Springer, 1998.

[11] P. Massart. *Concentration Inequalities and Model Selection : Ecole d'Été de Probabilités de Saint-Flour XXXIII - 2003*. Springer, 1 edition, 2007.

[12] X. Ni, D. Zhang, and H.H. Zhang. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66 :79–88, 2010.

[13] L. Nie. Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63 :123 – 143, 2006.

[14] L. Nie. Convergence rate of the mle in generalized linear and nonlinear mixed-effects models : Theory and applications. *Journal of Statistical Planning and Inference*, 137 : 1787–1804, 2007.

[15] L. Nie and M. Yang. Strong consitency of mle in nonlinear mixed-effects models with large cluster size. *Sankhya, The Indian Journal of Statistics*, 67 :736–763, 2005.

[16] A.E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25 : 111–163, 1995.

[17] A. Samson, M. Lavielle, and F. Mentré. The saem algorithm for group comparison tests in longitudinal analysis based on non-linear mixed-effects models. *Statistics in medecine*, 26 :4860–4875, 2007.

[18] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[19] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393) :82–86, 1986.

Chapitre 5

Discussion

5.1 Travaux annexes : valorisation des modèles à effets mixtes

5.1.1 Motivations

Les modèles à effets mixtes, les méthodes d'inférence qui ont été développées pour ces modèles à partir de l'algorithme SAEM et leur implémentation dans le logiciel MONOLIX fournissent une large gamme d'outils statistiques facilement utilisables dans de nombreuses disciplines. L'approche de population est cependant encore mal comprise et peu utilisée aujourd'hui en dehors de la pharmacologie. J'ai pu appréhender la diversité des applications possibles des modèles à effets mixtes en travaillant en parallèle de ma thèse avec des biologistes cellulaires de l'université d'Orsay. Dans cette collaboration, mon travail a été de présenter simplement la philosophie de ces modèles et d'en motiver l'utilisation, peu commune dans ce domaine de la biologie. Pour les expériences sur lesquelles j'ai été amenée à travailler, les données étaient des mesures longitudinales obtenues sur plusieurs unités expérimentales et se prêtaient naturellement à l'utilisation de modèles à effets mixtes. Pour comparer deux conditions à partir de telles données, la pratique des statistiques en biologie cellulaire est souvent simpliste voire peu rigoureuse. En effet, l'usage est de comparer des intervalles de confiance calculés en chaque temps de mesure à partir des observations de chaque condition. Cette approche ne tient pas compte des différentes sources de variabilité dans les données, en particulier de la variabilité inter-unités expérimentales dans chaque condition, induisant un biais potentiel dans la comparaison des groupes. Nous avons donc proposé une analyse basée sur des modèles à effets mixtes, permettant ainsi une évaluation beaucoup plus fine des différences entre plusieurs groupes d'unités expérimentales.

5.1.2 Exemple : construction d'un modèle non linéaire à effets mixtes pour des données de fluorescence

Les neutrophiles sont des cellules sanguines capables d'internaliser des microorganismes afin de lutter contre les infections. Pour ce faire, il se forme un compartiment appelé phagosome contenant le microorganisme à l'intérieur de la cellule. Dans le phagosome, le microorganisme est ensuite détruit grâce à des espèces réactives de l'oxygène qui sont produites à l'intérieur du phagosome par une enzyme. J'ai analysé des expériences destinées à comprendre les mécanismes sous-jacents à ce phénomène. J'ai en particulier travaillé sur des données de fluorescence mesurant la production d'espèces réactives de l'oxygène suite à la formation du

phagosome. Le production d'espèces réactives de l'oxygène est liée à la présence de certains lipides membranaires. Pour évaluer le rôle de ces lipides, la production des espèces réactives de l'oxygène a été mesurée par fluorescence en présence et en l'absence d'une sonde diminuant leur accessibilité. L'expérience est répliquée sur plusieurs cellules de chacune des deux conditions (avec ou sans sonde). La Figure 5.1 montre les mesures de fluorescence obtenues sur quelques cellules.

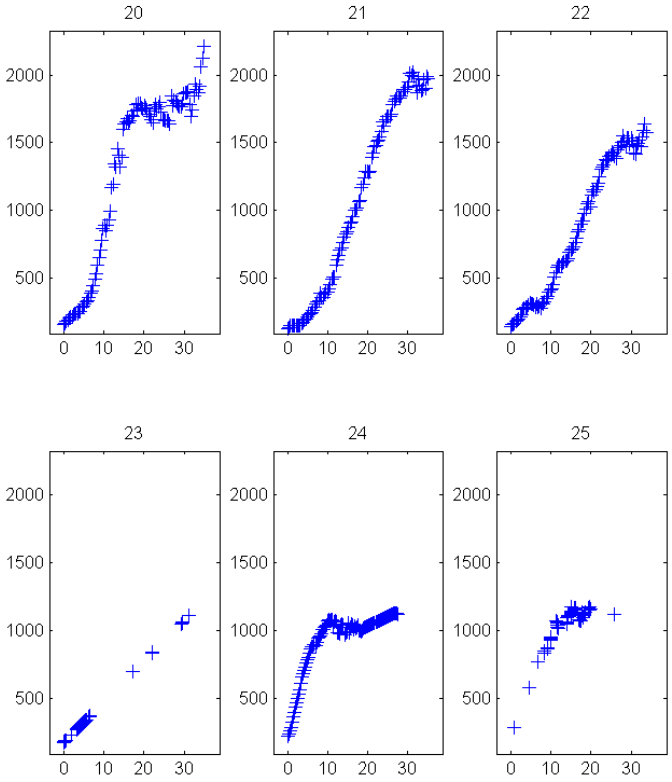

FIGURE 5.1 – Mesures de fluorescence de 6 cellules de l'échantillon.

Pour comparer la production d'espèces réactives de l'oxygène dans les cellules contrôle et les cellules marquées par une sonde, nous avons retenu le modèle suivant :

$$\begin{aligned} y_{ij} &= f_i(t_{ij}) + \xi_{ij}, \\ \xi_{ij} &\sim \mathcal{N}(0, a + b f_i(t_{ij})), \\ f_i(t_{ij}) &= F_{0i} + \frac{F_{\max,i} - F_0}{1 + e^{-\alpha_i (t_{ij} - T_{c,i})}}. \end{aligned}$$

Dans ce modèle, les paramètres F_{0i} et $F_{\max,i}$ représentent respectivement la fluorescence initiale et la fluorescence finale de la cellule i. $T_{c,i}$ est le temps auquel la fluorescence mesurée sur la cellule i atteint la moitié de son niveau maximal. α_i peut être compris comme un

Paramètres	Valeur estimée	s.e.	r.s.e.	p-value
F_0	63	5.500	9	
$F_{\max}$	1.12×10^3	87	8	
α	0.304	0.041	14	
β_α	-0.330	0.170	52	0.057
T_c	4.320	0.940	22	
β_{T_c}	0.726	0.280	38	0.008
ω_{F_0}	0	-	-	
$\omega_{F_{\max}}$	0.413	0.056	14	
$\omega_\alpha(CE=0)$	0.541	0.088	16	
$\omega_\alpha(CE=1)$	0.385	0.068	18	
$\omega_{T_c}(CE=0)$	0.865	0.140	16	
$\omega_{T_c}(CE=1)$	0.603	0.110	18	
a	15.6	0.780	5	
b	0.050	0.001	2	

TABLE 5.1 – Estimateurs du maximum de vraisemblance des paramètres du modèle final et leurs s.e. obtenus par l'algorithme SAEM sur les données de fluorescence.

paramètre indiquant la vitesse à laquelle le niveau fluorescence évolue dans le temps pour la cellule i. Les F_{0i} sont supposés identiques pour toutes les cellules : $F_{0i} = F_0$. Nous supposons que la différence entre les deux types de cellules est susceptible d'intervenir au niveau des paramètres α_i et $T_{c,i}$ du modèle. Nous définissons alors $F_{\max,i}$, α_i et $T_{c,i}$ comme des variables aléatoires de la façon suivante :

$$\begin{aligned}
\log(\alpha_i) &= \log(\alpha) + \beta_\alpha.CE_i + \omega_\alpha(CE_i)\ \eta_{1i}, \\
\log(T_{c,i}) &= \log(T_c) + \beta_{T_c}.CE_i + \omega_{T_c}(CE_i)\ \eta_{2i}, \\
\log(F_{\max,i}) &= \log(F_{\max}) + \omega_{F_{\max}}\ \eta_{3i},
\end{aligned}$$

où

$$\begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right),$$

et CE_i désigne la covariable indiquant si la cellule i est une cellule contrôle ($CE_i = 0$) ou une cellule marquée d'une sonde ($CE_i = 1$).

Nous avons estimé les paramètres de ce modèle par maximum de vraisemblance via l'algorithme SAEM. Les résultats figurent dans le Tableau 5.1. Nous avons ainsi pu mettre en évidence une différence significative dans la production d'espèces réactives de l'oxygène dans les deux conditions (Figure 5.2).

Ce travail, réalisé conjointement avec Marie-Cécile Faure, Jean-Claude Sulpice, Marc Lavielle, Magali Prigent, Marie-Hélène Cuif, Chantal Melchior, Eric Tschirhart, Oliver Nüsse et Sophie Dupré-Crochet est en cours de rédaction et devrait être soumis dans une revue de biologie cellulaire dans les prochains mois.

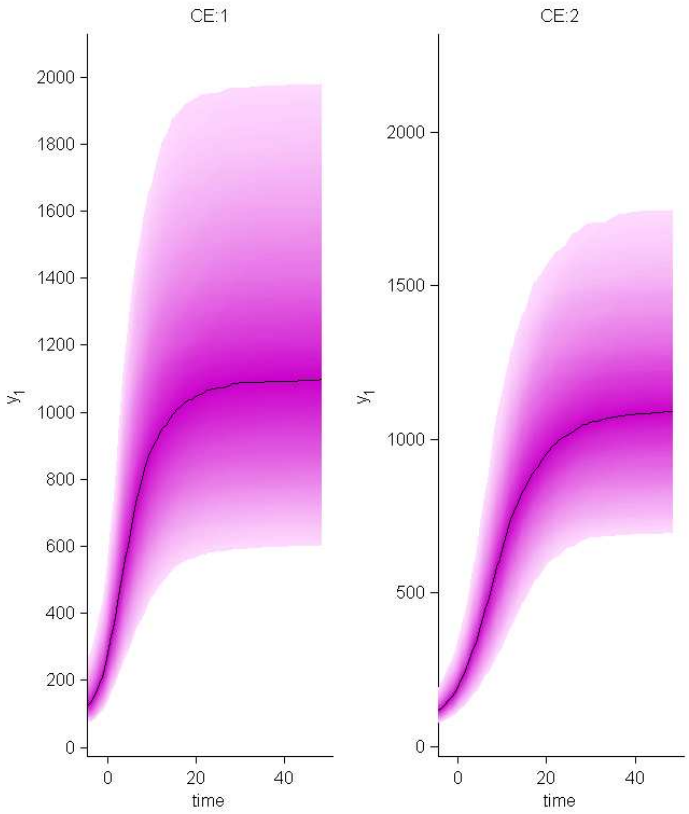

FIGURE 5.2 – Comparaison de la distribution des fluorescences dans les cellules contrôle (panneau gauche) et dans les cellules marquées par la sonde (panneau droit) prédite par le modèle final, sans erreur de mesure.

5.2 Bilan de la thèse et perspectives de recherche

L'approche de population a un succès croissant ces dernières années en modélisation statistique. Cette approche se traduit par l'utilisation de modèles à effets mixtes qui permettent une description rigoureuse des différentes sources de variabilité dans les échantillons composés de mesures répétées sur plusieurs sujets. Les modèles à effets mixtes sont largement utilisés en pharmacologie, notamment pour l'analyse de données d'essais cliniques pour l'évaluation de nouveaux traitements médicamenteux. L'utilisation des modèles à effets mixtes est encouragée par le développement de méthodes statistiques spécifiques à ces modèles et par leur implémentation dans des logiciels tels que NONMEM et MONOLIX, devenus les logiciels de référence pour l'approche de population dans l'industrie pharmaceutique. Néanmoins, la nature des données recueillies dans les divers contextes expérimentaux requiert l'utilisation de modèles structuraux de plus en plus complexes et l'implémentation de méthodologies spécifiques est très attendue par les industriels. Dans cette thèse, nous nous sommes intéressés à des modèles à effets mixtes à dynamique markovienne, dont nous avons présenté quelques applications en biologie. Nous avons consacré le Chapitre 2 aux modèles de Markov cachés à effets

mixtes dont nous avons illustré l'utilisation dans l'analyse de nombres de crises d'épilepsie et qui peuvent plus généralement s'appliquer à la description de symptômes relatifs à des maladies chroniques. Dans le Chapitre 3, nous avons étudié des modèles de diffusion à effets mixtes dont l'emploi est approprié en pharmacocinétique. Nous avons proposé des méthodes d'estimation pour ces modèles complexes. Ces méthodes reposent sur l'algorithme SAEM, dont les propriétés pratiques sont attractives, en termes de rapidité et de précision des estimateurs. Pour adapter l'algorithme aux modèles de Markov cachés à effets mixtes et aux modèles de diffusion à effets mixtes, la clé réside dans le choix d'une procédure de calcul efficace des vraisemblances complètes $p(\mathbf{y}_i, \phi_i; \theta)$. De cette manière, nous avons considéré les processus latents comme des paramètres de nuisance plutôt que des données non observées du modèle, évitant la simulation à chaque itération de SAEM d'une réalisation de chaque processus individuel latent, difficile en pratique et coûteuse en temps de calcul.

Dans le Chapitre 2, nous avons combiné l'algorithme SAEM à l'algorithme de Baum-Welch pour estimer les paramètres de population dans les modèles de Markov cachés à effets mixtes. Nous avons également proposé des procédures spécifiques pour l'estimation des paramètres individuels et la reconstruction des trajectoires individuelles données par les processus markoviens latents. Ces procédures ont été implémentées sous forme de codes MATLAB dans des versions de travail de MONOLIX. Nous avons contrôlé les propriétés de ces deux algorithmes sur des jeux de données simulés de taille "réaliste" et pouvant contenir plusieurs effets aléatoires. Les performances de SAEM sont très satisfaisantes, et les simulations mettent en évidence de bonnes propriétés de l'estimateur du maximum de vraisemblance dans ces modèles. L'adaptation de l'algorithme SAEM aux modèles de Markov cachés à effets mixtes nous a permis d'analyser les données d'un essai clinique concernant un traitement anti-épileptique. Nous avons en effet proposé de modéliser des données de nombres de crises d'épilepsie au moyen de modèles de Markov cachés à effets mixtes à émissions Poissonniennes et avons montré que l'ajout d'une structure markovienne latente permettait une meilleure description des données de nombres de crises d'épilepsie que les modèles de Poisson à effets mixtes proposés dans des travaux antérieurs. De nombreux compléments à ces travaux appliqués sont envisageables. Une piste exploitable pour améliorer encore la description de ces données serait de définir le nombre quotidien de crises survenant dans chaque état de la maladie comme des variables aléatoires de loi de Poisson généralisée de manière à mieux prendre en compte la sous-dispersion de certaines données individuelles. A la fin du Chapitre 2, nous avons ajusté un tel modèle aux données de screening. Il serait intéressant d'utiliser ces modèles pour affiner l'analyse que nous avons proposée de l'essai clinique sur la Gabapentine. Une autre approche possible pour décrire la dynamique de l'épilepsie lorsque les observations de comptage ne sont pas régulièrement espacées dans le temps ou lorsque seuls les temps des événements de crises sont connus, serait de définir le nombre de crises d'épilepsie comme un processus de Poisson.

Dans le Chapitre 3, nous avons combiné l'algorithme SAEM au filtre de Kalman étendu pour estimer les paramètres de population dans les modèles de diffusion à effets mixtes à observations discrètes et bruitées. Certaines questions doivent encore être résolues concernant l'estimation dans les modèles de diffusion à effets mixtes via SAEM. En effet, d'après les premières simulations conduites sur un modèle simple inspiré du bolus, les propriétés de l'estimateur obtenu par SAEM pour certains paramètres de ces modèles reste discutable. Plus précisément, les paramètres de volatilité et de variance des erreurs de mesure sont sous-

estimés. Nous avons cependant constaté que le biais d'estimation avait tendance à diminuer lorsque la taille de l'échantillon devenait grande, en particulier lorsque le nombre d'observations par sujet devient grand. Nous n'avons pour le moment pas réussi à établir clairement l'origine de ce biais. Plusieurs explications sont envisageables. Il est d'abord possible que ce résultat soit causé par un défaut de la méthodologie elle-même. Les propriétés du filtre de Kalman étendu sont très discutées et il se peut que l'approximation des vraisemblances conditionnelles $p(\mathbf{y}_i|\phi_i)$ donnée par cette méthode soit trop grossière lorsque le nombre d'observations par sujet est petit, correspondant à des cas où les pas de temps entre observations successives sont relativement grands dans le cadre de nos simulations. Dans ce cas, il serait possible d'améliorer les propriétés de l'estimateur donné par l'algorithme SAEM en introduisant des temps intermédiaires aux temps d'observations pour affiner l'approximation des vraisemblances conditionnelles. Le biais mis en évidence par l'algorithme SAEM est peut-être uniquement dû aux propriétés théoriques de l'estimateur du maximum de vraisemblance. Ce point mériterait d'être vérifié par l'étude théorique des propriétés du maximum de vraisemblance dans des modèles de diffusion à observations discrètes et bruitées à taille d'échantillon fixée puis infinie. Ce type de résultat sera peut-être à relier aux propriétés de l'estimateur du maximum de vraisemblance dans les modèles linéaires à effets mixtes, l'EMV des paramètres de variance des effets aléatoires et erreurs de mesure de ces modèles étant biaisé à taille d'échantillon fixée, mais asymptotiquement non biaisé lorsque le nombre de sujets tend vers l'infini. Dans les modèles linéaires à effets mixtes, ce biais est corrigé en estimant les paramètres de variance à partir de la vraisemblance restreinte des observations, méthode connue sous le nom de méthode REML (*REstricted Maximum Likelihood*). Il serait alors intéressant d'un point de vue pratique de travailler sur le développement d'une nouvelle méthode d'estimation pour les modèles de diffusion à effets mixtes basée sur la vraisemblance restreinte pour les paramètres de volatilité et de variance des erreurs de mesure, en adaptant par exemple l'algorithme SAEM à une estimation type REML de ces paramètres. Néanmoins, l'étude théorique du maximum de vraisemblance dans les modèles à effets mixtes est de manière générale très complexe, la vraisemblance des observations et l'EMV pour les paramètres de population n'ayant généralement pas d'expression explicite.

La question des propriétés théoriques de l'estimateur du maximum de vraisemblance est abordée dans le Chapitre 3, où nous avons étudié les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans des modèles de diffusions à effets mixtes continûment observés et dont la fonction de drift dépend linéairement d'effets aléatoires gaussiens. Dans cette classe de modèles, la vraisemblance et l'estimateur du maximum de vraisemblance ont une expression explicite. Nous avons démontré la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance des paramètres de population. Ces propriétés sont étendues au cas d'observations discrètes et non bruitées de tels modèles. Ce résultat suppose les paramètres de la fonction de volatilité non aléatoires et connus. Cette condition est restrictive et non réaliste en pratique. Les propriétés asymptotiques de l'estimateur du maximum de vraisemblance mériteraient d'être étendues à des modèles de diffusion dont la volatilité dépend de paramètres inconnus et/ou d'effets aléatoires dont la loi dépend de paramètres inconnus. De même, il serait intéressant de généraliser cette étude à des processus de diffusions plus réalistes dont les fonctions de drift et de volatilité seraient des fonctions non nécessairement linéaires d'effets aléatoires, d'envisager des observations bruitées des processus de diffuions ou encore d'autres distributions pour les effets aléatoires qu'une distribution gaussienne.

Dans le Chapitre 4, nous avons quitté le cadre des modèles mixtes à dynamique markovienne et avons abordé le problème de sélection de modèles dans une approche populationnelle. Que ce soit en pharmacologie ou dans un autre domaine d'application, il est primordial pour proposer la meilleure description possible d'un phénomène donné de pouvoir comparer les degrés d'adéquation à l'échantillon de plusieurs modèles à effets mixtes de structures différentes. Dans la pratique, le critère BIC est utilisé à cet effet, mais la définition de la taille de l'échantillon intervenant dans l'expression de la pénalité ne s'impose pas clairement. Nous avons proposé d'éclaircir cette pratique en revenant aux fondements théoriques du BIC et en nous plaçant dans une situation de double asymptotique où le nombre de sujets et le nombre d'observations par sujet tendent vers l'infini. Nos premiers résultats encouragent l'utilisation de la version du BIC pénalisée par le logarithme du nombre de sujets, et valide le choix de la version du BIC implémentée dans MONOLIX. Des simulations supplémentaires sont encore nécessaires. Ce travail n'a pour seul intérêt que de guider l'utilisation du BIC par des arguments théoriques dans le cadre des modèles à effets mixtes. La comparaison de modèles reste un problème très complexe dans une approche populationnelle. Les questions principales sont la sélection de covariables et la sélection de covariance. Par sélection de covariance, nous entendons sélection du nombre d'effets aléatoires dans le modèle et du nombre de corrélations non nulles entre effets aléatoires. Par construction, le BIC permet de comparer des modèles à structure de covariance donnée. Nous avons d'ailleurs pu vérifier au fil de nos simulations que les critères de type BIC étaient inadaptés au problème du choix d'une structure de covariance. La suite logique de cette contribution serait de travailler à un critère de vraisemblance pénalisée capable de résoudre simultanément les problèmes de sélection de covariables et de sélection de covariance.

Les modèles étudiés dans cette thèse pourraient aisément être appliqués à d'autres problèmes que la dynamique de maladies chroniques ou la pharmacocinétique de population. Les méthodes que nous avons développées pour ces modèles devraient à l'avenir être intégrées au logiciel MONOLIX, facilitant ainsi leur diffusion et leur utilisation pratique dans de nombreux domaines.