N° d'ordre:

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DES UNIVERSITÉS PARIS-SUD XI ET
YAOUNDE I

Spécialité : Mathématiques

par

Cyprien MBOGNING TCHINDA

Inference dans les modèles conjoints et de mélange non-linéaires à effets mixtes.

Soutenue le 19 Décembre 2012 devant la Commission d'examen:

Mme.	Hélène	JACQMIN-GADDA	(Membre)
M.	Marc	LAVIELLE	(Directeur de thèse)
M.	Jean-Michel	POGGI	(Membre)
M.	Henri	GWET	(Co-Directeur de thèse)
M.	Jean-Marc	BARDET	(Rapporteur)
Mme.	Elisabeth	GASSIAT	(Membre)

Rapporteurs:

M.	Jean-Marc	Bardet
M.	Fabrice	Gamboa

Thèse préparée dans les institutions suivantes :

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

École Nationale Supérieure Polytechnique (ENSP)
LIMSS
Université de Yaoundé I
Cameroun

Thèse financée par :

Contrat doctoral attribué par
Inria-Saclay Île de France

Remerciements

Cette Thèse a été réalisée dans le cadre d'une Co-tutelle internationale entre l'Université de Paris 11 Orsay (France) via le laboratoire de Mathématiques et l'université de Yaoundé I (Cameroun) via l'École Nationale Supérieure Polytechnique. Elle s'inscrit aussi dans le cadre du projet STAFAV (Statistiques pour l'Afrique Francophone et Applications au Vivant) et est l'aboutissement d'un long et laborieux travail ayant nécessité l'assistance et le soutien de plusieurs personnes.

Je remercie mon directeur de thèse, le Professeur Marc Lavielle pour tout ce qui m'est arrivé de bien pendant cette Thèse. Il est le principal initiateur des questions traitées dans ce document et a été déterminant pour le financement. J'ai été bluffé par sa capacité à switcher de la théorie aux applications, avec comme principale philosophie que les méthodologies développées doivent être implémentées et mises à la disposition des utilisateurs dans un logiciel convivial. Merci encore pour ta patience et ta disponibilité.

Je remercie le Professeur Henri Gwet, Co-encadreur de ce travail et responsable du master STAFAV de Yaoundé dont je suis un récipiendaire. Merci pour ta disponibilité et tes multiples conseils.

J'exprime toute ma gratitude à Kevin Bleakley avec qui on a eut une collaboration fructueuse lors de cette dernière année de thèse.

Merci aux Professeurs Jean-Marc Bardet et Fabrice Gamboa d'avoir gracieusement accepté de rapporter cette Thèse.

Je suis honoré de compter les Professeurs Elisabeth Gassiat, Hélène Jaqmin-Gadda et Jean-Michel Poggi parmi les membres de mon jury de thèse.

Je remercie tous les membres du projet STAFAV, particulièrement le principal initiateur, le Professeur Didier Dacuhna-Castelle qui avec Elisabeth Gassiat m'ont mis en contact avec Marc Lavielle.

Je remercie tous les membres du LMO. Je pense notamment aux doctorants ou ex-doctorants : Emmanuel, Hatem, Jean-Patrick, Jairo, Maud, Wilson et tous les autres pour leur accueil et les multiples discussions. Je remercie aussi tous les membres de l'équipe POPIX de l'Inria, particulièrement les ingénieurs Hector, Jean-François et Kaelig pour toutes les discussions, pour la plupart liées à la programmation.

Je tiens à remercier tous les membres du laboratoire LIMSS de l'ENSP de Yaoundé pour toute leur collaboration. Je pense particulièrement au Docteur Eugène-Patrice Ndong Nguema.

Je remercie Valérie Lavigne de l'école doctorale qui a effectué la quasi totalité des procédures administratives me concernant à Orsay en mon absence. Je remercie aussi Katia Evrat de l'INRIA qui s'occupait de l'organisation de mes voyages en France.

Merci à tous mes ami(e)s pour leur soutien et surtout la patience dont ils ont fait preuve pendant toute la durée de cette thèse.

Je remercie enfin toute ma famille pour tout leur soutient, les sacrifices consentis et leur patience. J'ai une pensée particulière pour ma mère qui souhaitait avoir un docteur dans sa famille. J'espère que cette réalisation t'offrira un immense sourire quelques mètres sous terre...

Principales abréviations

BSMM : Between-Subject Model Mixtures

EM : Expectation Maximization.

SAEM : Stochastic Approximation EM.

MSAEM : Mixture SAEM.

SEM : Stochastic EM.

MLE : Maximum Likelihood Estimate.

MNLEM : Modèles Non Linéaires à Effets Mixtes.

NLMEM : Non Linear Mixed Effects Models.

MCMC : Markov Chain Monte Carlo.

PKPD : PharmacoKinetic-PharmacoDynamic.

REE : Relative Estmation Error.

RRMSE : Relative Root Mean Square Errors.

RTTE : Repeated Time-To-Event.

WSMM : Within-Subject Model Mixtures.

Résumé

Cette Thèse est consacrée au développement de nouvelles méthodologies pour l'analyse des modèles non-linéaires à effets mixtes, à leur implémentation dans un logiciel accessible et leur application à des problèmes réels. Nous considérons particulièrement des extensions des modèles non-linéaires à effets mixtes aux modèles de mélange et aux modèles conjoints. L'étude de ces deux extensions constitue l'essentiel du travail réalisé dans ce document qui peut donc être subdivisé en deux grandes parties.

Dans la première partie, nous proposons, dans le but d'avoir une meilleure maîtrise de l'hétérogénéité liée aux données sur des patients issus de plusieurs clusters, des extensions des MNLEM aux modèles de mélange.

- ◇ Les modèles de mélanges de distributions, utiles pour caractériser des distributions de population qui ne sont pas suffisamment bien décrites par les seules covariables observées. Certaines covariables catégorielles non observées définissent alors les composantes du mélange.
- ◇ Les mélanges de modèles inter-sujets supposent également qu'il existe des sous-populations de patients. Ici, différents modèles structurels décrivent la réponse de chaque sous-population et chaque patient appartient à une sous-population.
- ◇ Les mélanges de modèles intra-sujets supposent qu'il existe des sous-populations (de cellules, de virus ,...) au sein du patient. Différents modèles structurels décrivent la réponse de chaque sous-population mais la proportion de chaque sous-population dépend du patient.

Les algorithmes d'estimation dans les modèles de mélange tels que l'EM ne peuvent pas être appliqués de manière directe dans ce contexte. En effet, en plus de la structure de latence induite par les labels non observés des individus, on a aussi des paramètres individuels expliquant une partie de la variabilité non-observée qui ne sont pas observés. Les algorithmes populaires pour l'estimation des paramètres dans les MNLEM tels que SAEM sont confrontés dans certains cas à des difficultés pratiques dues particulièrement à la propriété de "Label-switching" inhérente aux modèles de mélange. Nous proposons dans cette Thèse de combiner l'algorithme EM, utilisé traditionnellement pour les modèles de mélanges lorsque les variables étudiées sont observées, et l'algorithme SAEM, utilisé pour l'estimation de paramètres par maximum de vraisemblance lorsque ces variables ne sont pas observées. la procédure résultante, dénommée MSAEM, permet ainsi d'éviter l'introduction d'une étape de simulation des covariables catégorielles latentes dans l'algorithme d'estimation. Cet algorithme est extrêmement rapide, très peu sensible à l'initialisation des paramètres et converge vers un maximum (local) de la vraisemblance. Cette méthodologie est désormais disponible sous Monolix , l'un des logiciels les plus populaires dans l'industrie pharmacologique, qui est libre pour les étudiants et la recherche académique. Nous avons ensuite effectué une classification non supervisée des données longitudinales réelles de charges virales sur des patients ayant le VIH, en considérant des mélanges de modèles structurels.

La seconde partie de cette Thèse traite de la modélisation conjointe de l'évolution d'un marqueur biologique au cours du temps et les délais entre les apparitions successives censurées d'un évènement d'intérêt. Nous considérons entre autres, les censures à droite, les multiples censures par intervalle d'évènements répétés. Nous proposons d'utiliser un MNLEM pour l'évolution temporelle du marqueur et un modèle de risque mixte, permettant de prendre en compte l'hétérogénéité due aux évènements répétés ainsi que la relation entre le processus longitudinal et le processus à risque, pour une utilisation effi-

ciente des informations disponibles dans les données. Les paramètres du modèle conjoint résultant sont estimés en maximisant la vraisemblance jointe exacte par un algorithme de type MCMC-SAEM. La matrice de Fisher est estimée par approximations stochastiques. La méthodologie proposée est générale et s'étend facilement aux modèles conjoints usuels (modèle linéaire mixte pour la variable longitudinale et modèle de risque pour un unique évènement ne pouvant se manifester q'une seule fois au cours de l'étude) et aux modèles d'évènements récurrents ou encore les modèles de fragilité. L'application de cette méthodologie aux jeux de données simulées montre que l'algorithme converge rapidement vers la cible avec une bonne précision. La méthode est ensuite illustrée sur des jeux de données réelles des patients ayant des cirrhoses biliaires primitives ainsi que des patients épileptiques. Cette méthodologie est désormais disponible sous MONOLIX.

Mots-clefs : Algorithme MSAEM, Algorithme SAEM, Censures par intervalle, Évènements répétés, Maximum de vraisemblance, Modèles conjoints, Modèles de mélange, Modèles mixtes, MONOLIX..

Inference in non-linear mixed effects joints and mixtures models.

Abstract

The main goal of this thesis is to develop new methodologies for the analysis of non linear mixed-effects models, along with their implementation in an accessible software and their application to real problems. We consider particularly extensions of non-linear mixed effects model to mixture models and joint models. The study of these two extensions is the essence of the work done in this document, which can be divided into two major parts.

In the first part, we propose, in order to have a better control of heterogeneity linked to data of patient issued from several clusters , extensions of NLMEM to mixture models.

- ◇ Mixture models of distributions are useful to characterize distributions of population that are not adequately described by only observed covariates. Some unobserved categorical covariates then define components of the mixture.
- ◇ Between-subject model mixtures also assume the existence of subpopulations of patients. Here, different structural models describe the response of each subpopulation and each patient belongs to a sub-population.
- ◇ Within-subject model mixtures assume that there exist subpopulations (Of cells, viruses, ...) within the patient. Different structural models describe the response of each sub-population, but the proportion of each subpopulation depends on the patient.

The standard estimation algorithms in mixture models such as EM can not be applied directly in this context. Indeed, in addition to the latency structure induced by unobserved individual labels, we also have individual parameters explaining part of the unobserved variability that are not observed. Popular algorithms for parameter estimation in NLMEM, such as SAEM, face in some practical cases several difficulties particularly due to the well-known "Label-switching" phenomenon, inherent to mixture models. We suggest in this Thesis to combine the EM algorithm, traditionally used for mixtures models when the variables studied are observed, and the SAEM algorithm, used to estimate the maximum likelihood parameters when these variables are not observed. The resulting procedure, referred MSAEM, allows to avoid the introduction of a simulation step of the latent categorical covariates in the estimation algorithm. This algorithm appears to be extremely fast, very little sensitive to parameters initialization and converges to a (local) maximum of the likelihood. This methodology is now available under the Monolix software, one of the most popular in the pharmacological industry, which is free for students and academic research. We then performed a classification of a longitudinal real data on viral loads for patients with HIV, by considering mixtures of three structural models.

The second part of this thesis deals with the joint modeling of the evolution of a biomarker over time and the time between successive appearances of a possibly censored event of interest. We consider among other, the right censoring and interval censorship of multiple events. We propose to use a NLMEM for the evolution of the marker and a risk mixed model, in order to take into account the heterogeneity due to repeated events and the relationship between the longitudinal process and the risk process. The parameters of the resulting joint model are estimated by maximizing the exact joint likelihood by using a MCMC-SAEM algorithm. The Fisher matrix is estimated by stochastic approximations. The proposed methodology is general and can easily be extended to the usual joint models (Linear mixed model for longitudinal and risk model for a single event that can occur only

once during the study) and models of recurring events or frailty models. The application of this methodology to the simulated data sets shows that the algorithm converges quickly to the target with high accuracy. As an illustration, such an approach is applied on real data sets on primary biliary cirrhosis and epileptic seizures. The proposed methodology is now available under MONOLIX.

Keywords : Interval censoring, Maximum likelihood, MSAEM algorithm, Joint models, SAEM algorithm, Mixed-effects models, Mixture models, MONOLIX, Repeated time-to-events.

Table des matières

Chapitre 1

Introduction Générale

Contents

1.1 contexte et problématique

Pour étudier des phénomènes biologiques complexes comme la pharmacocinétique d'un médicament, la dynamique d'un virus ou encore l'effet d'un traitement, l'industrie pharmaceutique fait de plus en plus appel à des approches de modélisation et de simulation. La modélisation de ces phénomènes complexes nécessite le développement et la mise en oeuvre de méthodologies de plus en plus performantes. En particulier, les approches de population ont pour objectif de modéliser la variabilité inter-sujet des données recueillies dans un essai clinique. L'outil statistique de référence pour la modélisation PKPD (pharmacocinétique-pharmacodynamique) est les modèles à effets mixtes.

Il existe néanmoins des situations d'hétérogénéité où la variabilité ne pourra être complètement expliquée par la seule variabilité inter-sujet. Une population de patients est généralement hétérogène par rapport à la réponse à un même traitement. Dans un essai clinique, les patients qui répondent, ceux qui répondent en partie et ceux qui ne répondent pas présentent des profils très différents. Par conséquent, la variabilité des cinétiques observées ne peut pas être uniquement expliquée de façon satisfaisante par la variabilité inter-patient de certains paramètres et les mélanges constituent une alternative pertinente dans de telles situations

Il existe à ce jour très peu de méthodes statistiques permettant d'estimer les paramètres dans les mélanges de modèles non-linéaires à effets mixtes par le maximum de vraisemblance.

Les méthodes implémentées dans le package nlme du logiciel R et dans le logiciel NONMEM sont basées sur la linéarisation de la vraisemblance. Ces méthodes présentent néanmoins des problèmes pratiques réels dans plusieurs situations (biais, forte influence des valeurs initiales, mauvaise convergence,...). Plus généralement, les propriétés théoriques des estimateurs obtenus par ces méthodes sont inexistantes dans plusieurs situations.

Des méthodes de type EM ont été proposées, utilisant des intégrations Monte-Carlo pour l'approximation de l'étape E. L'algorithme MCEM utilise ces intégrations Monte-Carlo pour déterminer la distribution des données non-observées conditionnellement aux observations, tandis que (De la Cruz-Mesia et al., 2008) proposent d'intégrer la distribution des données complètes pour déterminer la distribution marginale des observations dans chaque cluster. Néanmoins, ces méthodes peuvent s'avérer extrêmement lentes et pénibles à mettre en pratique quand le modèle structurel est complexe, ce qui est généralement le cas des modèles PKPD dont les modèles structurels sont généralement des solutions d'équations différentielles ordinaires ou stochastiques.

La première partie de cette Thèse a pour principal but de developer une méthodologie d'estimation par maximum de vraisemblance, constituant une alternative efficace aux méthodes précédemment citées, et ayant de bonnes propriétés théoriques, dans le cadre des modèles de mélange non-linéaires à effets mixtes.

Dans la plupart des études biomédicales (essais cliniques), on observe souvent de manière simultanée une variable longitudinale (progression d'un marqueur) ainsi qu'un délai jusqu'à la survenue d'un évènement terminal d'intérêt. La question scientifique la plus fréquente émanant de telles données est l'étude de la relation entre les deux variables, plus précisément, l'impact de l'une sur l'autre. A ce jour, plusieurs chercheurs ont manifesté un intérêt pour ce problème et il apparaît que la méthode optimale d'un point de vu modélisation est la modélisation conjointe des deux processus. On peut citer entre autre (Dafny and Tsiatsis, 1998; Tsiatis and Davidian, 2004; Hsieh et al., 2006) ou plus récemment encore (Rizopoulos, 2012b). La plupart des auteurs utilisent des modèles linéaires mixtes pour modéliser la variable longitudinale, ce qui constitue tout de même une importante restriction au niveau des applications, principalement en pharmacologie.

Dans plusieurs situations médicales, on rencontre des sujets pouvant expérimenter des évènements récurrents, tels que les crises d'épilepsie, les crises d'asthme, les tumeurs récurrentes, les hémorragies récurrentes, ... Les premiers travaux sur ce type de données considéraient juste le délai jusqu'à la première occurrence (ignorant la multiplicité) en utilisant un modèle de Cox (Cox, 1972). Cette approche n'utilise néanmoins pas toute l'information disponible dans les données, telles que la variation du temps entre les divers évènements, la durée du traitement, etc., et induit donc à des conclusions très peu fiables. C'est ainsi que, (Kelly and Jim, 2000; Nelson, 2003) ont montrés que le modèle de Cox est biaisé et inefficace dans le contexte typique des évènements répétés. Il est donc apparu une nécessité de considérer toutes les sources de variabilité dans le modèle pour des évènements récurrents. Par exemple, à cause du style de vie ou des traits génétiques, certains individus sont plus prédisposés à manifester leur première, deuxième,troisième, (etc) recurrence plus rapidement que les autres. Cette caractéristique introduit de ce fait une hétérogénéité entre les individus et produit une corrélation entre les sujets dans l'occurrence des évènements récurrents. Dans le but de modéliser cette hétérogénéité, des chercheurs ont proposés des modèles de fragilité, en ajoutant un paramètre aléatoire à la fonction de risque, jouant un role de covariable latente dans le modèle de risque. Parmi eux, (Liu et al., 2004; Huang and Liu, 2007; Rondeau et al., 2007) adoptent une fragilité de Gamma. Ce type de modèle est présenté en détail dans (Duchateau and Janssen, 2008).

Il existe aussi des situations médicales où on observe simultanément une variable longitudinale et des délais jusqu'aux observations répétées d'un évènement récurrent. Une question scientifique émanent de ce type de données pourrait être de modéliser l'impact de la variable longitudinale sur les occurrences répétées de l'évènement d'intérêt. On dénote cependant une absence de travaux liés à cette question.

Dans la deuxième partie de cette Thèse, nous proposons une modélisation conjointe de l'évolution d'un marqueur biologique via un modèle non linéaire à effets mixtes et les délais successifs des évènements répétés éventuellement censurés à droite ou par intervalle via un modèle de risque mixte. La méthodologie

d'estimation prend aussi en compte les modèles conjoints ayant un évènement ne pouvant se réaliser q'une seule fois lors de l'étude (évènement terminal), et les modèles constitués uniquement de données récurrentes.

1.2 Organisation de la thèse

Chapitre 2 : État de l'art

Le Chapitre 2 de cette Thèse présente un état de l'art global sur les modèles mixtes, les modèles de mélange et les modèles conjoints. Il est introductif et contient les outils utilisés ainsi qu' une présentation détaillée des problèmes abordés dans la suite de cette Thèse. Il contient entre autres – une présentation générale des modèles mixtes ainsi que les diverses méthodes utilisées à ce jour pour l'estimation des paramètres – une présentation des modèles de mélange "classiques" et quelques extensions ainsi que les méthodes d'estimation paramétriques – une présentation des modèles conjoints disponibles dans la littérature ainsi qu'une présentation détaillée de la problématique traitée par la suite.

Chapitre 3 : Inference in mixtures of non-linear mixed effects models

Le Chapitre 3 de cette Thèse est une version modifiée de l'article ayant le même titre, effectué en collaboration avec Marc Lavielle [1] [2], et soumis pour publication à Statistics and Computing.

Nous proposons dans ce Chapitre un modèle général de mélange incluant aussi bien les modèles de mélange de distribution non linéaires à effets mixtes que les mélanges de modèles non linéaires à effets mixtes, constituant des extensions de MNLEM aux modèles de mélange, ou encore des extensions de modèles de mélange aux MNLEM :

- ◇ Les modèles de mélanges de distributions peuvent être utiles pour caractériser des distributions de population qui ne sont pas suffisamment bien décrites par les seules covariables observées. Certaines covariables catégorielles non observées définissent alors les composantes du mélange.
- ◇ Les mélanges de modèles inter-sujets supposent également qu'il existe des sous-populations de patients. Ici, différents modèles structurels décrivent la réponse de chaque sous-population et chaque patient appartient à une sous-population.
- ◇ Les mélanges de modèles intra-sujets supposent qu'il existe des sous-populations (de cellules, de virus ,...) au sein du patient. Différents modèles structurels décrivent la réponse de chaque sous-population mais la proportion de chaque sous-population dépend du patient.

Nous proposons de combiner l'algorithme EM, utilisé traditionnellement pour les modèles de mélanges lorsque les variables étudiées sont observées, et l'algorithme SAEM, utilisé pour l'estimation de paramètres par maximum de vraisemblance lorsque ces variables ne sont pas observées. la procédure résultante, dénotée MSAEM, permet ainsi d'éviter l'introduction d'une étape de simulation des covariables catégorielles latentes dans l'algorithme d'estimation. Cet algorithme est extrêmement rapide, très peu sensible à l'initialisation des paramètres et converge vers un maximum (local) de la vraisemblance. Cette méthodologie est désormais disponible sous MONOLIX , l'un des logiciels les plus populaires dans l'industrie pharmacologique, qui est libre pour les étudiants et la recherche académique.

Chapitre 4 : Between-subject and within-subject model mixtures for classifying HIV treatment response

Le chapitre 4 de cette Thèse est un article ayant le même titre, effectué en collaboration avec Kevin Bleakley [1 2] et Marc Lavielle [1 2] et publié dans le journal Progress in Applied Mathematics.

Nous présentons dans ce chapitre une classification des données longitudinales réelles de charges virales sur des patients ayant le VIH. Nous considérons un mélange de modèles structurels pour classer les patients en 3 groupes : Ceux qui répondent totalement au traitement (caractérisés par une décroissance continue de la charge virale), ceux qui répondent partiellement (ou rebondissent) au traitement (caractérisés par une décroissance de la charge virale suivie d'une phase de rebond), et ceux qui ne répondent pas au traitement (aucune décroissante significative de la charge virale). Les paramètres du modèle sont estimés par l'algorithme SAEM et les patients sont ensuite classifiés par la règle du maximum à posteriori (MAP). Nous proposons aussi un mélange de modèle intra-sujet,qui suppose que chaque patient a une probabilité non nulle d'appartenir à chacune des 3 classes. Les 3 classes utilisées précédemment sont désormais internes à chaque patient. Ce dernier modèle est meilleur que le précédent en terme d'estimation de densité, mais ne permet néanmoins pas de classifier les patients. Cependant, il permet une étude approfondie des patients ayant des réponses atypiques (relativement aux 3 classes considérées).

Chapitre 5 : Joint modeling of longitudinal and repeated time-to-event data with maximum likelihood estimation via the SAEM algorithm.

Le Chapitre 5 de cette Thèse constitue la deuxième partie des travaux de Thèse et est issu d'une collaboration avec Kevin Bleakley [1 2] et Marc Lavielle [1 2].

Nous proposons dans ce Chapitre de modéliser de manière conjointe une ré-

ponse longitudinale en utilisant un modèle non linéaire à effets mixtes, et une suite de délais successifs jusqu'à un évènement récurrent en utilisant un modèle de risque mixte. Nous admettons des censures à droite et par intervalle pour les évènements successifs. Les paramètres du modèle conjoint résultant sont estimés en maximisant la vraisemblance jointe par un algorithme de type MCMC-SAEM. Cette méthodologie est générale et s'étend facilement aux modèles conjoints usuels et aux modèles d'évènements récurrents ou encore les modèles de fragilité. L'application de cette méthodologie aux jeux de données simulées montre que l'algorithme converge rapidement vers la cible avec une bonne précision. Une application à deux jeux de données réelles est proposée. Le premier jeu de données est constitué de patients atteints de cirrhoses biliaires primitives ; le second de patients épileptiques. Ce Chapitre constitue une avancée importante dans l'état de l'art sur les modèles conjoints et la méthodologie résultante est désormais disponible via le logiciel MONOLIX .

1. Laboratoire de Mathématiques d'Orsay (LMO), Bat 425, 91405 Orsay cedex, France.
2. Inria Saclay, POPIX team.

Chapitre 2

État de l'art

Contents

Dans ce chapitre, nous dressons un état de l'art général sur les modèles mixtes, les modèles de mélange et les modèles conjoints. La plupart des outils abordés dans cette Thèse sont présentés dans ce Chapitre. La première section est consacrée à une présentation générale des modèles non-linéaires à effets mixtes ainsi que les méthodologies utilisées à ce jour pour l'estimation des paramètres. La seconde traite des modèles de mélange (mélanges classiques ainsi que des extensions) et présente les limites des approches usuelles, constituant la principale problématique de la première partie de cette thèse. La dernière section traite des modèles conjoints et aboutit à la problématique traitée dans la deuxième partie de cette thèse.

2.1 Modèles non-linéaires à effets mixtes

Durant la dernière décennie, on a observé une explosion de travaux relatifs aux modèles mixtes et leur applications. Ces modèles se sont avérés être très efficaces dans le cadre des données répétées (longitudinales) et dans plusieurs domaines d'application tels que la biologie, la pharmacologie, l'agronomie, les sciences sociales. Les données répétées sont des données dans lesquelles plusieurs individus sont soumis à de multiples mesures temporelles ou spatiales. L'analyse de ce type de données requiert des méthodes statistiques adaptées dans la mesure où, les données de chaque patient sont supposées indépendantes les unes des autres, autorisant tout de même une corrélation dans le temps sur les observations d'un même sujet. Ainsi, une méthode d'analyse de ce type de données nécessite de reconnaître et d'estimer divers types de variabilité : Une variabilité entre les individus, dite inter-individuelle, une variabilité des paramètres d'un même sujet au cours du temps, dite intra-individuelle et une variabilité résiduelle représentant l'écart par rapport au modèle utilisé. En général, les variabilités intra-individuelle et résiduelle sont confondues. Ces modèles mixtes permettent de plus d'évaluer la distribution des paramètres du système biologique au sein de l'ensemble de la population en considérant dans le modèle statistique les paramètres individuels comme des variables aléatoires (effets aléatoires) définies à travers des paramètres de population. Ainsi, en tenant compte de la nature des relations entre la variable réponse et les variables explicatives, on distingue comme dans le cadre classique, trois catégories de modèles mixtes :

- ◇ Le modèle linéaire mixte qui fut introduit par (Laird and Ware, 1982) est l'un des plus utilisé pour étudier l'évolution d'un critère quantitatif continu au cours du temps en considérant une relation linéaire en les paramètres entre des variables explicatives et la variable réponse. Nous ne développerons pas d'avantage ces modèles par la suite et les lecteurs pourront se référer aux multiples ouvrages disponibles parmi lesquels (Verbeke and Molenberghs, 2000; Vonesh and Chinchilli, 1997; Jiang, 2007)
- ◇ lorsque la linéarité est définie via une fonction de lien, on parle de modèle linéaire mixte généralisé, utilisé pour des réponses quantitatives, qualitatives ou discrètes (Gilmour et al., 1985; Breslow and Clayton, 1993; Jiang, 2007).

◇ Il n'est pas rare de trouver une relation non-linéaire en les paramètres entre des variable explicatives (le temps en général) et une variable expliquée (concentration d'un médicament). On parle alors de modèle non-linéaire à effets mixtes (MNLEM). Ces derniers représentent un outil pour l'analyse des données répétées dans lesquelles la relation entre les variables explicatives et la variable réponse peut être modélisée comme une unique fonction non-linéaire, permettant aux paramètres de différer entre les individus. Ces modèles sont généralement utilisés en pharmacologie où des statisticiens sont impliqués à tous les niveaux pour l'évaluation des données collectées au cours des essais thérapeutiques et pour aider à planifier les études suivantes en fonction des résultats obtenus. Ceci passe par une analyse minutieuse de l'ensemble des données physiologiques (concentrations, marqueurs biologiques, effets pharmacologiques, effets indésirables) ainsi que leur évolution au cours du temps et leur variabilité entre les patients afin de mieux comprendre l'ensemble de la relation dose réponse, nécessaire pour la planification des essais cliniques suivants prenant mieux en compte les sources de variabilité et d'incertitude, ceci via des simulations. Ces modèles sont utilisés pour une grande variété de données parmi lesquelles les données continues, de comptage, catégorielles ou encore de survie.

La suite de cette section sera consacrée au formalisme mathématique des MNLEM ainsi que les diverses méthodes d'estimation de paramètres présentes dans la littérature.

2.1.1 Modèles et notations

Les MNLEM peuvent être définis comme des modèles hiérarchiques. A un premier niveau, les observations de chaque individu peuvent être décrites par un modèle de régression paramétrique propre à l'individu, appelé généralement *modèle structurel*, défini de manière identique moyennant un ensemble de paramètres individuels inconnus fluctuant entre les individus. Au seçond niveau hiérarchique, chaque ensemble de paramètres individuels est considéré comme provenant d'une distribution paramétrique inconnue.

Le modèle se présente donc de la manière suivante dans le cadre des observations continues :

$$y_{ij} = f\left(x_{ij}, \varphi_i\right) + g\left(x_{ij}, \varphi_i, \theta_y\right) \varepsilon_{ij},\ 1 \leq i \leq N,\ 1 \leq j \leq n_i \tag{2.1}$$

avec

◇ $y_{ij} \in \mathbb{R}$ la $j$ème observation faite sur le sujet i.

◇ N le nombre d'individus présents dans l'étude, et n_i le nombre d'observations faites sur l'individu i.

◇ x_{ij} un vecteur de variables de régressions (contenant en général les temps d'observation t_{ij} dans le cadre des données longitudinales).

◇ φ_i un vecteur d-dimensionnel de paramètres individuels liés à l'individu i. On suppose que les φ_i sont générés par une même distribution de population

définie comme des transformations de gaussiennes :

$$\varphi_i = h\left(\mu, c_i, \eta_i\right) \tag{2.2}$$

- ⋆ h une fonction décrivant le modèle de covariable,
- ⋆ $c_i = (c_{ik}\ ;\ 1 \leq k \leq K)$ un vecteur de K covariables connues,
- ⋆ $\eta_i \sim_{i.i.d} \mathcal{N}(0, \Sigma)$ un vecteur inconnu d'effets aléatoires, Σ étant la matrice de variance-covariance inter-individuelle.
- ⋆ μ un vecteur inconnu d'effets fixes

◇ $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$ représentant les erreurs résiduelles qui sont considérées indépendantes des paramètres individuels φ_i,

◇ $f(\cdot)$ et/ou $g(\cdot)$, des fonctions non-linéaires des φ_i, la fonction f définissant le modèle structurel et la fonction g le modèle résiduel.

◇ $\theta = (\theta_y, \mu, \Sigma) \in \Theta \subset \mathbb{R}^p$ le vecteur de paramètres du modèle appelés paramètres de population.

Pour les autres types de données, la présentation est semblable à la précédente avec tout de même une définition adéquate pour le modèle des observations. On considère alors que la loi des observations y_i conditionnellement aux paramètres individuels φ_i est connue, et la densité de probabilité donnée par $l\left(y_i|\varphi_i; \theta_y\right)$.

La vraisemblance des observations sur un individu i est donnée par :

$$l\left(y_i; \theta\right) = \int_{\mathbb{R}^d} p\left(y_i, \varphi_i; \theta\right)\ d\varphi_i, \tag{2.3}$$

$p\left(y_i, \varphi_i; \theta\right)$ étant la vraisemblance des données complètes sur l'individu i. Les fonctions de régression f et ou g étant non-linéaires, la vraisemblance des données observées n'a pas d'expression analytique. L'estimation des paramètres par maximisation de la vraisemblance des observations ne pourrait donc pas s'effectuer de manière directe. On dénote ainsi plusieurs problèmes liés à l'utilisation des MNLEM : l'estimation des paramètres de population θ avec une mesure de l'erreur commise, le calcul de la vraisemblance des observations $l\left(y; \theta\right)$ pour des besoins éventuels de sélection de modèles ou encore de tests d'hypothèses, l'estimation des paramètres individuels,...

La section suivante sera consacrée aux méthodes d'estimations des paramètres des MNLEM présentes dans la littérature

2.1.2 Méthodes d'estimation pour les MNLEM

Comme nous le signalions dans la section précédente, la vraisemblance des observations $l\left(y_i; \theta\right)$ n'a pas d'expression analytique dans les MNLEM. Par conséquent, dans le cadre des MNLEM, on rencontre des méthodes d'estimations des paramètres basées sur des approximations du modèle initial (via des linéarisations ou encore des approximations de la vraisemblance par des techniques Monte Carlo ou des algorithmes numériques), les méthodes d'estimation Bayésiennes et les méthodes basées sur le maximum de vraisemblance du modèle original. Les premières maximisent plutôt un modèle approché.

Méthodes basées sur une approximation du modèle

Plusieurs algorithmes basés sur des approximations du modèle ont été proposés, présentant des estimateurs minimisant un critère sur le modèle approché. Les plus utilisés sont les algorithmes itératifs First Order (FO) et la plus générale, First Order Conditional Estimate (FOCE), développés par (Beal and Sheiner, 1982) et (Lindstrom and Bates, 1990) respectivement. Ce dernier s'effectue en deux étapes. La première étape consiste à une estimation des paramètres individuels $\hat{\varphi}_i$ par une méthode de maximum à posteriori maximisant la loi conditionnelle $p(\cdot|y;\theta)$ en utilisant la Formule de Bayes. (Pinheiro and Bates, 1995) montreront plutard que cette étape correspond à la minimisation d'un critère pénalisé non-linéaire de moindres carrés via quelques itérations de l'algorithme de Newton-Raphson. La seconde étape consiste en une expansion de Taylor d'ordre 1 de la fonction définissant le modèle structurel au voisinage des paramètres individuels précédemment estimés. Cette opération permet d'avoir une expression analytique de la vraisemblance des observations du modèle approché. Les paramètres de populations sont ensuite actualisés en maximisant la pseudo-vraisemblance obtenue par un algorithme de Newton-Raphson. Les méthodes FO et FOCE sont implémentées dans la fonction nlme de Splus et R, et dans le logiciel NONMEM. On distingue aussi entre autres, les méthodes numériques basées sur une approximation de Laplace ou une quadrature de Gauss de la vraisemblance des observations (Wolfinger, 1993; Vonesh, 1996). Les paramètres de populations sont ensuite obtenus en maximisant la vraisemblance approchée. Ces méthodes sont mises en oeuvre dans la procédure NLMIXED du logiciel SAS.

On note cependant que, pour toutes les méthodes sus-citées, aucun résultat théorique sur la convergence vers un maximum de vraisemblance n'est établi à ce jour. (Vonesh, 1996) donne un exemple pour lequel les estimateurs issus des algorithmes de linéarisation FO et FOCE sont inconsistant dès que le nombre d'observation par sujet n_i croît moins vite que le nombre de sujet N (cette hypothèse est la plus rencontrée dans la pratique des modèles mixtes où on considère généralement le nombre d'observations par sujet borné). Des comportements similaires sont présentés par (Ge et al., 2004) dans des cas d'approximation de la vraisemblance par des fonctions splines, lorsque la variance des effets aléatoires est trop grande. Les méthodes basées sur l'approximation de Laplace ou la quadrature de Gauss quant à elles souffrent, comme toutes les méthodes d'approximation numériques d'intégrales, d'un problème lié à la dimension de l'espace d'intégration, et donc dans le cas présent de la dimension des effets aléatoires.

Il est donc apparu un besoin réel de developer des méthodes effectuant le maximum de vraisemblance du modèle original, et non pas celui d'un modèle approché, et qui possèdent des propriétés de convergence ou de consistence sous des hypothèses réalistes. Comme alternative aux précédentes méthodes, les méthodes Bayésiennes présentent un cadre rigoureux et flexible pour l'estimation des paramètres dans les MNLEM. Ces dernières se heurtent tout de même au fait que la distribution à posteriori (proportionnelle au produit d'une distribution à priori introduite sur les paramètres de populations notée $\pi(\theta)$ avec la vraisemblance des

observations $l(y;\theta)$) est difficile à calculer dans le cadre des MNLEM. Cette difficulté étant engendrée par l'absence d'expression analytique de la vraisemblance des observations et des constantes de normalisation difficiles à calculer. Néanmoins, les méthodes Monte Carlo par Chaînes de Markov (MCMC) permettent de contourner ces difficultés. Le lecteur intéressés par des méthodes d'estimation Bayésiennes dans le cadre des MNLEM pourront se référer entre autres aux articles de (Racine-Poon, 1985; Wakefield et al., 1994; Wakefield, 1996). Les méthodes MCMC, à l'origine dévellopées dans un contexte Bayésien, sont de plus en plus utilisées dans des approches fréquentistes. Avant de présenter les approches fréquentistes de maximisation de la vraisemblance exacte du modèle, nous exposerons dans La section suivante les méthodes MCMC les plus populaires permettant de réaliser des simulations suivant la loi à posteriori.

Méthodes MCMC de Simulation suivant la loi à posteriori

Les algorithmes tels que ceux de Metropolis-Hastings ou encore ceux du Gibbs-sampling sont les algorithmes de calcul d'inférence les plus efficaces fondés sur les méthodes MCMC. Ils ont à cet effet provoqués des développements spectaculaires récents de la statistique bayésienne. On appelle algorithme Monte Carlo par Chaînes de Markov (MCMC) pour une loi de probabilité donnée toute méthode produisant une chaîne de markov ergodique de loi stationnaire la dite loi (Robert, 1996). Les possibilités d'application des méthodes MCMC pour l'estimation dans des modèles à données manquantes sont nombreuses. Dans le cadre précis des modèles non-linéaires à effets mixtes, la distribution à posteriori est en général inconnue et la simulation suivant cette distribution ne peut se faire de manière directe. On peut donc envisager de générer une chaîne de Markov ergodique dont la loi stationnaire serait celle du posterior. Nous décrirons dans la suite, de manière succincte les deux méthodes MCMC les plus populaires, à savoir l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs, pour la simulation de la loi à posteriori.

Échantillonneur de Metropolis-Hastings

En général, la loi cible π est une loi à posteriori obtenue suite à l'application de la formule de Bayes si bien qu'elle n'est connue qu'à une constante multiplicative près. La méthode de Metropolis-Hastings étant historiquement la première des méthodes MCMC, se fonde sur le choix d'une distribution de transition instrumentale conditionnelle $q\left(\tilde{\varphi}|\varphi^{(k-1)}\right)$ qui est une généralisation de la distribution indépendante $q(\tilde{\varphi})$ de l'algorithme d'acceptation-rejet. Elle jouit de la propriété remarquable de n'imposer que peu de limitations théoriques au choix de la fonction d'exploration q. Cependant, les comportements pratiques et notamment la rapidité d'atteinte de l'état limite ergodique doivent être considérés avec attention car ils dépendent fortement du choix de la loi instrumentale q. Il existe donc des lois instrumentales de transition pour lesquelles la convergence est extrêmement

lente et donc inutilisable en pratique. Le choix de cette loi instrumentale est donc fondamental pour l'atteinte de la loi cible en un temps "raisonnable". C'est ainsi que diverses formes de lois instrumentales ont été utilisées dans la littérature pour des situations précises et pour lesquelles la probabilité d'acceptation du candidat $\tilde{\varphi}$

$$\alpha\left(\varphi^{(k-1)}, \tilde{\varphi}\right) = \min\left(1, \frac{p\left(\tilde{\varphi}|y\right) q\left(\varphi^{(k-1)}|\tilde{\varphi}\right)}{p\left(\varphi^{(k-1)}|y\right) q\left(\tilde{\varphi}|\varphi^{(k-1)}\right)}\right)$$

se présente sous diverses formes spécifiques. les lois instrumentales les plus couramment utilisées sont les suivantes :

- $q\left(\tilde{\varphi}|\varphi^{(k-1)}\right) = q\left(\tilde{\varphi}\right)$. Le tirage du candidat se fait indépendamment du point de départ (ou précédent) $\varphi^{(k-1)}$(comme dans le cas de l'algorithme d'acceptation rejet). la probabilité d'acceptation du candidat se réduit en quelque sorte en un rapport de marginales et s'exprime comme suit dans le cas où on a par exemple comme loi instrumentale la loi à priori, $q\left(\tilde{\varphi}|\varphi^{(k-1)}\right) = p\left(\tilde{\varphi}\right)$

  $$\alpha\left(\varphi^{(k-1)}, \tilde{\varphi}\right) = \min\left(1, \frac{p\left(y|\tilde{\varphi}\right)}{p\left(y|\varphi^{(k-1)}\right)}\right) \tag{2.4}$$

  Dans ce cas, lorsque la loi instrumentale q est proche de la loi cible π, la convergence est rapide. Cet algorithme est néanmoins sensible dans certain cas aux valeurs initiales et aux états absorbants de la chaîne. Il est donc recommandé de ne pas l'utiliser seul.
- $q\left(\tilde{\varphi}|\varphi^{(k-1)}\right) = q\left(\tilde{\varphi} - \varphi^{(k-1)}\right)$ c'est à dire une marche aléatoire homogène ($\tilde{\varphi} = \varphi^{(k-1)} + \varepsilon$, ε étant généré par des tirages indépendants d'une loi fixée facile à simuler). le choix d'une fonction q symétrique dans ce cas à pour conséquence une définition plus simple de la probabilité d'acceptation qui se réduit en un rapport de densité de données complètes (généralement connues). Les marches aléatoires les plus utilisées dans la littérature sont les marches aléatoires gaussiennes et les marches aléatoires uniformes. Dans le cas précis des marches aléatoires gaussiennes, une grande attention doit être portée à la valeur de la variance car dans le cas d'une loi cible bi-modale par exemple, les valeurs trop faibles de variance ne permettent pas de visiter tous les modes et avec une variance trop forte, il y a systématiquement un risque de rejet. De manière générale, on considère pour les marches aléatoires un paramètre d'échelle κ qui doit être calibré de manière convenable. Si κ est trop grand, la plupart des candidats seront rejetés. Si par contre κ est très petit, la fenêtre d'exploration des éventuels candidats est fine et induit à un lent déplacement de la chaîne paramétrique. Lorsque la dimension de la chaîne est petite, différents auteurs (Gilks et al., 1996; Roberts and Rosenthal, 2001) recommandent d'adapter ce paramètre d'échelle afin d'assurer un taux d'acceptation de 30%. La probabilité d'acceptation est réduite à une expression simple et donnée par :

  $$\alpha\left(\varphi^{(k-1)}, \tilde{\varphi}\right) = \min\left(1, \frac{p\left(y, \tilde{\varphi}\right)}{p\left(y, \varphi^{(k-1)}\right)}\right) \tag{2.5}$$

Dans le but d'accélérer la convergence vers la loi cible, on peux envisager l'utilisation successive de plusieurs lois instrumentales. Cette procédure est utilisée dans les Chapitres 3 et 5 de cette Thèse.

Échantillonneur de Gibbs

Le principe des échantillonneurs de Gibbs est de substituer la simulation sur un espace de dimension d (par exemple), par plusieurs simulations sur des espaces ayant des dimensions réduites. Posons $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_d$ l'espace contenant la variable à simuler de sorte que tout élément $\psi \in \mathcal{D}$ s'écrive en bloc $\psi = (\psi_1, \ldots, \psi_d)$. Notons $\psi_{(-i)}$ le vecteur $(\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_d)$ constitué de tous les blocs à l'exception du bloc i.

Pour mettre en oeuvre l'échantillonneur de Gibbs pour la simulation de ψ, on suppose que l'on sait simuler sous les lois conditionnelles du bloc i sachant les autres i.e. sous les lois π_i sur $\mathcal{D}_i$ données par

$$\pi_i\left(\psi_i|\psi_{(-i)}\right) = \frac{\pi\left(\psi\right)}{\int\limits_{\mathcal{D}_i} \pi\left(\psi_1, \ldots, \psi_{i-1}, u, \psi_{i+1}, \ldots, \psi_d\right) du}.$$

l'itération k de l'échantillonneur de Gibbs consiste donc à réaliser l'étape suivante : Pour $i = 1, \ldots, d$, tirer $\psi_i^{(k)}$ suivant $\pi_i\left(.|\psi_1^{(k)}, \ldots, \psi_{i-1}^{(k)}, \psi_{i+1}^{(k-1)}, \ldots, \psi_d^{(k-1)}\right)$.

La convergence de cet algorithme et l'ergodicité géométrique ou uniforme de la chaîne simulée sont assurées, par exemple sous des conditions de minoration proposées par (Tierney, 1994).

Remark 2.1.1. *On rencontre souvent des situations où il existe $i \in \{1, \ldots, d\}$ tel que soit la constante de normalisation n'est pas évidente à calculer, soit la la loi $\pi_i\left(\psi_i|\psi_{(-i)}\right)$ est entièrement connue mais pas usuelle pour les simulations directes. Si on se retrouve dans l'un des deux cas précédent, on peut envisager l'utilisation d'un algorithme hybride Gibbs/Metropolis-Hastings qui consiste à utiliser au cours de l'échantillonneur de Gibbs l'algorithme de Metropolis-Hastings pour corriger les incertitudes répertoriées.*

Cet algorithme hybride est utilisé au chapitre 4 de cette thèse dans le cadre des modèles de mélange non-linéaires à effets mixtes, pour simuler les paramètres individuels non-observés.

La section suivante présente les méthodes d'estimation de paramètres dans les MNLEM par maximisation de la vraisemblance des observations, utilisant pour la plupart les méthodes MCMC précédemment décrites.

Estimation par maximum de vraisemblance

La vraisemblance des observations n'ayant pas d'expression analytique et les paramètres individuels n'étant pas observés, plusieurs chercheurs considèrent dorénavant que le problème d'estimation des paramètres dans les MNLEM est équivalent à un problème d'estimation dans des modèles à données incomplètes. Les

algorithmes les plus utilisés dans ce cadre sont des algorithmes de type Newton-Raphson ou encore de type Expectation-Maximisation. La suite de cette section est consacrée à une présentation de ces algorithmes ainsi que d'éventuelles extensions nécessaires pour l'estimation dans les MNLEM.

Estimations basées sur l'algorithme de Newton-Raphson

L'algorithme de Newton-Raphson constitue une méthode classique d'estimation par maximum de vraisemblance, ayant un structure itérative reposant sur la résolution d'une équation de score. Notons $\mathcal{L}(y;\theta)$ la log-vraisemblance des observations. Notons également $S(\theta) = \frac{\partial \mathcal{L}(y;\theta)}{\partial \theta}$ et $H(\theta) = \frac{\partial^2 \mathcal{L}(y;\theta)}{\partial \theta \partial \theta'}$ respectivement les fonctions de score la vraisemblance et hessienne de la log-vraisemblance. L'emv est obtenu de manière itérative comme la solution de l'équation $S(\theta) = 0$. L'itération d'ordre k de l'algorithme de Newton-Raphson actualise les paramètres de la manière suivante :

$$\theta^{(k)} = \theta^{(k-1)} + \left(H\left(\theta^{(k-1)}\right)\right)^{-1} S\left(\theta^{(k-1)}\right).$$

La vraisemblance des observations dans les MNLEM n'admettant pas d'expression analytique, il en est de même pour les fonctions score et hessienne de la vraisemblance. Cet algorithme nécessite donc des modifications pour une application aux MNLEM. Des versions stochastiques adaptées aux problèmes à données incomplètes ont été proposées, utilisant des relations liant les fonctions score et hessienne de la vraisemblance des observations à leur homologues pour la vraisemblance des données complètes (Louis, 1982). D'après le principe (Louis, 1982), on a les relations suivantes :

$$S(\theta) = \mathbb{E}\left(\frac{\partial \mathcal{L}(y,\varphi;\theta)}{\partial \theta} | y, \theta\right) \tag{2.6}$$

$$H(\theta) = \mathbb{E}\left(\frac{\partial^2 \mathcal{L}(y,\varphi;\theta)}{\partial \theta \partial \theta'} | y, \theta\right) + \mathrm{Var}\left(\frac{\partial \mathcal{L}(y,\varphi;\theta)}{\partial \theta} | y, \theta\right). \tag{2.7}$$

Les intégrales présentes dans les fonctions score (2.6) et hessienne (2.7) à chaque itération sont déterminées par des approximations empiriques de Monte-Carlo, basées sur un échantillon simulé de données non observées. Cette opération induit à l'algorithme Monte-Carlo Newton-Raphson (MC-NR) proposé par (Tanner, 1996; McCulloch, 1997).

Cet algorithme a néanmoins besoin d'un nombre important de simulations à chaque itération de Newton-Raphson pour converger. Des alternatives ont été proposées permettant de réduire la lourdeur numérique, et son basées essentiellement sur des approximations stochastiques (Robbins and Monro, 1951) de (2.6) et (2.7). Le lecteur pourra se référer à (Gu and Kong, 1998) pour une version stochastique de l'algorithme de Newton-Raphson. (Gu and Zhu, 2001) proposent de combiner la version stochastique de Newton-Raphson avec un algorithme MCMC permettant de simuler les données non observées dans les cas où la simulation n'est pas directe.

Estimations basées sur l'algorithme EM

En considérant le problème d'estimation dans les MNLEM comme un problème d'estimation à données incomplètes (les paramètres individuels inconnu étant considérés comme des données manquantes du modèle), l'algorithme Expectation Maximization (EM) de (Dempster et al., 1977) devient un candidat crédible. Il constitue la principale alternative à l'algorithme de Newton-Raphson pour l'estimation des paramètres dans des modèles à données incomplètes. La vraisemblance des observations n'ayant pas d'expression analytique, l'EM repose sur la maximisation d'un nouveau critère basé sur la vraisemblance des données complètes (elle est connue explicitement) via des itérations successives. Le critère en question étant l'espérance de la log-vraisemblance des données complètes ou augmentées (y, φ) par rapport à la distribution des données non-observées φ sachant observations ou données incomplètes y et la valeur courante du paramètre θ' :

$$Q(\theta, \theta') = \mathbb{E}\left(\mathcal{L}(y, \varphi; \theta) \,|y, \theta'\right).$$

L'algorithme EM proposé par (Dempster et al., 1977) alterne les deux étapes suivantes :

- ◇ **Étape E** dite *Expectation* : Sachant la valeur courante du paramètre $\theta^{(k)}$ à l'itération k, la phase E consiste e n la détermination de la fonction

$$Q\left(\theta, \theta^{(k)}\right) = \mathbb{E}\left(\mathcal{L}(y, \varphi; \theta) \,|y, \theta^{(k)}\right).$$

- ◇ **Étape M** dite *Maximization* : La valeur courante du paramètre est actualisée en maximisant la fonction obtenue à l'étape E par rapport à θ, soit

$$\theta^{(k+1)} = \arg\max_{\theta} Q\left(\theta, \theta^{(k)}\right).$$

La principale relation liant la maximisation de la log-vraisemblance des observations $\mathcal{L}(y; \theta)$ à la maximisation de $Q(\theta, \theta')$ est donnée dans la proposition suivante :

Proposition 2.1.1. *Si* $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)}, \theta^{(k+1)}, \ldots,$ *constituent une suite d'itérations EM, on a la relation suivante :*

$$\mathcal{L}\left(y; \theta^{(k+1)}\right) \geq \mathcal{L}\left(y; \theta^{(k)}\right), \ \forall k.$$

Cette propriété de l'algorithme est fondamentale car elle garantit à l'utilisateur une bonne évolution des valeurs de la log-vraisemblance des observations. En effet, toute maximisation de Q engendre une maximisation de la vraisemblance des observations. Ainsi, lorsque la maximisation de Q est plus simple que celle de la vraisemblance des observations, des maximisations successives de Q peuvent permettre d'atteindre un maximum de la vraisemblance.

Des résultats de convergence de l'EM ont été proposés par (Dempster et al., 1977; Wu, 1983). (Delyon et al., 1999) présentent un résultat de convergence de

l'EM avec des hypothèses plus simples, dans le cadre des modèles appartenant à la famille exponentielle des modèles. En dépit de son caractère intuitif et des bonnes propriétés liées à l'algorithme EM, il est néanmoins sujet à plusieurs difficultés d'ordre pratique. Il existe ainsi des situations où, la quantité $Q(\theta, \theta')$ ne possède pas d'expression analytique, la maximisation de $Q(\theta, \theta')$ est extrêmement complexe, la convergence de l'algorithme est assez lente. Néanmoins, les problèmes persistants sont liés à la détermination de $Q(\theta, \theta')$. C'est d'ailleurs le cas dans les MNLEM où la loi à posteriori $l(\varphi|y, \theta')$ des paramètres individuels est inconnue, rendant impossible le calcul direct de $Q(\theta, \theta')$. Des extensions stochastiques de l'EM ont été proposées pour remédier au problème.

L'Algorithme MCEM

(Wei and Tanner, 1990) proposent un algorithme MCEM (Monte-Carlo EM) permettant d'approcher la quantité $Q(\theta, \theta')$ par une méthode Monte-Carlo. A chaque itération, un nombre L de variables aléatoires $\varphi^{(k)}$ est généré à partir de la loi à posteriori $l\left(\varphi|y, \theta^{(k)}\right)$, la fonction Q est ensuite approchée par une moyenne empirique

$$Q\left(\theta, \theta^{(k)}\right) \approx \frac{1}{L} \sum_{k=1}^{L} \mathcal{L}\left(y, \varphi^{(k)}; \theta\right).$$

On dénote cependant à ce jour une quasi-absence de résultats théoriques de convergence de cet algorithme. Il peut avoir des problèmes numériques tels que une convergence lente ou inexistante. Le nombre L de réplications joue un role important pour la convergence, mais son choix reste un problème ouvert. En général, cet algorithme nécessite un nombre de réplications assez important à chaque itération et principalement lors des dernières itérations, ce qui rend la méthode assez lourde numériquement avec des temps d'exécution très élevés.

La simulation de φ suivant la loi à posteriori $l(\varphi|y, \theta)$ n'étant pas directe dans les MNLEM, (Walker, 1996; Wu, 2004) proposent de combiner l'algorithme MCEM avec une procédure MCMC permettant de simuler les paramètres individuels φ. Ils rapportent les mêmes problèmes de convergence numérique et ne proposent aucun résultat théorique.

L'Algorithme SAEM

Une solution aussi bien théorique que numérique aux problèmes sus-cités repose sur une approximation stochastique de l'étape E et a été proposée par (Delyon et al., 1999). L'algorithme SAEM (Stochastic Approximation EM) proposé par (Delyon et al., 1999) incorpore une étape de simulation et une étape d'approximation à l'algorithme EM. Partant d'une position initiale $\theta^{(0)}$, une itération de l'algorithme SAEM qui à $\theta^{(k)}$ associe $\theta^{(k+1)}$ est donnée par :

- ◇ **Étape S** : Simuler les paramètres individuels inconnus $\varphi^{(k)}$ suivant la distribution conditionnelle $l\left(\varphi|y, \theta^{(k)}\right)$.

◇ **Étape AE** : Approcher stochastiquement $Q\left(\theta, \theta^{(k)}\right)$ par

$$Q_{k+1}(\theta) = (1 - \delta_k) Q_k(\theta) + \delta_k \mathcal{L}\left(y, \varphi^{(k)}; \theta\right).$$

◇ **Étape M** : Mettre à jour les paramètres en maximisant $Q_{k+1}(\theta)$ par rapport à θ

$$\theta^{(k+1)} = \arg\max_{\theta} Q_{k+1}(\theta).$$

δ_k étant une suite décroissante de pas positifs tendant vers 0 et vérifiant des conditions supplémentaires exposées dans (Delyon et al., 1999) permettant d'assurer la convergence presque sûre de l'algorithme vers un maximum (local) de la vraisemblance. Comme nous l'avons souligné dans les précédentes sections, l'étape S ne peut se faire de manière directe dans le contexte des MNLEM. C'est ainsi que (Kuhn and Lavielle, 2005) ont proposés de combiner SAEM avec une procédure MCMC (SAEM-MCMC) permettant de simuler les paramètres individuels non-observés. Ils montrent aussi que l'algorithme garde ses bonnes propriétés (convergence presque sûre vers un maximum (local) de la vraisemblance) en considérant des hypothèses supplémentaires. La différence significative entre SAEM et MCEM réside dans le fait que SAEM a besoin d'une seule réalisation des données manquantes et utilise les variables simulées lors des itérations précédentes au cours des itérations suivantes.

Au vu du nombre grandissant de publications intégrant désormais l'algorithme SAEM et la disponibilité de ce dernier dans de nombreux logiciels de référence, on peut dire qu'il constitue clairement un pivot concernant l'estimation des paramètres dans les MNLEM par le maximum de vraisemblance. Cet algorithme, initialement implementé dans le logiciel MONOLIX est désormais disponible sous R via la package saemix, sous la version 7 de NONMEM (longtemps considéré comme logiciel de référence dans le cadre des modèles pharmacocinétiques-pharmacodynamiques (PKPD)) sous Matlab via la procédure nlmefitsa. Comme l'attestent plusieurs publications parmis lesquelles (Samson et al., 2007; Snoeck et al., 2010; Chan et al., 2011; Dubois et al., 2011; Savic et al., 2011), l'algorithme SAEM peut-être utilisé pour plusieurs types de données et dans plusieurs champs d'application. De nombreuses extensions de l'algorithme SAEM ont été développées pour s'adapter à des situations pratiques complexes. (Samson et al., 2006) proposent une extension de l'algorithme SAEM permettant de prendre en compte les données censurées à gauche. Plus récemment, (Delattre and Lavielle, 2012) ont proposés une extension de l'algorithme SAEM pour l'estimation des paramètres dans les modèles de Markov cachés à effets mixtes.

Nous proposons dans cette thèse des extensions de l'algorithme SAEM aux modèles de mélange et conjoints non linéaires à effets mixtes.

2.2 Modèles de mélange fini

2.2.1 Modèle de mélange de distributions

Depuis l'article de (Pearson, 1894) sur l'estimation des paramètres d'un mélange de deux lois gaussiennes univariées, les mélanges finis de distribution de probabilité ont fait l'objet de nombreux travaux. Cette attention étant due au fait que ces mélanges reflètent l'idée intuitive qu'une population est composée de plusieurs groupes, caractérisés chacun par une distribution de probabilité. Leur flexibilité permet en outre de modéliser une large variété de phénomènes aléatoires. L'attention portée aux modèles de mélange s'est amplifiée grace à l'apparition d'une méthode efficace d'estimation des paramètres par maximum de vraisemblance : l'algorithme EM (Dempster et al., 1977). Plusieurs ouvrages de référence existent à présent sur les questions liées aux mélanges finis de distribrtions, aussi bien sur le plan théorique que pratique. Il s'agit entre autres des livres de (Everitt and Hand, 1981), (Titterington and Smith, 1985), (McLachland and Basford, 1988), (McLachland and Peel, 2000) ou plus récemment encore (Frühwirth-Schnatter, 2006). Un exemple intéressant provenant de la Biologie a été donné par (Titterington and Smith, 1985) pour l'analyse des données sur les longueurs de 256 snappers, montrant comment une distribution de melange survient quand une hétérogénéité non- observée est présente dans la population pour laquelle une caractéristique aléatoire particulière est observée. De manière similaire, on retrouve des données dans plusieurs autres domaines tels que le marketing (Rossi et al., 2005), ou encore la santé publique (Spieghelalter et al.2003).

Modèle

Une variable aléatoire ou un vecteur aléatoire Y prenant ses valeurs dans un espace $\mathcal{Y} \subset \mathbb{R}^{d_y}$, continue ou discrète suit une loi de mélange (fini) si sa densité est une combinaison convexe d'un nombre M de densités :

$$\begin{array}{rccc} f & : & \mathcal{Y} & \to & \mathbb{R} \\ & & y & \longmapsto & \sum_{m=1}^{M} \pi_m f_m(y) \end{array}$$

$f_1(.), \ldots, f_M(.)$ étant les densités de probabilité de chacun des composants du mélange. $\pi_1, \ldots, \pi_M$ sont les proportions du mélange et vérifient les relations

$$\forall m \in \{1, \ldots, M\}, \ \pi_m \in [0,1] \text{ et } \sum_{m=1}^{M} \pi_m = 1. \tag{2.8}$$

M est appelé le nombre de composants. Dans la plupart des applications,on considère que toutes les densités des composants $f_m(.)$ appartiennent à une même

famille de distribution paramétrique $\mathcal{P}(\vartheta)$ de densité $f(y|\vartheta)$, indexée par le paramètre $\vartheta \in \Theta$:

$$f(y|\vartheta) = \sum_{m=1}^{M} \pi_m f_m(y|\theta_m). \tag{2.9}$$

La densité de probabilité du melange de distribution $f(y|\vartheta)$ est indexée par le paramètre $\vartheta = (\pi_1, \ldots, \pi_M, \theta_1, \ldots, \theta_M)$ prenant ses valeurs dans l'espace paramétrique $\Theta_M = \Pi_M \times \Theta^M$, Π_M étant l'ensemble des M-uplets $\pi_1, \ldots, \pi_M$ vérifiant la condition (2.8).

Il existe dans la littérature plusieurs types de mélanges de lois (mélange de lois uniforme, binomiales, exponentielles, gaussiennes,...). D'un point de vu historique, le mélange de deux distributions gaussiennes univariées de moyennes différentes et de variance différentes est la plus ancienne application connue du modèle de mélange de distribution (Pearson, 1894). On rencontre par la suite des modèles de mélange de distributions de Poisson (Feller, 1943), modèles de mélange de distributions exponentielles (Teicher, 1963). Ces exemples sont des cas spéciaux des modèles de mélange de distributions issues de la famille exponentielle traités par (Barndoff-Nielsen, 1978). (Shaked, 1980) présente un traitement mathématique général sur les mélanges de la famille exponentielle. le mélange de gaussiennes reste néanmoins le plus populaire dans la littérature statistique. Dans le cas particulier d'un mélange de gaussiennes, les f_m sont remplacées par les densités d'une gaussienne de moyenne μ_m et de matrice de variance Σ_m. On aura donc

$$\forall y \in \mathbb{R}^{d_y}, \;\; f(y|\theta) = \sum_{m=1}^{M} \pi_m \Phi_{d_y}(y; \mu_m, \Sigma_m),$$

$\Phi_d(.; \mu, \Sigma)$ étant la densité d'une gaussienne d-dimensionnelle de moyenne μ et de matrice de variance Σ. Le vecteur des paramètres est donné ici par $\vartheta = (\pi_1, \ldots, \pi_M, \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M)$. Pour des nécessités de modélisation, les matrices de variances covariance des composantes Σ_m peuvent être restreintes à des structures particulières. (Banfield and Raftery, 1993) et par la suite (Celeux and Govaert, 1995) ont proposés de manière analogue, la décomposition suivante de Σ_m dans chaque composante :

$$\Sigma_m = \lambda_m H_m A_m H_m',$$

où λ_m représente la plus grande valeur propre de Σ_m dans (Banfield and Raftery, 1993) et $\lambda_m = \left|\Sigma_m^{\frac{1}{d_y}}\right|$ dans (Celeux and Govaert, 1995). Suivant cette dernière convention, λ_m est le volume de Σ_m. $\lambda_m A_m$ est une matrice diagonale contenant les valeurs propres de Σ_m rangées par ordre décroissant sur la diagonale : cette quantité représente la forme des composantes. H_m est la matrice des vecteurs propres de Σ_m et représente l'orientation des composantes. En faisant varier ou non les proportions du mélange, les volumes, les formes et les orientations, on obtient une collection de 28 modèles présentés dans (Celeux and Govaert, 1995).

Identifiabilité

En considérant l'équation (2.9) décrivant la densité d'un mélange fini de distribution à M composantes, il existe $c = 1, \ldots, M!$ manières équivalentes d'arranger les composantes. Chacune d'elles pouvant être décrite par une permutation $\mathfrak{p}_c : \{1, \ldots, M\} \to \{1, \ldots, M\}$. Soit $\vartheta = (\pi_1, \ldots, \pi_M, \theta_1, \ldots, \theta_M)$ un point arbitraire de l'espace paramétrique $\Theta_M = \Pi_M \times \Theta^M$; considérons le sous-ensemble $\mathcal{V}(\vartheta)$ de Θ_M défini par :

$$\mathcal{V}(\vartheta) = \bigcup_{c=1}^{M!} \left\{ \vartheta^* \in \Theta_M \ : \ \vartheta^* = \left(\pi_{\mathfrak{p}_c(1)}, \ldots, \pi_{\mathfrak{p}_c(M)}, \theta_{\mathfrak{p}_c(1)}, \ldots, \theta_{\mathfrak{p}_c(M)} \right) \right\} \tag{2.10}$$

Tout point ϑ^* de $\mathcal{V}(\vartheta)$ génere la même distribution de mélange que ϑ. En effet, ϑ^* est obtenu en rearrangeant les composantes de la distribution de mélange (2.9) via la permutation $\mathfrak{p}_c$ utilisée dans la définition de ϑ^* :

$$\begin{aligned} f(y|\vartheta) &= \pi_1 f(y|\theta_1) + \ldots + \pi_M f(y|\theta_M) \\ &= \pi_{\mathfrak{p}_c(1)} f\left(y|\theta_{\mathfrak{p}_c(1)}\right) + \ldots + \pi_{\mathfrak{p}_c(M)} f\left(y|\theta_{\mathfrak{p}_c(M)}\right) \\ &= f(y|\vartheta^*). \end{aligned}$$

Si les paramètres $\theta_1, \ldots, \theta_M$ sont tous distincts, alors $\mathcal{V}(\vartheta)$ contient $M!$ paramètres ϑ^* distincts. Ainsi, pour chaque paramètre $\vartheta \in \Theta_M$ tel que les paramètres d'au moins deux des composants θ_k et θ_l diffèrent, l'ensemble $\mathcal{V}(\vartheta)$ est un sous ensemble non-identifiable de Θ_M. Cette propriété des modèles de mélange fini de distribution est connue sous le nom de "label switching" (Redner and Walker, 1984) caractérisant l'invariance de la distribution de mélange aux permutations des composantes.

La non identifiabilité dans un modèle de mélange fini de distribution peut aussi être due à un potentiel surapprentissage des données disponibles (Crowford 1994). Considérons un mélange fini de distribution à M composantes, défini comme (2.9), où $\vartheta = (\pi_1, \ldots, \pi_M, \theta_1, \ldots, \theta_M) \in \Theta_M = \Pi_M \times \Theta^M$. Considérons ensuite un modèle de mélange fini de distribution de la même famille paramétrique, mais avec $M-1$ composantes au lieu de M. (Crowford 1994) a montré que tout mélange à $M-1$ composantes représente un sous ensemble non identifiable de Θ_M, correspondant aux mélanges à M composantes, dans lesquels soit l'une des composants est vide, soit deux composantes sont identiques. Par example, tout mélange de deux gaussiennes peut-être considéré comme un mélange de trois gaussiennes avec $\pi_3 = 0$.

Plusieurs auteurs ont proposés des solutions pour résoudre les problèmes d'identifiabilité sus-cités liés aux modèles de mélange fini de distribution en restreignant l'espace paramétrique Θ_M via des contraintes sur les paramètres des composantes dans les approches fréquentistes. (McLachland and Peel, 2000) suggèrent par exemple de définir un ordre sur les paramètres et de ne retenir que le plus petit paramètre correspondant à chaque distribution pour fixer le problème lié à l'invariance de la distribution par relabellisation des composantes. (Aitkin and Rubin, 1985) proposent dans le but de résoudre la seconde difficulté d'imposer que les composantes doivent être distinctes les unes des autres. Un mélange de

distribution à M composantes sera considéré comme un mélange de distribution à $M-1$ composantes dès que deux de ses composantes seront identiques. Dans le même esprit, les proportions de mélange nulles ne sont pas acceptées. (Yakowitz and Spragins, 1968) définissent une notion faible d'identifiabilité qui accepte le "label switching". Cette définition est suffisante lorsque les grandeurs d'intérêt ne dépendent pas de l'ordre des composantes. Dans le but de construire un test de rapport de vraisemblance pour les modèles de mélange de distribution, (Dacunha-Castelle and Gassiat, 1999) ont définis une paramétrisation appelées "locally conic parametrization" permettant de séparer la partie identifiable des paramètres et la partie non-identifiable. Dans un cadre Bayésien, le phénomène de "label switching" a attiré l'attention de plusieurs auteurs. On peut citer entre autres (Celeux, 1998; Yao and Lindsay, 2009; Papastamoulis and Iliopoulos, 2010; Yao, 2012).

Nous proposons, dans le Chapitre 3, une procédure qui permet entre autre d'éviter le "label switching" lors de l'estimation des paramètres du modèle.

Estimation des paramètres

La méthode d'estimation des paramètres d'un modèle de mélange fini de distributions par le maximum de vraisemblance la plus populaire est celle que réalise l'algorithme Espérance-Maximisation (McLachland and Peel, 2000). Il existe aussi des approches d'inférence Bayésiennes, basées sur des méthodes MCMC (Monte Carlo par chaîne de Markov) telles que l'échantillonneur de Gibbs. Nous n'en dirons néanmoins pas plus à ce sujet puisque cette méthode ne rentre pas dans le cadre de cette Thèse. L'algorithme EM a été introduit par (Dempster et al., 1977) pour calculer les estimateurs du maximum de vraisemblance des paramètres d'un modèle lorsque celui-ci comporte des données manquantes ou des variables latentes. De ce fait, il est particulièrement adapté à l'estimation des modèles de mélange de distrbutions, car il prend en compte la structure latente inhérente au problème de classification en complétant ou en augmentant les données observées avec des données non observées qui indiquent les appartenance inconnues aux classes. Nous présentons dans la suite l'algorithme EM dans le cadre d'un mélange de distributions gaussiennes, le procédé restant le même pour les autres distributions.

Algorithme EM pour un mélange de gaussiennes

Considérons un vecteur $y = (y_1, \ldots, y_N)$ d'observations indépendantes de N individus issus de M groupes distincts et dont la loi est un mélange de gaussiennes. l'objectif est d'estimer le vecteur de paramètre ϑ par maximisation de la vraisemblance des observations donnée par

$$l(y;\vartheta) = \prod_{i=1}^{N} \sum_{m=1}^{M} \pi_m \Phi_{d_y}(y|\mu_m, \Sigma_m).$$

Dans toute la suite, le sigle $\mathcal{L}$ indiquera la log-vraisemblance. l'algorithme EM (expectation Maximization) proposé par (Dempster et al., 1977) est le plus couramment utilisé pour la détermination de $\hat{\vartheta}$. Il est basé sur la maximisation par

itérations successives de l'espérance de la log-vraisemblance complète conditionnellement aux observations y et à une valeur courante du vecteur de paramètres. La log-vraisemblance complète est donnée par

$$\mathcal{L}(y, z; \vartheta) = \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \log\left(\pi_m \phi\left(y|\mu_m, \Sigma_m\right)\right)$$

z_i étant une variable latente non-observée indiquant la classe de l'individu i.
Après initialisation du vecteur des paramètres ϑ par $\vartheta^{(0)}$, cet algorithme alterne les deux étapes suivantes à l'itération d'ordre k :

- ◇ **Étape E** : Cette étape consiste à calculer $Q_k(\vartheta) = \mathbb{E}\left(\mathcal{L}(y, z; \vartheta)\,|y, \vartheta^{(k-1)}\right)$. Ce qui revient à calculer les probabilités conditionnelles que l'individu i soit issu du composant m :

$$\begin{aligned}
\gamma_m^{(k-1)}(y_i) &= \mathbb{E}\left(\mathbb{1}_{z_i=m}|y, \vartheta^{(k-1)}\right) \\
&= \mathbb{P}\left(z_i = m|y, \vartheta^{(k-1)}\right) \\
&= \frac{\pi_m^{(k-1)} \phi\left(y_i|\mu_m^{(k-1)}, \Sigma_m^{(k-1)}\right)}{\sum_{r=1}^{M} \pi_r^{(k-1)} \phi\left(y_i|\mu_r^{(k-1)}, \Sigma_r^{(k-1)}\right)}
\end{aligned}$$

- ◇ **Étape M** : Cette étape de maximisation consiste à déterminer le vecteur de paramètres $\vartheta^{(k)}$ maximisant $Q_k(\vartheta)$ sous la contrainte $\sum_{m=1}^{M} \pi_m = 1$. Étant donné le fait qu'on est amené à faire une maximisation sous une contrainte linéaire, nous utiliserons la méthode de Lagrange et on aura donc

$$\left(\vartheta^{(k)}, \hat{\lambda}\right) = \arg\max_{\vartheta, \lambda} \left(Q_k(\vartheta) - \lambda\left(\sum_{m=1}^{M} \pi_m - 1\right)\right),$$

λ étant le multiplicateur de Lagrange. Ceci est équivalent à déterminer le vecteur des proportions maximisant

$$(\lambda, \pi_1, \ldots, \pi_M) \mapsto \sum_{i=1}^{N} \sum_{m=1}^{M} \gamma_m^{(k-1)}(y_i) \log(\pi_m) - \lambda\left(\sum_{m=1}^{M} \pi_m - 1\right)$$

et les moyennes et matrices de variances minimisant

$$(\mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M) \mapsto \sum_{i=1}^{N} \sum_{m=1}^{M} \gamma_m^{(k-1)}(y_i)\left(y_i - \mu_m\right)' \Sigma_m^{-1}\left(y_i - \mu_m\right).$$

Les proportions sont donc données pour chaque composant m par

$$\pi_m^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_m^{(k-1)}(y_i),$$

les vecteurs moyennes par :

$$\mu_m^{(k)} = \frac{\sum_{i=1}^{N} \gamma_m^{(k-1)}(y_i)\, y_i}{\sum_{i=1}^{N} \gamma_m^{(k-1)}(y_i)}$$

et les matrices de variance par

$$\Sigma_m^{(k)} = \frac{\sum_{i=1}^{N} \gamma_m^{(k-1)}(y_i) \left(y_i - \mu_m^{(k)}\right) \left(y_i - \mu_m^{(k)}\right)'}{\sum_{i=1}^{N} \gamma_m^{(k-1)}(y_i)}$$

Le calcul de $\Sigma_m^{(k)}$ dépend des conditions imposées par la forme du mélange, celui présenté ci dessus ayant été effectué dans le cas de la forme la plus générale de la matrice de variance. Les calculs selon les diverses formes sont développés dans (Celeux and Govaert, 1995). L'une des principales caractéristiques de cet algorithme est le fait que la croissance de $Q_k(\vartheta)$ à chaque étape de l'algorithme implique celle de la log-vraisemblance observée $\mathcal{L}(y;\vartheta)$. elle est matérialisée par la relation

$$Q_k(\vartheta) = \mathcal{L}(y;\vartheta) + H_k(\vartheta)$$

où $H_k(\vartheta) \equiv \mathbb{E}\left(\mathcal{L}(z|y;\vartheta)\,|y, \vartheta^{(k-1)}\right)$ vérifie l'inégalité $H_k(\vartheta) \leq H_k\left(\vartheta^{(k-1)}\right)$, pour tout ϑ dans l'espace paramétrique d'après l'inégalité de Jensen. Tout accroissement de Q engendre donc un accroissement de vraisemblance. Ainsi, lorsque la maximisation de Q est plus simple que celle de la vraisemblance, on peut espérer atteindre l'emv en procédant à des maximisations successives de Q.

Cet algorithme est néanmoins fortement lié à l'initialisation considérée et nécessite des temps de calcul assez importants en pratique. Des solutions à ces problèmes sont apparues par le biais d'une approche stochastique alliant la simulation des classes latentes à l'algorithme EM. Une des premières variantes stochastiques de l'EM fut la version SEM (Simulated EM) proposée par (Celeux and Diebolt, 1986) pour l'identification de mélanges finis de densités. A l'itération k de l'étape E s'ajoute une étape S de simulation d'une réalisation de la classe z_k suivant la loi conditionnelle $p(.|y;\vartheta_k)$. L'actualisation de θ se faisant donc sur la log-vraisemblance complète $\log(f(y, z_k, \vartheta))$. Le caractère stochastique de cet algorithme permet ainsi de réduire la dépendance de l'estimateur de ϑ (la limite de la suite ϑ_k) en les conditions initiales : la simulation de la variable z_k à chaque itération laisse à la suite θ_k une certaine souplesse en autorisant une exploration plus générale des modes de la vraisemblance. L'algorithme n'est ainsi plus systématiquement contraint à explorer un voisinage du mode le plus proche des conditions initiales. Dans le but d'établir la convergence presque sûre de la suite ϑ_k vers un maximum local de la vraisemblance, les mêmes auteurs ont améliorés le précédent algorithme. Ils considèrent dorénavant une version de type recuit simulé, combinant l'algorithme EM et la version SEM précédemment proposée (Celeux and Diebolt, 1990; Celeux and Diebolt, 1992).

Sélection de modèles

Le problème de sélection de modèles dans les modèles de mélange se résume le plus souvent au choix du nombre de classe M et des contraintes adéquates (une famille de contraintes dans le cas des mélanges Gaussiens est donnée dans (Celeux and Govaert, 1995)) sur les paramètres des distributions de probabilité des composantes permettant d'expliquer de la meilleure des façons l'hétérogénéité dans la population cible. Ce choix se fait généralement à travers des critères asymptotiques de vraisemblance pénalisée tels que le critère BIC (Bayesian Information Criterion) proposé par (Schwarz, 1978) ou encore de vraisemblance complète pénalisée ICL(Integrated Completed Likelihood) proposé par (Biernacki et al., 2000).

2.2.2 Melanges fini de modèles de régression

La plupart des algorithmes standards de classification sont basés sur l'hypothèse que les vecteurs à classifier sont des réalisations de vecteurs aléatoires suivant des modèles statistiques paramétriques. Ces modèles n'admettent en général aucune restriction sur la structure de la moyenne via les covariables par exemple. Néanmoins, dans la plupart des applications telles que les études médicales longitudinales, où les mesures prises au cours du temps sur les individus présentent une structure hautement déséquilibrée, il semble naturel de modéliser la moyenne via un modèle de régression. Traditionnellement, les algorithmes non-paramétriques tels que les K-means sont utilisés pour la classification des vecteurs de taille fixe. Les dispositifs étant généralement déséquilibrés dans le cadre des données longitudinales, ces algorithmes ne pourront pas être utilisés. Ainsi, les méthodes de classification basées sur un modèle possèdent des avantages intrinsèques comparativement aux techniques de classification non-probabilistes (Li, 2006). Depuis les travaux de (Quandt, 1972) , les mélanges fini de modèles de régression ont été utilisés de manière intensive dans la littérature statistique et sont appliqués dans plusieurs disciplines telles que l'épidémiologie, la médecine, la génétique, l'économie, l'ingénierie, le marketing,... La présentation du mélange fini de modèles de regression qui sera proposée par la suite est similaire à celle de (Frühwirth-Schnatter, 2006).

Soit $(\mathrm{Y}_i, \mathrm{X}_i)$ un couple constitué d'une variable aléatoire Y_i et d'un vecteur de variables explicatives $\mathrm{X}_i = (\mathrm{X}_{i1}, \ldots, \mathrm{X}_{id_x}) \in \mathbb{R}^{d_x}$. Supposons qu'il existe une dépendance linéaire entre Y_i et X_i caractérisée par le modèle de régression :

$$\mathrm{Y}_i = \mathrm{X}_i\mu + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{2.11}$$

μ et σ_ε^2 étant des paramètres inconnus. On suppose par la suite que le coefficient de regression μ et la variance résiduelle σ_ε^2 ne sont pas homogènes pour toutes les paires $(\mathrm{Y}_i, \mathrm{X}_i)$ possible. Une manière de capturer de telles fluctuations au niveau des paramètres est de considérer un mélange fini de modèles de régression. Un mélange fini de modèles de régression suppose l'existence de M modèles de régression caractérisés par les paramètres $\left(\mu_1, \sigma_{\varepsilon,1}^2\right), \ldots, \left(\mu_M, \sigma_{\varepsilon,M}^2\right)$, et d'une variable latente

$z_i \in \{1, \ldots, M\}$ contenant la classe d'appartenance de l'individu i de sorte que l'on ait le modèle génératif suivant pour Y_i :

$$\mathrm{Y}_i = \mathrm{X}_i \mu_{z_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon, z_i}\right). \tag{2.12}$$

$\mu_1, \ldots, \mu_M$ et $\sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,M}$ sont les paramètres inconnus à estimer. Tout comme dans les modèles de mélange de distributions, l'application des mélanges finis de modèles de regression nécessite des hypothèses supplémentaire sur le mécanisme permettant de générer la variable z_i. En l'absence d'informations supplémentaires, on suppose généralement une indépendance mutuelle des z_i pour tous les individus de la population, et une distribution de probabilité inconnu $\pi = (\pi_1, \ldots, \pi_M)$. Le vecteur de paramètres a estimer est donc donné par

$$\vartheta = \left(\mu_1, \sigma^2_{\varepsilon,1}, \pi_1, \ldots, \mu_M, \sigma^2_{\varepsilon,M}, \pi_M\right).$$

La vraisemblance des observations est donnée par :

$$l\left(y_i; \vartheta\right) = \sum_{m=1}^{M} \pi_m p\left(y_i | z_i = m; \vartheta\right) = \sum_{m=1}^{M} \pi_m \phi\left(y_i | \mathrm{X}_i \mu_m; \sigma^2_{\varepsilon,m}\right),$$

$\phi\left(\cdot | a, b\right)$ étant la densité de probabilité d'une gaussienne de moyenne a et de variance b. Ainsi, pour chaque valeur du vecteur de variable explicative X_i, la distribution marginale de Y_i est une distribution de mélange fini de gaussiennes unidimensionnelles de moyenne $\mu_{m,i} = \mathrm{X}_i \mu_m$ et de variance $\sigma^2_{\varepsilon,m}$. Un melange fini de modèles de régression peut ainsi être vu comme une extension d'un modèle de mélange de gaussiennes unidimensionnelles avec des moyennes dépendant des variables explicatives. D'autre part, un modèle de mélange fini de gaussiennes unidimensionnelles est un cas particulier de mélange fini de modèles de regression où $\mathrm{X}_i \equiv 1$ pour tout i. Les mélanges finis de modèles de régression sont reconnus dans la littérature statistique sous divers noms : les modèles de regression du changement ("switching regression models") en économie (Quandt, 1972) ; modéles de régression à classe latente en marketing (DeSarbo and Cron, 1988) ; mélange de modèles expert en apprentissage machine (Jacobs et al, 1991) ; modèles mixtes en biologie (Wang et al., 1996).

Les mélanges finis de modèles de régression souffre tout comme les modèles de mélange classique du "label switching" et plusieurs autres obstacles à l'identifiabilité détaillés dans (Hurn et al., 2003; Frühwirth-Schnatter, 2006). L'inférence sur les paramètres des mélanges finis de modèles de régression se fait généralement à travers le maximum de vraisemblance ou encore le maximum à posteriori via les méthodes Bayesiennes. L'un des premiers travaux liés aux mélanges fini de modèles de régression fut celui de (Quandt, 1972) qui maximise la vraisemblance des observations de manière numérique en utilisant un algorithme de gradient conjugué. L'auteur signale néanmoins que la méthodologie est sujette à de gros soucis de convergence dans le cas des expériences répétées sur des données générées artificiellement. Plutard, (Quandt and Ramsey, 1978) ont développés une procédure basée sur l'estimateur des moments pour estimer les paramètres. (Hosmer, 1974) a

développé un algorithme EM pour l'estimation des paramètres par maximization de la vraisemblance d'un mélange de modèles de régression à deux composantes dans un cas univarié . Plutard, (DeSarbo and Cron, 1988) généralise cet algorithme à un nombre quelconque de composants. (Jones and McLachlan, 1992) étendent les travaux de DeSarbo aux observations multivariées. (Hurn et al., 2003) discutent des solutions au problème de "label switching" pour l'inférence Bayesienne avec des mélanges de modèles de régression, et (Viele and Tong, 2002) présentent des résultats de consistance pour la distribution à posteriori.

Il existe à ce jour dans la littérature plusieurs extensions aux mélanges fini de modèles de régression :

- ◇ *Melanges finis de modèles de regression à effets mixtes* : Ces modèles permettent de combiner les coefficients de régression qui sont fixes suivant toutes les réalisations $(\mathrm{Y}_i, \mathrm{X}_i)$ avec ceux qui varient. En utilisant les mêmes notations que précédemment, on a dans ce cas la spécification suivante :

$$\mathrm{Y}_i = \mathrm{X}_i^f \beta + \mathrm{X}_i^a \mu_{z_i} \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon, z_i}\right)$$

  où X_i^f représente l'effet fixe et X_i^a l'effet aléatoire. Nous faisons remarquer tout de même qu'ici, le sens de "effets mixtes" n'est pas le même que celui de la Section 2.1. Le lecteur pourra se reporter à l'ouvrage de (Frühwirth-Schnatter, 2006) pour plus ample informations sur l'identifiabilité de tels modèles ainsi que les méthodes d'inférence qui pour la plupart sont Bayesiennes.
- ◇ Récemment, plusieurs auteurs ont incorporés des effets aléatoires dans une grande variété de modèles de régression pour des réponses corrélées et à multiples sources de variation. Dans le contexte des mélanges, (Gaffney and Smith, 2003) ont développés un modèle de mélange de modèles de regression (linéaire) à effets aléatoires et dérivés une méthode de type EM basée sur le maximum à posteriori pour l'inférence paramétrique. (Celeux et al., 2005) ont développés un mélange de modèles linéaires à effets mixtes pour la classification des profils de l'expression génétique provenant des expériences répétées. Ils utilisent l'algorithme EM pour l'estimation des paramètres. (Pfeifer, 2004) considère le problème de classification basé sur un modèle à effets mixtes semi-paramétrique. Plus récemment, Booth et al. (2007) ont proposés une approche Bayesienne pour la classification des données multivariées, basée sur un model linéaire mixte multi-niveau. (De la Cruz-Mesia et al., 2008) proposent une approche basée sur un modèle pour la classification des données longitudinales. Ils proposent un algorithme EM pour l'approche fréquentiste et une méthode MCMC pour l'estimation Bayesienne. Néanmoins, Dans l'approche fréquentiste, à chaque itération EM et pour chaque individu, une intégration Monte-Carlo de la loi marginale est calculée dans chaque classe pour déterminer les probabilités à posteriori d'appartenance aux classes, suivi d'une autre intégration Monte-Carlo suivant la loi à posteriori pour résoudre la problématique inhérente à l'étape E de l'algorithme EM dans les MNLEM.

Nous proposons dans cette Thèse au Chapitre 3 des extensions de modèle de mélanges dans le cadre des modèles non linéaires à effets mixtes, ainsi qu'une nouvelle méthodologie d'estimation des paramètres basée sur l'estimation par le maximum de vraisemblance. Le Chapitre 4 constitue une application à un jeu de données réelles sur le VIH.

2.3 modèles conjoints

2.3.1 Motivations

Lors des études médicales, plusieurs type de données sont généralement collectées sur chaque individu. on peut citer entre autres, des réponses de type longitudinal et le temps d'attente jusqu'à l'apparition d'un événement d'intérêt particulier. Les questions de recherche inhérentes à ce type de données sont diverses et variées. L'une d'entre elles consisterait en une analyse séparée des diverses réponses. Néanmoins, dans plusieurs situations, l'intérêt principal de l'étude consiste à étudier la structure associative des diverses réponses. Un exemple fréquemment rencontré pour ce type de problème pourrait se trouver dans la recherche sur les marqueurs biologiques, où la plupart des études cliniques consistent en l'identification des biomarqueurs ayant de fortes capacités pronostiques pour l'événement d'intérêt. Parmi les exemples standards, on peut citer entre autres, les recherches sur le VIH où l'intérêt est généralement lié à l'association entre le nombre de CD4 ou la charge virale et le temps d'attente jusqu'à l'apparition du SIDA, les études sur les cirrhoses, destinées à évaluer la relation entre le sérum bilirubin et le temps jusqu'au décès. Une caractéristique inhérente importante de ces conditions médicales est leur nature dynamique. Ainsi, le taux de progression vers l'évènement d'intérêt est non seulement différent d'un patient à l'autre, mais change aussi dynamiquement dans le temps pour un même patient. Ainsi, le vrai potentiel d'un biomarqueur dans la description de la progression d'une maladie et son association avec la survie ne peut-être révélée qu'en considérant des évaluations répétées du marqueur dans l'analyse.

Plusieurs auteurs se sont penchés sur la question de modélisation de la relation entre une variable longitudinale et le temps jusqu'à l'apparition d'un événement d'intérêt. La première approche fut celle de (Kalbfleisch and Prentice, 2002) qui proposent de considérer un modèle de survie à risque proportionnel pour le délai jusqu'à l'évènement d'intérêt, en considérant la variable longitudinale comme une covariable dépendant du temps. Si on note T_i la variable aléatoire contenant le délai jusqu'à l'évènement d'intérêt sur le sujet i, $Y_i(t), t > 0$ le processus longitudinal sur le sujet i, et Z_i un vecteur de covariables mesurées sur l'individu i. L'association entre le processus longitudinal, le délai à l'évènement et les covariables est caractérisée par la relation entre T_i, $Y_i(t)$, et Z_i via le modèle de risque

proportionnel

$$\begin{aligned}\lambda_i(t) &= \lim_{dt\to 0}\frac{\mathbb{P}\left(t \le T_i < t+dt | T_i \ge t, Y_i^H(t), Z_i\right)}{dt} \\ &= \lambda_0(t)\exp\left(\gamma Y_i(t) + \beta' Z_i\right) \end{aligned} \tag{2.13}$$

$Y_i^H(t) = \{Y_i(u), 0 \le u \le t\}$ étant l'historique du processus longitudinal jusqu'au temps t. Ainsi, la fonction de risque dépend linéairement de l'historique du processus longitudinal via la valeur courante $Y_i(t)$. Ce modèle, très intuitif et simple à appréhender théoriquement, s'avère néanmoins compliqué à mettre en oeuvre en pratique car les données idéales ne sont généralement pas disponibles. En dépit du fait que le modèle de risque donné par 2.13 nécessite les valeurs de la réponse longitudinale à tout instant t, cette dernière n'est mesurée que de manière intermittente à des instants $t_{ij} \le T_i,\ i = 1, \ldots, n_i$. De plus, les valeurs observées $Y_i(t_{ij})$ pourraient différer des vraies valeurs de la variable longitudinale à cause des erreurs de mesure et au lieu de $Y_i(t_{ij})$, on observerait plutôt $Y_i(t_{ij}) + \varepsilon_{ij}$, où ε_{ij} représente une erreur intra-individuelle. Le temps d'évènement pouvant être censuré, la valeur de la variable longitudinale aux instants de censure est considérée comme manquante. Dans l'idéal des cas où $Y_i(t)$ est observable pour tout $t \le T_i$, la seul difficulté inhérente au modèle (2.13) est la présence des censures. Soit C_i la variable aléatoire contenant l'instant de censure de l'évènement d'intérêt. Ainsi, on l'observation sur l'évènement d'intérêt est constituée des variables $B_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{(T_i \le C_i)}$. Suivant (Cox, 1972) et sous les conditions discutées dans (Kalbfleisch and Prentice, 2002), l'inférence sur β et γ se fait en maximisant la vraisemblance partielle

$$\prod_{i=1}^{N}\frac{\exp\left(\gamma Y_i(B_i) + \beta' Z_i\right)}{\sum_{k=1}^{N}\exp\left(\gamma Y_k(B_i) + \beta' Z_k\right)\mathbb{1}_{(B_k \ge B_i)}}. \tag{2.14}$$

Il apparaît donc clairement à travers (2.14) que l'inférence nécessite les valeurs du processus longitudinal $Y_i(t)$ pour tous les individus $i = 1, \ldots, N$ et à tous les instants observés d'évènement. Comme nous l'avons mentionné précédemment, l'implémentation de ce modèle en pratique est compliquée de par le fait qu'on a pas toutes les valeurs nécessaires de $Y_i(t)$.

Les premières approches qu'on pourrait considérer comme naïves consistaient généralement à imputer la valeur manquante $Y_i(t)$ par la valeur observée la plus proche pour l'individu i. Cette propriété est connu dans littérature sous le nom LOCF (Last Observation Carried Forward). Néanmoins, cette approche résulte en des estimations biaisées (Prentice, 1982), particulièrement quand le délai entre les diverses mesures est long ou les erreurs de mesure sont importantes. Des auteurs proposent dans le but de s'affranchir des erreurs de mesures et obtenir les valeurs de la variable longitudinale à tout instant, d'estimer les paramètres d'un modèle mixte sur la variable longitudinale, et d'ensuite utiliser les prédictions aux instants non observés dans le modèle de survie. D'après (Dafny and Tsiatsis, 1998), cette approche permet de réduire le biais par rapport à la précédente sans toute fois

l'éradiquer. Ainsi, pour répondre aux questions de recherche concernant l'association entre des mesures répétées et le temps jusqu'à l'apparition d'un événement d'intérêt de manière optimale, une nouvelle classe de modèles statistiques a vue le jour : modèles conjoints des données longitudinales et de survie.

De nombreux auteurs se sont intéressés aux modèles conjoints, parmi lesquels (Tsiatis and Davidian, 2004) qui donne un apercu global des modèles conjoints et montre que l'estimations simultanée des paramètres d'un modèle linéaire mixte pour l'évolution de la variable longitudinale et d'un modèle de survie pour le délai d'attente jusqu'à l'apparition de l'évènement, en maximisant la vraisemblance conjointe permet d'obtenir des estimations sans biais des paramètres. Ces modèles sont désormais largement utilisés dans plusieurs domaines et ont fait l'objet de plusieurs ouvrages, l'un des plus récents étant celui de (Rizopoulos, 2012b).

2.3.2 Formalisation des modèles Conjoints

Nous considérons ici la description faite par (Tsiatis and Davidian, 2004). Comme dans la section précédente, notons pour $i = 1, \ldots, N$, T_i et C_i respectivement le délai à l'évènement et la censure ; $Y_i(t)$ le processus longitudinal à un instant t, Z_i un vecteur de covariables, $B_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{(T_i \leq C_i)}$.

Un modèle conjoint est constitué de deux sous-modèles liés entre eux. L'un modélisant le processus longitudinal $Y_i(t)$ et l'autre pour le délai à l'évènement T_i, avec des hypothèses et des spécifications supplémentaires permettant une représentation complète de la distribution jointe des données observées $\{Y_i, V_i, \Delta_i\}$.

Modèle des données longitudinales

Un modèle linéaire mixte (Laird and Ware, 1982) est généralement utiliser pour les données longitudinales :

$$Y_{ij} = f(\varphi_i, t_{ij}) + \varepsilon_{ij}, \tag{2.15}$$

où $Y_{ij} = Y_i(t_{ij})$, f est une fonction linéaire en les paramètres individuels φ_i et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ représente l'erreur intra-individuelle.

Modèle de survie

Le modèle de survie considéré est généralement de la forme (2.13) en remplaçant $Y_i(t)$ par des valeurs dépourvues d'erreurs de mesure :

$$\lambda_i(t) = \lambda_0(t) \exp(\gamma f(\varphi_i, t) + \beta' Z_i). \tag{2.16}$$

(Schluchter, 1992) et (Pawitan and Self, 1993) proposent des modèles conjoints de la forme (2.15) et (2.16) avec un modèle de risque paramétrique. (Self and

Pawitan, 1992) proposent un modèle longitudinal de la forme (2.15) et un modèle de risque similaire à (2.13) en remplaçant le terme $\exp(\gamma Y_i(t))$ par $1+\gamma Y_i(t)$ pour avoir un modèle de risque linéaire en φ_i. Ils utilisent une stratégie inférentielle à deux étapes où l'estimateur des moindres carrés des paramètres individuels φ_i est utilisé pour imputer les valeurs manquantes de $Y_i(t)$. (Tsiatis et al., 1995) proposent un modèle conjoint de la forme conjoint de la forme (2.15) et (2.16) pour effectuer l'inférence sur la relation entre le taux de CD4 $Y_i(t)$ et la mortalité T_i dans une étude sur le VIH. Ils proposent de linéariser la fonction de risque et d'utiliser les estimations des paramètres individuels pour prédire les valeurs non-observées de la variable longitudinale dans la vraisemblance partielle (2.14). (Dafny and Tsiatsis, 1998) montre que cette approche résulte en une estimation biaisée des paramètres, mais réduit néanmoins le biais par rapport aux méthodes naïves d'imputation.

Vraisemblance conjointe

La vraisemblance conjointe issue des modèles (2.15) et (2.16) est donnée par :

$$\begin{aligned} l\left(\lambda_0(\cdot),\gamma,\beta,\sigma^2,\theta_\varphi\right) &= \prod_{i=1}^{N}\int \left[\lambda_0(B_i)\exp\left\{\gamma f(\varphi_i,B_i)+\beta' Z_i\right\}\right]^{\Delta_i} \\ &\quad \exp\left[-\int_0^{B_i}\lambda_0(u)\exp\left\{\gamma f(\varphi_i,u)+\beta' Z_i\right\}\ du\right] \\ &\quad \frac{1}{(2\pi\sigma^2)^{n_i/2}}\exp\left[-\sum_{j=1}^{n_i}\frac{\left(Y_{ij}-f(\varphi_i,t_{ij})\right)^2}{2\sigma^2}\right] p(\varphi_i;\theta_\varphi)\ d\varphi_i. \end{aligned} \tag{2.17}$$

(DeGruttola and Tu, 1994) proposent un modèle longitudinal de la forme (2.15) et un modèle paramétrique (lognormal) pour T_i et proposent un algorithme EM pour l'estimation des paramètres. Plus récemment, (Rizopoulos, 2012a) a proposé une méthode d'estimation basée sur quadrature de Gauss pour la détermination de l'intégrale présente dans l'expression de la vraisemblance conjointe (2.17). Cette méthode est désormais implémentée dans le package *JM* (Rizopoulos, 2010) du logiciel R.

Nous proposons dans le cadre de cette Thèse, principalement au Chapitre 5, plusieurs extensions des modèles conjoints ainsi que des méthodes d'inférence optimales.

- ◇ comme on l'a vu dans le précédent argumentaire, les modèles linéaires mixtes sont considérés pour modéliser la variable longitudinale. Cette hypothèse est néanmoins rarement vérifiée dans les modèles de pharmacologie qui sont généralement issus des solutions d'équations différentielles ordinaires ou stochastiques. Nous proposons donc à la place des modèles mixtes linéaires, des modèles non-linéaires permettant d'expliquer un très grande variété de phénomènes biologiques, incluant ceux définis par une relation linéaire.

- ◇ On remarque aussi que toutes les méthodologies évoquées précédemment sont valables dans le cas de la modélisation jointe d'un marqueur longitudinal avec le délai jusqu'à l'apparition d'un *unique* évènement ou encore un évènement terminal, pouvant se produire une seule fois durant l'étude. Nous proposons de modéliser de manière conjointe les données longitudinales et les délais successifs d'apparition répétée des évènements en considérant en plus des censures à droite, des censures par intervalle. Un modèle non-linéaire à effets mixtes est proposé pour modéliser l'évolution du marqueur et un modèle mixte paramétrique de risque est proposé pour modéliser les évènements répétés.
- ◇ Une procédure d'inférence basée sur l'algorithme SAEM est proposée permettant de maximiser la vraisemblance conjointe exacte et non pas des approximations.

Chapitre 3

Inference in mixtures of non-linear mixed effects models

Contents

Abstract

We propose several extensions of non linear mixed-effects models (NLMEM) for dealing with mixture models. These extensions include among others, mixtures of distributions, mixtures of structural models and mixtures of residual error models. A new methodology is proposed to carry out estimation of parameters of these mixture models by maximizing the likelihood of the observations. Since the individual parameters inside the NLMEM are not observed, we propose to combine the EM algorithm usually used for mixtures models when the mixture structure concerns an observed variable, with the SAEM (Stochastic Approximation EM) algorithm, which is known to be suitable for estimating parameters in NLMEM and also has nice theoretical properties. The main advantage of this hybrid procedure is to avoid a simulation step of unknown labels required by a "full" version of SAEM. Indeed, such a simulation step may introduce numerical instability into the algorithm. The resulting MSAEM (Mixture SAEM) algorithm is now implemented in the MONOLIX software. Several criteria for classification of subjects and estimation of individual parameters are also proposed. Numerical experiments on simulated data showed that MSAEM performs well in a general framework of mixtures of NLMEM. Indeed, MSAEM provides an estimator close to the maximum likelihood estimator in very few iterations and is robust with regard to initialization.

Keywords: *SAEM algorithm - Maximum likelihood estimation - Mixture models - Non linear mixed effects model* - MONOLIX .

3.1 Introduction

Mixed model are statistical models containing both fixed effects and random effects. They are well suited for the analysis of the wide variety of data for which observations come from a population of differing individuals. Here, there are two sources of variability: intra-subject and inter-subject. Part of the inter-subject variability can be explained by known covariates (age, weight, sex, etc.). The random effects are used for modelling the non-explained part of the inter-subject variability of the individual parameters.

Mixed effects models are useful in a wide variety of disciplines such as for example agronomy, biology and pharmacology. The use of nonlinear mixed-effects for quantifying both within and between-subject variability in a drug's pharmacokinetics (PK) was first proposed in (Sheiner et al., 1972). Following this work, the population approach had a significant impact on the quantification of the variability of pharmacokinetics and pharmacodynamics data to better understand the variability of responses in a given population to the same treatment.

A set of patients is usually heterogeneous with respect to response to a drug therapy. In any clinical efficacy trial, patients who respond, partially respond or do not respond present quite different profiles. Thus, diversity of the observed kinetics cannot be explained adequately only by inter-patient variability of certain parameters.

It was recognized as well in (Evans and Relling, 1999) that genetic polymorphisms in drug metabolism and in the molecular targets of drug therapy can also have a significant influence on the efficacy and toxicity of medications. There is therefore a need for population modeling approaches that allow taking into account the existence of sub-populations, with the aim of helping to identify, for example, unknown genetic determinants of observed pharmacokinetic/pharmacodynamic (PK/PD) phenotypes.

Introducing a categorical covariate (sex, genotype, treatment, etc.) in the model assumes that the whole population can be decomposed into several sub-populations. Mixture models usually refer to such models when the categorical covariate is unknown. They are relevant to model heterogenous populations when observed data is treated as outcomes from different populations. Mixture models are well studied; we refer the reader to (Redner and Walker, 1984; McLachland and Basford, 1988; Bryant, 1991; Roeder and Wasserman, 1997; McLachland and Peel, 2000; Frühwirth-Schnatter, 2006) and references therein for more details.

There exist several types of mixture models in the context of mixed effects models: *Mixtures of distributions* assume that non-observed individual parameters come from different sub-populations Such models are considered for instance in (Frühwirth-Schnatter, 2006; De la Cruz-Mesia et al., 2008). *Between-subject model mixtures* (BSMM) also assume that there exist sub-populations of patients. Here, various structural models describe the response of each sub-population, and each

patient belongs to one sub-population. *Within-subject model mixtures* (WSMM) assume that there exist sub-populations (of cells, viruses, etc.) within each patient. Differing structural models describe the response of each sub-population, and the proportions of each sub-population depend on the patient.

Our goal is to propose new methods for maximum likelihood estimation (MLE) of population parameters in the context of a mixture of NLMEM. Since the individual parameters and the latent variables, which contain individual labels, are not observed, we can suppose that we are in a classical framework of incomplete data, and EM-type algorithms can be envisaged. The EM algorithm (Dempster et al., 1977; Wu, 1983) and several EM-based algorithms have been used in "classical" mixture models for different goals (density estimation, clustering). (Celeux and Diebolt, 1986) proposed a stochastic EM (SEM) algorithm which incorporated a stochastic step (S step) between the E and the M steps to deal with some limitations of the EM algorithm (limiting position strongly dependent on starting position, extremely slow rate of convergence in some cases, etc.). (Celeux and Govaert, 1992) proposed a Classification EM (CEM) algorithm for clustering which incorporated a classification step (C step) between the E and M steps. (Wang et al., 2007; Wang et al., 2009) used a Monte Carlo EM (MCEM) algorithm with importance sampling to deal with the intractable E step of the EM algorithm in non-linear mixtures. In the NLMEM framework, (Kuhn and Lavielle, 2005) proposed the SAEM (Stochastic Approximation EM) algorithm which incorporates a simulation step of the unobserved individual parameters and an approximation of several statistics between the E and M steps. SAEM is a very powerful tool for NLMEM, known to accurately estimate population parameters and also to have good theoretical properties.

In this paper, we propose to combine the EM algorithm typically used for mixtures models when the mixture structure concerns some observed variable, with the SAEM algorithm used in NLMEM. The use of the resulting Mixed SAEM (MSAEM) instead of the SAEM itself avoid a simulation step of the unobserved categorical covariates and significantly improves results. Section 2 of this paper describes NLMEM and mixtures of NLMEM, and Section 3 is devoted to a description of the proposed methods. Several numerical examples in Section 4 illustrate the performance of MSAEM. A discussion is provided in section 5.

3.2 Mixtures in non linear mixed-effects models

3.2.1 Non linear mixed-effects model

Mixed-effects models are used in a wide variety of applications, including population pharmacology (see for example (Wakefield et al., 1998)). Data in this particular field consist of repeated measurements on a number of individuals with specific individual parameters. In the population approach, mixed effects mod-

els are commonly used for modelling inter-subject variability of these individual parameters.

Mixed-effects models can address a wide class of data including continuous, count, categorical and time-to-event data. Modelling such data leads to using NLMEM as hierarchical models. At a first level, each individual has their own parametric regression model, known as the structural model, each identically defined up to a set of unknown individual parameters. At a second level, each set of individual parameters is assumed to be randomly drawn from some unknown population distribution.

We will focus here on continuous data models which can be described as follows:

$$y_{ij} = f\left(x_{ij};\varphi_i\right) + g\left(x_{ij};\varphi_i,\theta_y\right)\varepsilon_{ij}, \tag{3.1}$$

where

- $y_{ij} \in \mathbb{R}$ denotes the j-th observation for the i-th individual, $1 \leq i \leq N$ and $1 \leq j \leq n_i$. $y_i = (y_{ij})$ is the vector of observations for the i-th individual.
- N is the number of individuals and n_i the number of observations for the i-th individual.
- x_{ij} denotes a vector of regression variables (for longitudinal data, x will generally be time).
- φ_i is the d-vector of individual parameters of individual i. We assume that all the φ_i are drawn from the same population distribution. We limit ourselves to Gaussian models of the form:

$$\varphi_i = h(\mu, c_i) + \Sigma^{-1/2}\eta_i, \tag{3.2}$$

  where h is a function which describes the covariate model: μ a vector of fixed-effects and c_i a vector of known covariates, $\eta_i \sim_{i.i.d} \mathcal{N}\left(0, I_d\right)$ a vector of standardized random effects and Σ the inter-individual variance-covariance matrix.
- $\varepsilon_{ij} \sim \mathcal{N}\left(0,1\right)$ denote the residual errors and are independent of individual parameters φ_i.
- f is a function defining the structural model and g a function defining the (possibly heteroscedastic) residual error model.
- $\theta = (\mu, \Sigma, \theta_y)$ denotes the complete set of population parameters.

The model is therefore completely defined by the joint probability distribution of the observations $y = (y_i)$ and the individual parameters $\varphi = (\varphi_i)$ which admits this hierarchical decomposition:

$$p(y, \varphi; \theta) = p(y|\varphi; \theta)p(\varphi; \theta) \tag{3.3}$$

Many problems are related with the use of these models: estimation of the population parameters θ with their standard errors, calculation of the likelihood of the observations $p(y;\theta)$ for model selection or hypothesis testing purpose, estimation of the individual parameters (φ_i),...

When the model is linear, *i.e.* when the observations (y_{ij}) are Normally distributed, then the likelihood of the observations can be computed in a closed form and the EM algorithm can be used for maximizing this likelihood. On the other hand, when the structural model is not a linear function of the random effects and/or when the residual error model is not a linear Gaussian model, then the model of the observations is not linear anymore and the likelihood cannot be maximized using EM. A complete methodology for NLMEM is implemented in the MONOLIX software, including the Stochastic Approximation of EM (SAEM) proposed in (Delyon et al., 1999). This algorithm is becoming a reference method for maximum likelihood estimation in NLMEM. Indeed, it is now implemented in NONMEM (software widely used for PKPD applications), Matlab (nlmefitsa.m) and R (saemix package).

3.2.2 Mixtures of mixed effects models

Mixture models can be useful for representing the presence of sub-populations within an overall population. The simplest way to model a finite mixture model is to introduce a label sequence $(z_i; 1 \leq z_i \leq N)$ that takes its values in $\{1, 2, \ldots, M\}$ and is such that $z_i = m$ if subject i belongs to sub-population m.

In some situations, the label sequence (z_i) is known and can then be used as a categorical covariate in the model. We will address in the following the more challenging situation where this sequence is unknown. We therefore consider that (z_i) is a sequence of independent random variables taking values in $\{1, 2, \ldots, M\}$. A simple model might assume that the (z_i) are identically distributed: for $m = 1, \ldots, M$,

$$\mathbb{P}(z_i = m) = \pi_m. \tag{3.4}$$

But more complex models deserve to be considered for practical applications, for instance, the introduction of covariates for defining each individual's probabilities.

In its most general form, a mixture of mixed effects models assumes that there exist M joint distributions $p_1, \ldots, p_M$ and M vector of parameters $\theta_1, \ldots \theta_M$ such that the joint distribution defined in (3.3) now decomposes into

$$p(y_i, \varphi_i; \theta) = \sum_{m=1}^{M} \mathbb{P}(z_i = m)\, p_m(y_i, \varphi_i, \theta_m) \tag{3.5}$$

The mixture can then concern the distribution of the individual parameters $p(\varphi_i; \theta)$ and/or the conditional distribution of the observations $p(y_i|\varphi_i; \theta)$. Let us see some examples of such mixtures models:

i) A latency structure can be introduced at the level of the individual parameters assuming a Gaussian mixture model. This mixture model assumes that there

exist $\mu_1, \Sigma_1, \ldots, \mu_M, \Sigma_M$ such that

$$\varphi_i = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} h(\mu_m, c_i) + \Sigma_m^{-1/2} \eta_i. \tag{3.6}$$

The shape of the mixture depends on the structure of the variance-covariance matrices Σ_m. In the standard case of Gaussian mixture models, $h(\mu_m, c_i) = \mu_m$. Then, $p(\varphi_i; \theta) = \sum_{m=1}^{M} \pi_m \Phi(\varphi_i; \mu_m, \Sigma_m)$, where Φ denotes the d-dimensional Gaussian probability distribution function (pdf). We refer to (Banfield and Raftery, 1993) or (Celeux and Govaert, 1995) for a detailed presentation of such models. Gaussian mixture models are widely used for supervised and unsupervised classification in many applications. But even if the model itself is standard, its use in the context in NLMEM requires particular attention since the individual parameters are not observed. In other words, we aim to create clusters of non observed parameters.

ii) A latency structure can also be introduced at the level of the conditional distribution of the observations (y_{ij}):

$$y_{ij} = f(x_{ij}; \varphi_i, z_i) + g(x_{ij}; \varphi_i, z_i, \theta_y) \varepsilon_{ij}. \tag{3.7}$$

A mixture of conditional distributions therefore reduces to a mixture of residual errors and/or a mixture of structural models.

A mixture of residual error models has the general form:

$$g(x_{ij}; \varphi_i, z_i, \xi) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} g_m(x_{ij}; \varphi_i, \xi_m). \tag{3.8}$$

As an example, a mixture of constant error models assumes that

$$y_{ij} = f(x_{ij}; \varphi_i) + \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \xi_m \varepsilon_{ij}. \tag{3.9}$$

Between subject model mixtures (BSMM) assume that the structural model is a mixture of M different structural models:

$$f(\cdot; \varphi_i, z_i) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} f_m(\cdot; \varphi_i). \tag{3.10}$$

This model is relevant for example to distinguish different types of response to the same treatment. See the next chapter for an application to HIV where different viral kinetics models are used to classify treated patients into responders, non-responders and rebounders on the basis of their viral load profiles.

Remark 3.2.1. *Within subject model mixtures (WSMM) assume a mixture of structural models for each individual, but with differing weights* $\pi_i = (\pi_{i1}, \dots \pi_{iM})$*:*

$$f(\cdot; \varphi_i, p_i) = \sum_{m=1}^{M} \pi_{im} f_m(\cdot; \varphi_i). \tag{3.11}$$

Here, π_i *is an additional vector of individual parameters. This model is a classical NLMEM since there is no latent categorical covariate, so no specific methodology needs to be developed for WSMM.*

3.2.3 Log-likelihood of mixture models

The completed data is (y, φ, z), where y and (φ, z) are respectively the observed and unobserved data. The log-likelihood of the complete data of subject i is

$$\mathcal{L}(y_i, \varphi_i, z_i; \theta) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \left(\mathcal{L}_m(y_i, \varphi_i; \theta_m) + \log \mathbb{P}(z_i = m) \right), \tag{3.12}$$

$\mathcal{L}_m(y_i, \varphi_i; \theta_m)$ being the log-likelihood of the pair of variables (y_i, φ_i) in group G_m defined by

$$G_m = \{i, 1 \leq i \leq N \text{ such that } z_i = m\}.$$

In the case of a mixture of Gaussian distributions as described in (3.6), $\theta_m = (\xi, \mu_m, \Sigma_m)$ and the complete log-likelihood becomes

$$\mathcal{L}_m(y_i, \varphi_i; \theta_m) = \mathcal{L}(y_i | \varphi_i; \xi) + \mathcal{L}_m(\varphi_i; \mu_m, \Sigma_m).$$

For the mixture of structural models (BSMM) defined in (3.10), $\theta_m = (\xi, \mu, \Sigma)$ and

$$\mathcal{L}_m(y_i, \varphi_i; \theta_m) = \mathcal{L}_m(y_i | \varphi_i; \xi) + \mathcal{L}(\varphi_i; \mu, \Sigma),$$

while for the mixture or error models defined in (3.8), $\theta_m = (\xi_m, \mu, \Sigma)$ and

$$\mathcal{L}_m(y_i, \varphi_i; \theta_m) = \mathcal{L}(y_i | \varphi_i; \xi_m) + \mathcal{L}(\varphi_i; \mu, \Sigma).$$

The likelihood of any combination of these different mixture models is straightforward to derive.

In the following, for the sake of clarity, we will make the assumption that the likelihood $\mathcal{L}_m$ belongs to the exponential family: there exists a function ψ of θ_m and a minimal sufficient statistic $T(y_i, \varphi_i)$ such that

$$\mathcal{L}_m(y_i, \varphi_i; \theta_m) = \langle T(y_i, \varphi_i), \theta_m \rangle - \psi(\theta_m). \tag{3.13}$$

According to (3.4), if we assume that $\pi_m = \mathbb{P}(z_i = m)$, then

$$\mathcal{L}(y_i, \varphi_i, z_i; \theta) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \left(\langle T(y_i, \varphi_i), \theta_m \rangle + \log(\pi_m) - \psi(\theta_m) \right). \tag{3.14}$$

Then, the likelihood of the complete model (y, φ, z) also belongs to the exponential family:

$$\mathcal{L}(y, \varphi, z; \theta) = \langle S(y, \varphi, z), \theta \rangle - \psi(\theta), \tag{3.15}$$

where

$$S(y, \varphi, z) = \left(\sum_{i=1}^{n} \mathbb{1}_{z_i=m}, \sum_{i=1}^{n} \mathbb{1}_{z_i=m} T(y_i, \varphi_i); 1 \leq m \leq M \right). \tag{3.16}$$

We will take advantage of this representation of the log-likelihood for our description of the proposed stochastic EM-like algorithms. Indeed, computing any (conditional) expectation of $\mathcal{L}(y, \varphi, z; \theta)$ reduces to computing the (conditional) expectation of $S(y, \varphi, z)$.

Some statistical properties of the MLE for NLMEM can be derived (Online Resource).

3.3 Algorithms proposed for maximum likelihood estimation

We aim to estimate θ by maximizing the likelihood of the observations (y_i). As mentioned above, we are in the general framework of incomplete data where EM-type algorithms are known to be efficient.

First of all, we assume that the complete likelihood $\mathcal{L}(y, \varphi, z; \theta)$ can be maximized when the complete data is observed. In other words, there exists a function $\hat{\theta}$ such that for any (y, φ, z),

$$\hat{\theta}(S(y, \varphi, z)) = \arg\max \{ \langle S(y, \varphi, z), \theta \rangle - \psi(\theta) \}. \tag{3.17}$$

3.3.1 The EM algorithm

Since φ and z are not observed, the EM algorithm replaces $S(y, \varphi, z)$ by its conditional expectation (Dempster et al., 1977; Wu, 1983). Then, given some initial value $\theta^{(0)}$, iteration k of the EM algorithm updates $\theta^{(k-1)}$ into $\theta^{(k)}$ with the two following steps:

- **E-step :** evaluate the quantity

$$s_k = \mathbb{E}\left(S(y, \varphi, z) | y; \theta^{(k-1)} \right).$$

- **M-step:** with respect to (3.17), compute

$$\theta^{(k)} = \hat{\theta}(s_k).$$

Each EM iteration increases the likelihood of observations and the EM sequence $\theta^{(k+1)}$ converges to a stationary point of the observed likelihood under mild regularity conditions (Wu, 1983).

Unfortunately, in the framework of non linear mixed-effects models, there is no explicit expression for the E-step since the relationship between observations y and individual parameters φ is non linear. Several authors have proposed stochastic versions of the EM algorithm which attempt to solve the problem. (Wei and Tanner, 1990) proposed the Monte Carlo EM (MCEM) algorithm in which the E-step is replaced by a Monte Carlo approximation based on a large number of independent simulations of the missing data. In recent work, (Wang et al., 2007; Wang et al., 2009) also proposed an MCEM algorithm with importance sampling.

Another EM type algorithm for mixtures of mixed effects models was proposed in (De la Cruz-Mesia et al., 2008). They use an extensive Monte-Carlo integration procedure during the E step for computing the marginal distribution of the observations in each cluster. Unfortunately, the computational effort required by this method is prohibitive for most practical application since the structural model f needs to be evaluated T times (here T is the Monte-Carlo size), at each iteration of the algorithm and for each patient. Furthermore, the authors claim that their procedure converges if conditions that ensure the convergence of EM are fulfilled. This is true in "theory", with an infinite Monte-Carlo size, but nothing can be said in realistic conditions.

We will see in the next sections that the proposed modified SAEM algorithm offers appealing practical and theoretical properties. Indeed, convergence of the algorithm is demonstrated under general conditions. Moreover it is extremely fast and can be used for complex problems.

3.3.2 The SAEM algorithm

The stochastic approximation version of the EM algorithm, proposed by (Delyon et al., 1999), consists of replacing the E-step by a stochastic approximation obtained using simulated data. Given some initial value $\theta^{(0)}$, iteration k of SAEM consists of the three following steps:

- **S-step:** draw $\left(z^{(k)}, \varphi^{(k)}\right)$ with the conditional distribution $p\left(z, \varphi | y, \theta^{(k-1)}\right)$.
- **AE-step:** update s_k according to

$$s_k = s_{k-1} + \delta_k \left(S\left(y, \varphi^{(k)}, z^{(k)}\right) - s_{k-1} \right). \tag{3.18}$$

- **M-step:** compute $\theta^{(k)} = \hat{\theta}(s_k)$.

Here, (δ_k) is a decreasing sequence. In the case of NLMEM, the simulation step cannot be directly performed, and a MCMC procedure can be used (Kuhn and Lavielle, 2004). Convergence of the parameter sequence $(\theta^{(k)})$ toward a (local) maximum of the likelihood is ensured under general conditions (see (Delyon et al., 1999; Kuhn and Lavielle, 2004; Allassonnière et al., 2010)).

This version of SAEM for mixtures of NLMEM was first implemented in the MONOLIX software. We have noticed that the algorithm tends to become unstable and produces poor estimations when the problem becomes difficult: small sample sizes, heteroscedastic models, overlap between mixture components, etc. This poor behavior is mainly due to the fact that the S-step of SAEM requires simulation of the categorical variable (z_i), which then impacts the M-step, leading to inference problems. Due to the well known label-switching phenomenon, as pointed out by (Celeux et al., 2000), uniform ergodicity of the Markov chain $\left(\varphi^{(k)}, z^{(k)}\right)$ is no longer guaranteed and convergence of SAEM can not be ensured. We also have noticed that some components of the mixture can disappear during iterations, mainly when these components are not well separated. In the next section, we propose a methodology that avoids simulation of these latent categorical covariates and exhibits improved practical behavior.

3.3.3 The MSAEM algorithm

We have seen that the E-step of the EM algorithm requires evaluating $\mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta^{(k-1)}\right)$. We have the following relation:

$$\mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta\right) = \mathbb{E}\left(\mathbb{E}\left(S\left(y,\varphi,z\right)|y,\varphi,\theta\right)|y;\theta\right).$$

Then, by setting

$$H\left(y,\varphi,\theta\right) = \mathbb{E}\left(S\left(y,\varphi,z\right)|y,\varphi,\theta\right), \tag{3.19}$$

the E-step of the EM algorithm at iteration k reduces to calculating

$$\mathbb{E}\left(H\left(y,\varphi,\theta^{(k-1)}\right)|y;\theta^{(k-1)}\right).$$

The underlying idea in this operation is to use a conditional distribution that only depends on φ and y but not the latent categorical covariates z. Then, φ becomes the only unobserved variable of the model. Nevertheless, introduction of the latent categorical variable remains very useful since it allows one to derive a manageable expression of the complete likelihood. Then, iteration k of MSAEM requires us to calculate $H\left(y,\varphi;\theta^{(k-1)}\right)$:

- **S-step:** draw $\varphi^{(k)}$ with the conditional distribution $p\left(.|y,\theta^{(k-1)}\right)$.
- **E-step:** compute $H\left(y,\varphi^{(k)};\theta^{(k-1)}\right)$ using (3.19).
- **AE-step:** update s_k according to

$$s_k = s_{k-1} + \delta_k\left(H\left(y,\varphi^{(k)},\theta^{(k-1)}\right) - s_{k-1}\right). \tag{3.20}$$

- **M-step :** compute $\theta^{(k)} = \hat{\theta}(s_k)$.

The simulation step of the MSAEM algorithm at iteration k consists of a few MCMC iterations with $p(\phi|y;\theta^{(k)})$ as the stationary distribution. More precisely, we propose to use the Hasting-Metropolis algorithm, with various proposal kernels. Here, the N subjects are assumed to be independent and the same procedure is used for the N subjects, i.e., for $i = 1, 2, \ldots, N$.

A first kernel consists in using the marginal distribution $p(\phi_i)$ for generating a candidate ϕ_i^c. Then, the probability of acceptance, i.e., the probability to move from ϕ_i to ϕ_i^c, reduces to

$$\alpha\left(\varphi_i, \varphi_i^c\right) = \min\left(1, \frac{p\left(y|\varphi_i^c; \theta^{(k)}\right)}{p\left(y|\varphi_i; \theta^{(k)}\right)}\right).$$

Another possible kernel is the random walk: $\varphi_i^c \sim \mathcal{N}\left(\varphi^{(k-1)}, \Omega\right)$, where Ω is a diagonal matrix which is adaptively adjusted in order to reach a given acceptance rate (typically 0.3). Different directions can be used by setting different elements of the diagonal of Ω to 0 during iteration. Here, the probability of acceptance is

$$\alpha\left(\varphi_i, \varphi_i^c\right) = \min\left(1, \frac{p\left(y, \varphi_i^c; \theta^{(k)}\right)}{p\left(y, \varphi_i; \theta^{(k)}\right)}\right).$$

Practical implementation of this algorithm requires computing $p(\varphi_i)$ and $p(y|\varphi_i)$. Depending on the type of mixture model considered (mixture of distributions, mixture of structural models, mixture of residual error models,etc.), these two terms can easily be computed in a closed form.

Certain parameters need to be well chosen to improve the convergence of the algorithm, such as the total number of iterations K, the number of iterations of the MCMC procedure during the S-step, and the step-size sequence (δ_k). We remark that selection of the various settings of the algorithm is not a problem related to the particular extension of SAEM to mixture models considered here, but a general issue for practical implementation of SAEM. We will give some leads, but an in-depth discussion of the choice of these settings is beyond the scope of the paper.

Sequence (δ_k) has a strong impact on the speed of convergence of the algorithm. Fast convergence towards a neighborhood of the solution is obtained with a constant sequence $\delta_k = 1$ during the first K_1 iterations of SAEM. Then, the M-step of SAEM reduces to maximizing the complete log-likelihood: for $k = 1, \ldots, K_1$,

$$\theta^{(k)} = \arg\max \mathcal{L}\left(y, \varphi^{(k)}; \theta\right).$$

Thus, if we consider a mixture model, the M-step consists of estimating the components of the mixtures using the observations y and the simulated individual parameters $\varphi^{(k)}$. An EM can be used at iteration k for computing $\theta^{(k)}$. After converging to a neighborhood of the MLE, a decreasing step-size sequence (δ_k) will permit almost sure convergence of the algorithm to a maximum of the observed likelihood (Section 3..2). For the numerical experiments presented below, (δ_k) decreases as $1/k$.

When the number of subjects N is small, convergence of the algorithm can be improved by combining the stochastic approximation with Monte-Carlo, i.e., by

running R Markov chains in parallel instead of only one chain. The S-step now consists of generating R sequences $\varphi^{(k,1)}, \ldots, \varphi^{(k,R)}$, and (3.20) becomes

$$s_k = s_{k-1} + \delta_k \left(\frac{1}{R} \sum_{r=1}^{R} H\left(y, \varphi^{(k,r)}, \theta^{(k-1)}\right) - s_{k-1} \right).$$

For the numerical experiments, we have set $R = 5$ with $N = 100$.

3.3.4 Some examples

Mixtures of Gaussian distributions

We consider here that the distributions of the individual parameters is a mixture of Gaussian distributions

$$\varphi_i \sim_{\text{iid}} \sum_{m=1}^{M} \pi_m \mathcal{N}\left(\mu_m, \Sigma_m\right)$$

Let $\pi = (\pi_1, \ldots, \pi_M)$, $\mu = (\mu_1, \ldots, \mu_M)$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_M)$. Then, the conditional log-likelihood of the individual parameters φ is given by

$$\begin{aligned} \mathcal{L}\left(\varphi | z; \mu, \Sigma\right) &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \left(d \log(2\pi) + \log\left|\Sigma_m\right| \right) \\ &- \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{z_i=m} (\varphi_i - \mu_m)' \Sigma_m^{-1} (\varphi_i - \mu_m), \end{aligned}$$

and the log-likelihood of the labels z is:

$$\mathcal{L}\left(z; \pi\right) = \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \log(\pi_m). \tag{3.21}$$

On the other hand, we consider a proportional residual model:

$$y_{ij} = f\left(x_{ij}, \varphi_i\right) + \xi f\left(x_{ij}, \varphi_i\right) \varepsilon_{ij}.$$

Then, the conditional log-likelihood of the observations y is given by

$$\begin{aligned} \mathcal{L}\left(y | \varphi, z; \xi\right) &= -\sum_{i,j} \log\left(\xi f\left(x_{ij}, \varphi_i\right)\right) - \frac{N_{\text{tot}}}{2} \log\left(2\pi\right) \\ &- \frac{1}{2\xi^2} \sum_{i,j} \left(\frac{y_{ij} - f\left(x_{ij}, \varphi_i\right)}{f\left(x_{ij}, \varphi_i\right)} \right)^2, \end{aligned} \tag{3.22}$$

where $N_{\text{tot}} = \sum_{i=1}^{N} n_i$ is the total number of observations. Sufficient statistics of the complete model are:

$$S = (S_{1,m}, S_{2,m}, S_{3,m}, S_4; 1 \leq m \leq M),$$

where

$$S_{1,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m} \tag{3.23}$$

$$S_{2,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m} \varphi_i \tag{3.24}$$

$$S_{3,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m} \varphi_i \varphi_i' \tag{3.25}$$

$$S_4 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(y_{ij} / f\left(x_{ij}, \varphi_i\right) - 1 \right)^2, \tag{3.26}$$

and the function $\hat{\theta}$ is given by:

$$\hat{\pi}_m(S) = S_{1,m}/N \tag{3.27}$$

$$\hat{\mu}_m(S) = S_{2,m}/S_{1,m} \tag{3.28}$$

$$\hat{\Sigma}_m(S) = \frac{S_{3,m}}{S_{1,m}} - \left(\frac{S_{2,m}}{S_{1,m}}\right)\left(\frac{S_{2,m}}{S_{1,m}}\right)' \tag{3.29}$$

$$\hat{b}(S) = \sqrt{S_4/N_{\text{tot}}}. \tag{3.30}$$

At iteration k, SAEM requires using simulated sequences $\varphi^{(k)}$ and $z^{(k)}$ for updating the set of statistics defined in (3.23-3.26) using the stochastic approximation scheme defined in (3.18).

Instead, at iteration k, MSAEM consists of using only the simulated sequence $\varphi^{(k)}$ for computing $H\left(y, \varphi; \theta^{(k-1)}\right)$ using (3.19), and updating the set of statistics using the stochastic approximation scheme defined in (3.20).

The minimal sufficient statistic here is $\mathbb{1}_{z_i=m}$, and the E-step of iteration k of MSAEM reduces to the evaluation of:

$$\begin{aligned}
\gamma_{i,m}^{(k)} &= \mathbb{E}\left(\mathbb{1}_{z_i=m} | y_i, \varphi_i^{(k)}; \theta^{(k-1)}\right) \\
&= \mathbb{P}\left(z_i = m | y_i, \varphi_i^{(k)}; \theta^{(k-1)}\right) \\
&= \mathbb{P}\left(z_i = m | \varphi_i^{(k)}; \mu^{(k-1)}, \Sigma^{(k-1)}, p^{(k-1)}\right) \\
&= \frac{\pi_m^{(k-1)} \ell(\varphi_i^{(k)}; \mu_m^{(k-1)}, \Sigma_m^{(k-1)})}{\sum_{r=1}^{M} \pi_r^{(k-1)} \ell(\varphi_i^{(k)}; \mu_r^{(k-1)}, \Sigma_r^{(k-1)})},
\end{aligned}$$

where ℓ is the pdf of a Gaussian vector.

The zero-one variable $\mathbb{1}_{z_i=m}$ present in expressions when applying SAEM is replaced at iteration k in MSAEM by the probability $\mathbb{P}\left(z_i = m | \varphi_i^{(k)}, \theta^{(k-1)}\right)$, and

permits us to tackle the problems mentioned before. Then, the A-step of MSAEM reduces to:

$$
\begin{aligned}
s_{k,1,m} &= s_{k-1,1,m} + \delta_k \left(\sum_{i=1}^{N} \gamma_{i,m}^{(k)} - s_{k-1,1,m} \right) \\
s_{k,2,m} &= s_{k-1,2,m} + \delta_k \left(\sum_{i=1}^{N} \gamma_{i,m}^{(k)} \varphi_i^{(k)} - s_{k-1,2,m} \right) \\
s_{k,3,m} &= s_{k-1,3,m} + \delta_k \left(\sum_{i=1}^{N} \gamma_{i,m}^{(k)} \varphi_i^{(k)} \varphi_i^{(k)'} - s_{k-1,3,m} \right)
\end{aligned}
$$

$$
s_{k,4} = s_{k-1,4} + \delta_k \left(\sum_{i,j} \left(\frac{y_{ij} - f(x_{ij}, \varphi_i^{(k)})}{f(x_{ij}, \varphi_i^{(k)})} \right)^2 - s_{k-1,4} \right).
$$

Parameters are then updated using the function $\hat{\theta}$ defined above.

Mixtures of residual error models

Suppose now a proportional residual model in each group, given by $g_m\left(x_{ij}, \varphi_i, \theta_{y,m}\right) = \xi_m f\left(x_{ij}, \varphi_i\right)$, where $\xi = (\xi_1, \ldots, \xi_m)$. Then, the conditional log-likelihood of the observations in group G_m is now:

$$
\begin{aligned}
\mathcal{L}_m\left(y_i|\varphi_i;\xi\right) &= \mathcal{L}\left(y_i|\varphi_i;\xi_m\right) \\
&= -\frac{1}{2\xi_m^2} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - f\left(x_{ij},\varphi_i\right)}{f\left(x_{ij},\varphi_i\right)} \right)^2 \\
&\quad - \sum_{j=1}^{n_i} \log\left(\xi_m f\left(x_{ij},\varphi_i\right)\right) - \frac{n_i}{2}\log\left(2\pi\right),
\end{aligned}
$$

On the other hand, assume a unique Gaussian distribution for the individual parameters, then the likelihood of the individual parameters is

$$
\begin{aligned}
\mathcal{L}\left(\varphi_i;\mu,\Sigma\right) &= -\frac{1}{2}\left(\varphi_i - \mu\right)' \Sigma^{-1} \left(\varphi_i - \mu\right) \\
&\quad -\frac{d}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(|\Sigma|\right).
\end{aligned}
$$

Here, the E-step requires computing:

$$
\begin{aligned}
\gamma_{i,m}^{(k)} &= \mathbb{P}\left(z_i = m | y_i, \varphi_i^{(k)}, \theta^{(k-1)}\right) \\
&= \frac{\pi_m^{(k-1)} \ell\left(y_i|\varphi_i^{(k)};\xi_m^{(k-1)}\right)}{\sum_{r=1}^{M} \pi_r^{(k-1)} \ell\left(y_i|\varphi_i^{(k)};\xi_m^{(k-1)}\right)}.
\end{aligned}
$$

In the A-step, we approximate several minimal sufficient statistics as follows:

$$\begin{aligned} s_{k,i,1,m} &= s_{k-1,i,1,m} + \delta_k \left(\sum_{i=1}^{N} \gamma_{i,m}^{(k)} - s_{k-1,i,1,m} \right) \\ s_{k,2} &= s_{k-1,2} + \delta_k \left(\sum_{i=1}^{N} \varphi_i^{(k)} - s_{k-1,2} \right) \\ s_{k,3} &= s_{k-1,m} + \delta_k \left(\sum_{i=1}^{N} \varphi_i^{(k)} \varphi_i^{(k)\prime} - s_{k-1,m} \right) \\ s_{k,4,m} &= s_{k-1,4,m} \\ +\delta_k \left(\sum_{i,j} \gamma_{i,m}^{(k)} \left(\frac{y_{ij} - f_m\left(x_{ij}, \varphi_i^{(k)}\right)}{f_m\left(x_{ij}, \varphi_i^{(k)}\right)} \right)^2 - s_{k-1,4,m} \right). \end{aligned}$$

In the M-step, we update parameters according to:

$$\begin{aligned} \pi_m^{(k)} &= \frac{1}{N} \sum_{i=1}^{N} s_{k,1,i,m} \\ \mu^{(k)} &= \frac{s_{k,2,i}}{N} \\ \Sigma^{(k)} &= \frac{s_{k,3}}{N} - \left(\frac{s_{k,2,i}}{N}\right)\left(\frac{s_{k,2,i}}{N}\right)' \\ \xi_m^{(k)} &= \sqrt{\frac{s_{k,4,m}}{\sum_{i=1}^{N} n_i s_{k,1,i,m}}}. \end{aligned}$$

3.3.5 Estimation of the individual parameters

For a given set of population parameters θ, we use each individual conditional distribution $p(z_i, \phi_i | y_i, \theta)$ for estimating the latent variable z_i and the vector of individual parameters ϕ_i.

A first estimate is the Maximum a Posteriori (MAP) which is obtained by maximizing this joint conditional distribution with respect to (z_i, ϕ_i):

$$(\hat{z}_i, \hat{\varphi}_i) = \arg\max_{(z_i, \varphi_i)} p\left(z_i, \varphi_i | y, \theta\right) \tag{3.31}$$

Such a maximization is not straightforward and requires performing a two-step procedure:

1) For $m = 1, \ldots, M$ compute

$$\hat{\phi}_{i,m} = \arg\max_{\phi_i} p(y_i | \phi_i, z_i = m; \theta) p(\phi_i | z_i = m; \theta) \tag{3.32}$$

2) Compute

$$\hat{m}_i = \arg\max_m p(y_i, \hat{\phi}_{i,m}|z_i = m;\theta)\mathbb{P}\left(z_i = m;\theta\right) \tag{3.33}$$

and set

$$(\hat{z}_i, \hat{\varphi}_i) = (\hat{m}_i, \hat{\varphi}_{i,\hat{m}_i})\,. \tag{3.34}$$

Another estimate of the latent covariate z_i maximizes the marginal conditional distribution:

$$\hat{z}_i \;=\; \arg\max_m \mathbb{P}\left(z_i = m|y_i;\theta\right), \tag{3.35}$$

where

$$\begin{aligned}\mathbb{P}\left(z_i = m|y_i,\theta\right) &= \mathbb{E}\left(\mathbb{P}\left(z_i = m|y_i,\varphi_i,\theta\right)|y_i,\theta\right)\\ &= \mathbb{E}\left(\gamma_{i,m}|y_i,\theta\right),\end{aligned} \tag{3.36}$$

which can be estimated using the stochastic approximation procedure described in Section 3.3.4.

Instead of maximizing the conditional distribution for estimating φ_i, an alternative is to compute the conditional mean

$$\hat{\varphi}_i \;=\; \mathbb{E}\left(\varphi_i|y_i;\theta\right). \tag{3.37}$$

We remark that

$$\mathbb{E}\left(\varphi_i|y_i;\theta\right) = \sum_{m=1}^{M} \mathbb{E}\left(\varphi_i|y_i, z_i = m;\theta\right)\mathbb{P}\left(z_i = m|y_i;\theta\right).$$

Then, estimating the conditional expectation of the individual parameters requires estimating the conditional probabilities $\mathbb{P}\left(z_i = m|y_i;\theta\right)$ and the conditional means in each group, $\mathbb{E}\left(\varphi_i|y_i, z_i = m;\theta\right)$.

We have seen above how to estimate $\mathbb{P}\left(z_i = m|y_i;\theta\right)$ using stochastic approximation. On the other hand, $\mathbb{E}\left(\varphi_i|y_i, z_i = m;\theta\right)$ can easily be estimated by MCMC.

3.4 Numerical experiments

A simulation study was conducted to evaluate the performance of the proposed algorithm for estimating the parameters of the different non linear mixed-effects mixture models. We used a pharmacokinetics (PK) model for these numerical experiments. The vector of individual PK parameters of subject i is

$$\varphi_i = \left(\log(ka_i), \log(V_i), \log(Cl_i)\right), \tag{3.38}$$

where V_i is the volume of distribution, Cl_i the clearance and ka_i the absorption constant rate of the subject. We define φ_i as the set of log-parameters, since

log-normal distributions will be used for describing the inter-subject variability of these PK parameters.

The structural model is an oral administration PK model with one compartment, first order absorption and linear elimination. The plasmatic concentration of drug predicted by the model is:

$$f\left(\varphi_i, x_{ij}\right) = \frac{D_i ka_i}{V_i\left(ka_i - \frac{Cl_i}{V_i}\right)}\left(e^{-\frac{Cl_i}{V_i}x_{ij}} - e^{-ka_i x_{ij}}\right). \tag{3.39}$$

Here, (x_{ij}) are the measurement times of subject i, and D_i the dose administered at time 0. We use the same amount of drug $D_i = 1000$mg and the same measurement times for the N subjects (times are in hours):

$$x_i = (0.25,\ 1,\ 2.5,\ 6,\ 16,\ 26,\ 72)$$

For each of the four examples considered, $L = 100$ datasets were simulated and the parameters were estimated using MSAEM. Let $\theta^\star$ be the parameter used for the simulation and $\hat{\theta}_\ell$ be the estimated parameter obtained with the ℓ-th simulated dataset. The following quantities were computed:

◇ the relative estimation errors (in %): for $1 \leq \ell \leq L$,

$$\text{REE}_\ell = \frac{\hat{\theta}_\ell - \theta^\star}{\theta^\star} \times 100.$$

◇ The relative root mean square error (in %):

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}_\ell - \theta^\star)^2}}{\theta^\star} \times 100.$$

3.4.1 Mixtures of distributions

We assume a proportional error model for the observed concentration:

$$y_{ij} = f\left(\varphi_i, x_{ij}\right) + \xi f\left(\varphi_i, x_{ij}\right)\varepsilon_{ij}, \tag{3.40}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and $\xi = 0.2$. We assume that ka_i and Cl_i are log-normally distributed:

$$\begin{aligned} \log\left(ka_i\right) &\sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), \\ \log\left(Cl_i\right) &\sim \mathcal{N}\left(\mu_3, \sigma_3^2\right), \end{aligned}$$

with $\mu_1 = 1$, $\mu_3 = 4$, $\sigma_1^2 = 0.04$ and $\sigma_3^2 = 0.04$. Mixtures of distributions will be used for V_i. Scenarios 1 and 2 assume a homoscedastic model:

$$\log\left(V_i\right) \sim p_1\mathcal{N}\left(\mu_{21}, \sigma_2^2\right) + p_2\mathcal{N}\left(\mu_{22}, \sigma_2^2\right),$$

with the following numerical values

S1: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 70$, $\sigma_2^2 = 0.04$,
S2: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 50$, $\sigma_2^2 = 0.04$.
The difference between the two means is significantly reduced in Scenario 2 compared to Scenario 1.

Scenario 3 assumes an heteroscedastic model:

$$\log(V_i) \sim p_1\mathcal{N}\left(\mu_{21}, \sigma_{21}^2\right) + p_2\mathcal{N}\left(\mu_{22}, \sigma_{22}^2\right),$$

with
S3: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 50$, $\sigma_{21}^2 = 0.08$, $\sigma_{22}^2 = 0.04$.
Fig. 5.1 displays the probability distribution functions of $\log(V_i)$ under the three scenarios. Distributions are well separated in Scenario 1. Overlapping between the two distributions increases in Scenario 2 since the two distribution become closer. Increasing one of the variances in Scenario 3 increases further this overlapping.

Fig. 5.2 displays the distribution of the observed concentration in both groups under each scenario. Medians and 90% confidence intervals are used to summarize these distributions.

Results obtained with the MSAEM algorithm are displayed Fig. 5.3. We show the distribution of the relative estimation errors (REE_ℓ) for each parameter under each scenario, with $N = 100$ and $N = 1000$ subjects. Relative root mean square errors are presented in Table 4.1. These are compared with those obtained in differing situations, i.e., when φ and/or z are known.

Results obtained with scenario S1 are very similar whether or not z is known. Indeed, the two components of the mixture are well separated here, and the conditional probabilities of belonging to each class are close to 0 or 1. The results deteriorate with scenarios S2 and S3 for the parameters of the mixture which are much less-well estimated when z is unknown. It is also interesting to notice that the other model parameters are little affected by knowledge of z.

Lastly, we remark that the differences when individual parameters (φ_i) are known or not have little impact on the estimation of the parameters of the mixture. The most difficult parameter to estimate when (φ_i) is unknown is the variance of $\log(ka_i)$. This is a purely statistical issue related to the quantity of information in the data, and independent of the mixture model: few observations are available during the absorption phase which makes it difficult to estimate the absorption rate constant ka. The boxplots confirm that parameters are better estimated with $N = 1000$, but even with $N = 100$, we do not see any bias in the estimation of the parameters, with the exception perhaps of the variances of the mixture in scenario S3, which are poorly estimated.

Fig. 5.4 provides a graphical illustration of the probability of correct classification in both groups for the different scenarios and $N = 1000$ subjects. For each of the $K = 100$ runs, the probabilities of correct classification were computed as well as the median of these $K = 100$ sequences for each individual in each group.

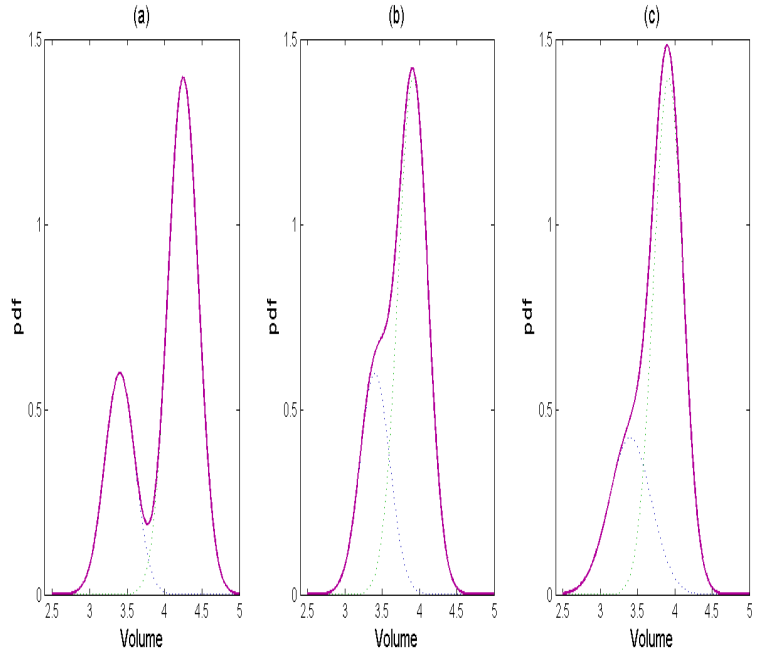

Figure 3.1: Probability distribution function of the log-volume, (a) scenario S1: $\mu_1 = 30$, $\mu_2 = 70$, $\sigma^2_{21} = \sigma^2_{22} = 0.04$; (b) scenario S2: $\mu_1 = 30$, $\mu_2 = 50$, $\sigma^2_{21} = \sigma^2_{22} = 0.04$; (c) scenario S3: $\mu_1 = 30$, $\mu_2 = 50$, $\sigma^2_{21} = 0.08$, $\sigma^2_{22} = 0.04$.

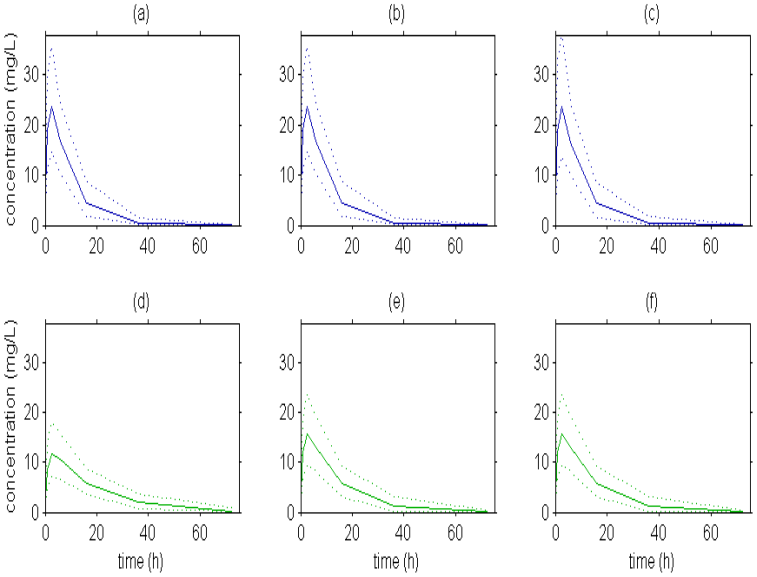

Figure 3.2: Median (solid line) and 90% prediction interval (dotted line) of the observed concentration in different groups: (a-c) group 1, (d-f) group 2, and with different scenarios: (a)&(d) S1, (b)&(e) S2, (c)&(f).

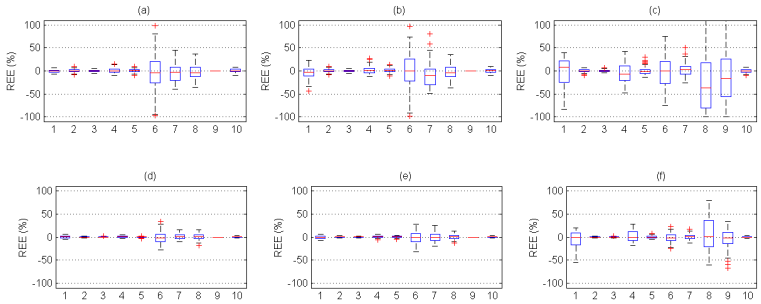

Figure 3.3: Empirical distribution of the relative estimation errors (REE_ℓ) with different sample sizes: (a-c) $N = 100$, (d-f) $N = 1000$, and different scenarios: (a)&(d) S1, (b)&(e) S2, (c)&(f) S3. The estimated parameters are 1: p_2; 2: μ_1; 3: μ_{21}; 4: μ_{22}; 5: μ_3; 6: σ_1^2; 7: σ_{21}^2; 8: σ_{22}^2 (only in S3, *i.e.* (c)&(f)); 9: σ_3^2; 10: ξ.

		N=100			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	6.74	6.95	6.07	6.87
μ_1	1	1.98	1.98	2.67	2.96
μ_{21}	30	3.30	3.95	4.25	5.35
μ_{22}	70	2.26	2.42	2.57	3.19
μ_3	4	1.75	1.75	2.13	2.24
σ_1^2	0.04	15.80	15.80	38.93	40.46
σ_2^2	0.04	13.48	13.99	14.67	18.91
σ_3^2	0.04	13.88	13.88	19.93	16.07
b	0.20	2.58	2.58	3.87	4.00
		N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	2.04	2.17	2.19	2.21
μ_1	1	0.68	0.68	1.07	0.93
μ_{21}	30	1.24	1.40	1.50	1.73
μ_{22}	70	0.75	0.78	0.91	0.95
μ_3	4	0.62	0.62	0.69	0.65
σ_1^2	0.04	4.18	4.18	12.80	11.44
σ_2^2	0.04	5.51	5.95	5.09	5.38
σ_3^2	0.04	4.45	4.45	6.93	6.38
b	0.20	0.91	0.91	1.23	1.22

Table 3.1: Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 1, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

		N=100			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	7.06	10.84	6.07	12.34
μ_1	1	1.92	1.92	2.64	2.98
μ_{21}	30	3.83	6.32	4.15	7.91
μ_{22}	50	2.63	3.83	2.45	4.91
μ_3	4	1.91	1.91	2.15	2.29
σ_1^2	0.04	14.23	14.23	37.85	38.09
σ_2^2	0.04	15.56	21.64	19.57	26.54
σ_3^2	0.04	13.91	13.91	14.92	16.11
b	0.20	2.59	2.59	3.88	4.08
		N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	1.88	3.06	2.19	3.73
μ_1	1	0.56	0.56	1.11	1.08
μ_{21}	30	1.13	1.83	1.53	2.33
μ_{22}	50	0.71	1.04	0.93	1.65
μ_3	4	0.59	0.59	0.69	0.66
σ_1^2	0.04	4.48	4.48	12.48	11.34
σ_2^2	0.04	5.02	6.62	5.17	9.37
σ_3^2	0.04	4.64	4.64	6.57	5.18
b	0.20	0.84	0.84	1.23	1.28

Table 3.2: Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 2, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

		N=100			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	0.00	18.81	6.06	32.17
μ_1	1	1.95	1.95	2.55	3.27
μ_{21}	30	6.01	15.77	5.75	22.80
μ_{22}	50	2.56	4.80	2.44	7.85
μ_3	4	1.85	1.85	2.16	2.24
σ_1^2	0.04	14.15	14.15	35.05	36.90
σ_{21}^2	0.08	24.62	53.45	30.95	64.80
σ_{22}^2	0.04	23.93	34.14	21.22	60.60
σ_3^2	0.04	15.77	15.77	14.64	15.20
b	0.20	3.07	3.07	3.82	3.30
		N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	0.00	10.76	2.18	17.30
μ_1	1	0.60	0.60	1.05	1.02
μ_{21}	30	1.62	8.34	1.98	12.12
μ_{22}	50	0.74	1.96	0.92	2.61
μ_3	4	0.63	0.63	0.70	0.74
σ_1^2	0.04	4.58	4.58	12.52	9.94
σ_{21}^2	0.08	8.18	24.84	11.14	34.71
σ_{22}^2	0.04	7.34	13.01	7.49	20.75
σ_3^2	0.04	4.02	4.02	5.32	5.96
b	0.20	0.78	0.78	1.27	1.17

Table 3.3: Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 3, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

These medians were then ranked in increasing order in each group. We repeated then the same procedure, but assuming that the individual parameters (φ_i) were known. Fig.5.4 compares these two medians. The y-axis of Fig. 5.4 represents the Monte Carlo median probability of correct classification ranked in increasing order. As expected, for each scenario, the probabilities of correct classification are greater when (φ_i) is known, but it is interesting to notice that the difference is relatively small. As already mentioned, the difficulty of the estimation problem increases from Scenario 1 to Scenario 3. We see that the difficulty of the classification problem also increases: it is obviously much more difficult to correctly classify the subjects under Scenario 3, where there is a lot of overlap, than under Scenario 1, where the two distributions are well separated.

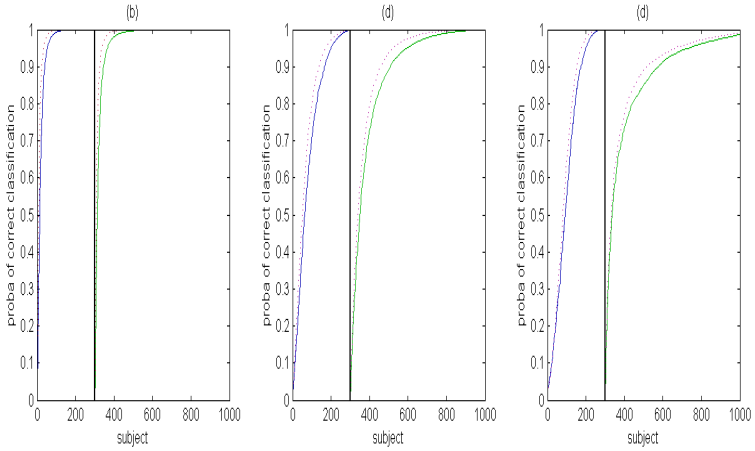

Figure 3.4: Medians of the probabilities of correct classification ranked in increasing order in both groups for the different scenarios and $N = 1000$ subjects. Blue: group 1 ; green: group 2. Solid line: the individual parameters (φ_i) are unknown ; dotted line: the individual parameters (φ_i) are known.

3.4.2 Mixtures of error models

We still use the same PK model, but assuming now a mixture of residual error models:

$$\begin{aligned} \text{if } z_i = 1 \quad , \qquad & y_{ij} = f(\varphi_i, x_{ij}) + \xi_1 f(\varphi_i, x_{ij})\, \varepsilon_{ij}, \\ \text{if } z_i = 2 \quad , \qquad & y_{ij} = f(\varphi_i, x_{ij}) + \xi_2 f(\varphi_i, x_{ij})\, \varepsilon_{ij}, \end{aligned}$$

with $\mathbb{P}(z_i = 1) = 0.3$ and $\xi_1 = 0.1$, $\xi_2 = 0.2$.

We assume that ka_i, V_i and Cl_i are log-normally distributed:

$$\begin{aligned} \log(ka_i) &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ \log(V_i) &\sim \mathcal{N}(\mu_2, \sigma_2^2) \\ \log(Cl_i) &\sim \mathcal{N}(\mu_3, \sigma_3^2). \end{aligned}$$

with $\mu_1 = 1$, $\mu_2 = 30$, $\mu_3 = 4$, $\sigma_1^2 = 0.04$, $\sigma_2^2 = 0.04$ and $\sigma_3^2 = 0.04$.

Numerical results are summarized Table 3.4. Comments we can make about these results are similar to those for the previous examples: the fact that z is known or unknown affects mainly the estimation of parameters of the mixture model: proportions p_1 and p_2 and standard deviations ξ_1 and ξ_2.

		N=100			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.3	6.73	11.76	6.06	20.97
μ_1	1	1.94	1.94	2.70	2.92
μ_2	30	1.98	1.98	2.28	2.29
μ_3	4	1.84	1.84	2.18	2.23
σ_1^2	0.04	15.94	15.94	26.00	36.91
σ_2^2	0.04	12.36	12.36	14.21	13.62
σ_3^2	0.04	15.38	15.38	15.23	16.07
b_1	0.10	5.36	9.41	6.38	19.60
b_2	0.20	3.17	4.05	4.28	14.35
		N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.3	2.01	2.93	2.18	4.80
μ_1	1	0.60	0.60	0.96	0.87
μ_2	30	0.64	0.64	0.72	0.73
μ_3	4	0.56	0.56	0.67	0.65
σ_1^2	0.04	4.67	4.67	10.66	9.33
σ_2^2	0.08	4.80	4.80	5.89	5.26
σ_3^2	0.04	4.74	4.74	4.95	4.57
b_1	0.10	1.84	2.55	2.49	5.08
b_2	0.20	0.96	1.14	1.30	1.97

Table 3.4: Relative Root Mean Square Errors (RRMSE) of parameter estimates in Scenario 4, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

3.5 An application to PK data

For confidentiality reasons, neither the drug nor the pharma company can be cited in this application to real-word PK data, but the example is interesting and challenging. A drug X was orally administrated to 199 patients. Each patients received one dose per day, during a period that varies between 1 and 14 days. The pharmacokinetics model that was shown to better describe the process is a 2

compartment models with linear absorption and linear elimination:

$$\begin{aligned}\dot{A}_d(t) &= -k_a A_d(t)\\ \dot{A}_c(t) &= k_a A_d(t) - k_e A_c(t) - k_{12} A_c(t) + k_{21} A_p(t)\\ \dot{A}_p(t) &= k_{12} A_c(t) - k_{21} A_p(t)\end{aligned}$$

where A_d is the amount of drug in the depot compartment, A_c the amount in the central compartment and A_p the amount in the peripheral compartment. There is no drug in any compartment before the administration of the drug: for any $t < 0$, $A_d(t) = A_c(t) = A_p(t) = 0$. If an amount D of drug is administrated at time τ, then $A_d(\tau^+) = A_d(\tau^-) + D$.

The concentration of drug $Cc = Ac/V$ is measured in the central compartment, where V is the volume of the central compartment.

Here, the individual PK parameters $(k_a, V, k_e, k_{12}, k_{21})$ are log-normally distributed. Then, the Gaussian vector φ_i is the vector of log-parameters. The variance-covariance matrix of φ_i is assumed to be diagonal.

The measured log-concentration is assumed to be normally distributed with a constant error model:

$$\log(y_{ij}) = \log(Cc(t_{ij}; \phi_i)) + a\varepsilon_{ij}.$$

We first used MONOLIX to fit this NLMEM to the PK data. Observed concentration data from 4 patients with their concentrations predicted by the model are displayed Figure 3.5.

We then used a mixture of two log-normal distributions for modelling the distribution of the PK parameters. We used a forward strategy for selecting the best mixture model: we first assumed that only one of the five PK parameter distributions was a mixture of two distributions and we compared the five possible models with only one component modelled as a mixture. The model with the highest likelihood value was selected (k_e was selected). We then looked for a second parameter among the four remaining ones (k_{12} was selected), then a third parameter among the three remaining ones (V was selected), and lastly the best combination of four mixtures (k_a was selected). We then used the BIC criteria for comparing the six selected models, including the model without any mixture component and that with all the parameters modelled with mixtures. All the results are summarized Table 3.5. The final model was a model assuming that the distributions of V, k_e and k_{12} are mixtures of log-normal distributions. The five estimated distributions of the five PK parameters are displayed Figure 3.6.

3.6 Discussion

There exist very few methods available for maximum likelihood estimation in mixtures of mixed effects models. Methods implemented in the nlme R package

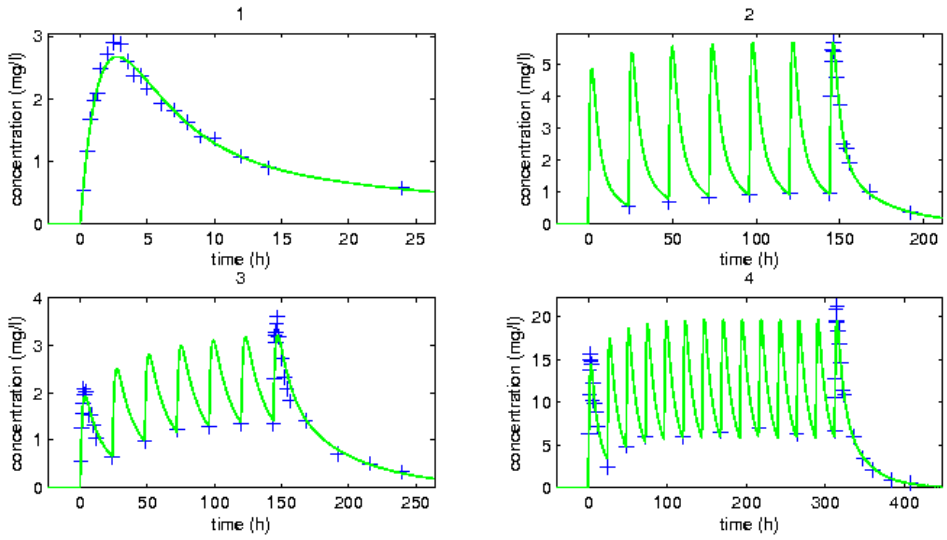

Figure 3.5: Observed concentration data from 4 patients with their concentrations predicted by the model.

Parameters with a mixture distribution	BIC
–	1586.4
k_e	1559.4
(k_e, k_{12})	1534.2
(k_e, k_{12}, V)	1372.2
(k_e, k_{12}, V, k_a)	1376.4
(k_e, k_{12}, V, k_a, k_{21})	1391.4

Table 3.5: The six best models obtained for different number of mixture distributions and the corresponding Bayesian Information Criteria.

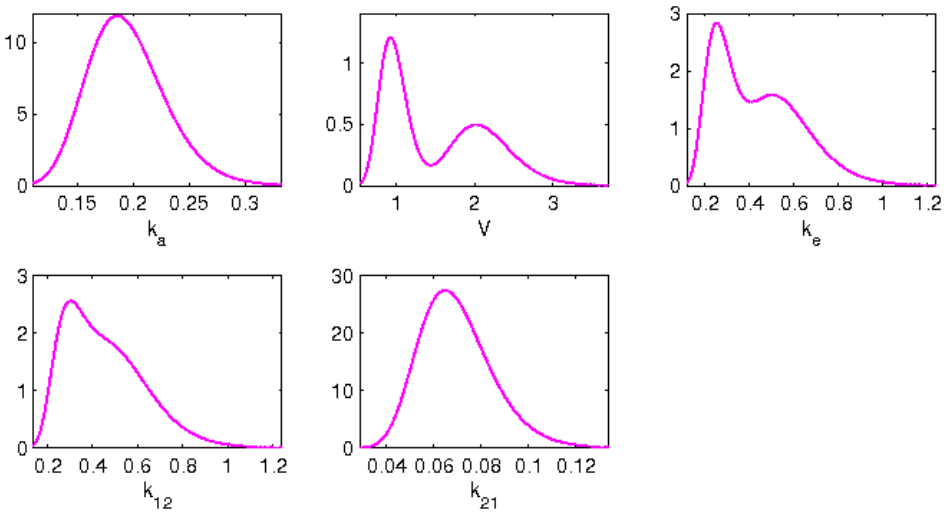

Figure 3.6: Probability distribution functions of the five PK parameters. Distributions of V, k_e and k_{12} are mixtures of log-normal distributions ; distributions of k_a and k_{21} are log-normal distributions.

and in NONMEM are based on a linearization of the likelihood. These methods are known to pose real practical problems in several situations (bias, strong influence of the initial guess, poor convergence,...). Moreover the theoretical properties of the estimates obtained with these methods are unknown in most situations.

Some EM-types methods were developed for mixed models, including mixtures of NLMEM. These methods usually intend to approximate the E step of EM by a Monte-Carlo integration. The MCEM algorithm uses this Monte-Carlo integration for computing the conditional distribution $p(\varphi|y;\theta)$ while the procedure proposed in (De la Cruz-Mesia et al., 2008) aims to integrate the joint distribution $p(\varphi, y;\theta)$ for computing the marginal distribution of y in each cluster. These methods can be drastically time-consuming when the structural model is complex, which is the case for most PKPD applications for instance.

We have proposed an extension of the SAEM algorithm for mixtures of mixed effects models. The model is very general, including mixtures of distributions, mixtures of residual error models and mixtures of structural models. Convergence of MSAEM toward a (local) maximum of the observed likelihood is obtained under very general conditions. The algorithm is fast mainly because the Monte-Carlo integration is replaced by a Stochastic Approximation. Indeed, only one Markov Chain needs to be drawn since the integration is performed over the iterations of the algorithm and not over the chains. Moreover, it is shown to be very few sensitive to the initial value, which is a very valuable property for practical applications. This algorithm for mixtures of NLMEM is now implemented in the MONOLIX software. MONOLIX is free for academic research and for students.

Several demo examples including mixtures of NLMEM are available with the software.

Several extensions of the proposed method would be of particular interest. First, an optimal strategy for model building would be very useful, for selecting both the mixture structure and the number of clusters. Some specific tools for model assessment are also required for demonstrating that the selected model is capable to generate data similar to the observed ones. Lastly, we have only considered here the Maximum Likelihood approach for these models. Estimation in a Bayesian framework can also be done using posterior simulation via Markov chain Monte Carlo (MCMC) methods, see for example (Frühwirth-Schnatter, 2006; De la Cruz-Mesia et al., 2008).

3.7 Appendix: Some important results

3..1 Estimation of several quantities of interest

Maximization step in special cases

For the examples considered in Section 3.4, as in many application fields, it is often reasonable to assume that the mechanism of genetic polymorphism for example applies to only part of the system (for example drug metabolism or drug target in PK/PD field). It is therefore desirable to partition individual parameters φ_i into two components ; one ($\varphi_{i\zeta}$) that follows a mixture and the second ($\varphi_{i\nu}$) defined without mixtures, such that $\varphi_i = (\varphi_{i\zeta}, \varphi_{i\nu})$ with an independence assumption on $\varphi_{i\zeta}$ and $\varphi_{i\nu}$. If we consider the first example in Section 3.3.4, the updates of the parameters at iteration k of MSAEM for this special case are as follow

$$
\begin{aligned}
\mu_{\zeta_m}^{(k)} &= \frac{S_{2,k}^m}{S_{1,k}^m}, \\
\mu_\nu^{(k)} &= \frac{S_{4,k}}{N}, \\
\Sigma_{\zeta_m}^{(k)} &= \frac{1}{S_{1,k}^m}\left(S_{3,k}^m - S_{2,k}^m \mu_{\zeta_m}^{(k)\prime} - \mu_{\zeta_m}^{(k)} S_{2,k}^{m\,\prime} + S_{1,k}^m \mu_{\zeta_m}^{(k)} \mu_{\zeta_m}^{(k)\prime}\right) \\
\Sigma_\nu^{(k)} &= \frac{1}{N}\left(S_{5,k} - S_{4,k}\mu_\nu^{(k)\prime} - \mu_\nu^{(k)} S_{4,k}^{m\,\prime} + \mu_\nu^{(k)} \mu_\nu^{(k)\prime}\right)
\end{aligned}
$$

With the following approximating statistics

$$
\begin{aligned}
S_{1,k}^m &= S_{1,k-1}^m + \delta_k \left(\sum_{i=1}^N \gamma_m \left(\varphi_i^{(k)} \right) - S_{1,k-1}^m \right) \\
S_{2,k}^m &= S_{2,k-1}^m + \delta_k \left(\sum_{i=1}^N \gamma_m \left(\varphi_i^{(k)} \right) \varphi_{i\zeta}^{(k)} - S_{2,k-1}^m \right) \\
S_{3,k}^m &= S_{3,k-1}^m + \delta_k \left(\sum_{i=1}^N \gamma_m \left(\varphi_i^{(k)} \right) \varphi_{i\zeta}^{(k)} \varphi_{i\zeta}^{(k)\prime} - S_{3,k-1}^m \right) \\
S_{4,k} &= S_{4,k-1} + \delta_k \left(\sum_{i=1}^N \varphi_{i\nu}^{(k)} - S_{4,k-1} \right) \\
S_{5,k} &= S_{5,k-1} + \delta_k \left(\sum_{i=1}^N \varphi_{i\nu}^{(k)} \varphi_{i\nu}^{(k)\prime} - S_{5,k-1} \right).
\end{aligned}
$$

The updates of $\pi_m, m = 1, \ldots, M$ and θ_y are the same as in Section 3.3.4.

Estimation of the Fisher information matrix

Once the maximum likelihood estimator have been obtained with the MSAEM algorithm, the next target is the estimation of standards errors on estimated parameters. Assuming some regularity conditions, the standard errors could be obtained by determining the expected Fisher information matrix expressed as

$$
I(\theta) = \mathbb{E}\left(-\nabla_\theta^2 \mathcal{L}(y;\theta) \right), \tag{3.41}
$$

∇_θ^2 being the hessian operator w.r.t the parameter vector θ and $\mathcal{L}(y;\theta)$ being the log-likelihood of observations. In practice, (3.41) is generally computed by the observed information matrix at $\hat{\theta}$ ((Efron and Hinkley, 1978) provided a justification of a such approximation) expressed as $\mathcal{J}\left(\hat{\theta}, y\right)$, such that

$$
I(\theta) = \mathbb{E}_y \left(\mathcal{J}(\theta, y) \right).
$$

Since we are in a typical framework of incomplete data, the computation of $\mathcal{J}(\theta, y)$ could be weakened by the loss of information principle introduces by (Woodbury, 1971). Using this principle, (Louis, 1982) showed that the observed information matrix in a framework of incomplete data could be determined via the EM algorithm with the following relation

$$
\begin{aligned}
-\mathcal{J}(\theta, y) &= \nabla_\theta^2 \mathcal{L}(y;\theta) \\
&= \mathbb{E}\left(\nabla_\theta^2 \mathcal{L}(y, \zeta;\theta) \,|y, \theta \right) + \mathrm{Cov}\left(\nabla_\theta \mathcal{L}(\mathrm{y}, \zeta;\theta) \,|\mathrm{y}, \theta \right),
\end{aligned} \tag{3.42}
$$

ζ representing the unobserved data. (3.42) have an advantage of computing the hessian matrix and the score vector of the completed data which are indeed known

explicitly. In our model, the unobserved vector ζ is given by (φ, z) and (3.42) then becomes

$$-\mathcal{J}(\theta, y) = \mathbb{E}\left(\nabla_\theta^2 \mathcal{L}(y, \varphi, z; \theta) | y, \theta\right) + \mathrm{Cov}\left(\nabla_\theta \mathcal{L}(\mathrm{y}, \varphi, \mathrm{z}; \theta) | \mathrm{y}, \theta\right). \tag{3.43}$$

Remark 3..1. *We have to notice that the introduction of the latent covariates z is really fundamental in order to have a manageable expression of the completed likelihood.*

We will use the same scheme used to carry out parameters estimation here to approximate the observed fisher information matrix. We have the following relation

$$\mathbb{E}\left(\nabla_\theta^2 \mathcal{L}(y, \varphi, z; \theta) | y, \theta\right) = \mathbb{E}\left(\mathbb{E}\left(\nabla_\theta^2 \mathcal{L}(y, \varphi, z; \theta) | y, \varphi; \theta\right) | y, \theta\right).$$

By the same way, we have

$$\mathbb{E}\left(\nabla_\theta \mathcal{L}(y, \varphi, z; \theta) | y, \theta\right) = \mathbb{E}\left(\mathbb{E}\left(\nabla_\theta \mathcal{L}(y, \varphi, z; \theta) | y, \varphi; \theta\right) | y, \theta\right)$$

By setting

$$H(y, \varphi, \theta) = \mathbb{E}\left(\nabla_\theta^2 \mathcal{L}(y, \varphi, z; \theta) | y, \varphi; \theta\right)$$

and

$$\Delta(y, \varphi, \theta) = \mathbb{E}\left(\nabla_\theta \mathcal{L}(y, \varphi, z; \theta) | y, \varphi; \theta\right).$$

The observed information matrix is obtained using the following approximation scheme

$$\begin{aligned} \Delta_{k+1} &= \Delta_k + \delta_k\left(\Delta\left(y, \varphi^{(k+1)}, \theta\right) - \Delta_k\right) \\ H_{k+1} &= \delta_k\left(H\left(y, \varphi^{(k+1)}, \theta\right) + \Delta\left(y, \varphi^{(k+1)}, \theta\right)\Delta^t\left(y, \varphi^{(k+1)}, \theta\right)\right) + H_k(1 - \delta_k). \end{aligned}$$

Finally, the observed information matrix is approximated by

$$\mathcal{J}_{k+1} = -H_{k+1} + \Delta_{k+1}\Delta_{k+1}^t. \tag{3.44}$$

3..2 Convergence result on MSAEM

We shall make the following assumptions on the model:

(C1) The parameter space Θ is an open subset of $\mathbb{R}^p$. The complete data log-likelihood function for a given unit i is given by:

$$\mathcal{L}(y_i, \varphi_i, z_i; \theta) = \sum_{m=1}^{M} \mathbb{1}_{z_i = m}\left(\mathcal{L}_m(y_i, \varphi_i; \theta_m) + \log \pi_m\right), \tag{3.45}$$

with

$$\mathcal{L}_m(y_i, \varphi_i; \theta_m) = \langle T(y_i, \varphi_i), \phi(\theta_m)\rangle - \psi(\theta_m), \ m = 1, \ldots, M \tag{3.46}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, T is a Borel function on $\mathbb{R}^d$ in the second variable taking its values in an open subset $\mathcal{T}$ of $\mathbb{R}^n$. According to (3.45) and (3.46), it appears that the complete data belongs to the exponential family and the log-likelihood of the complete data is given by

$$\mathcal{L}(y, \varphi, z; \theta) = \langle S(y, \varphi, z), \phi(\theta) \rangle - \psi(\theta), \tag{3.47}$$

where

$$S(y, \varphi, z) = (S_m(y, \varphi, z))_{m=1,\dots,M}$$

with

$$S_m(y, \varphi, z) = \left(\sum_{i=1}^{n} \mathbb{1}_{z_i=m}, \sum_{i=1}^{n} \mathbb{1}_{z_i=m} T(y_i, \varphi_i) \right).$$

By setting $\gamma_m(y_i, \varphi_i) = \mathbb{P}(z_i = m | y_i, \varphi_i, \theta)$, with θ considered here as fix or known, we define an intermediate set of statistic $H(y, \varphi) = (H_m(y, \varphi))_{m=1,\dots,M}$ with

$$H_m(y, \varphi) = \left(\sum_i \gamma_m(y_i, \varphi_i), \sum_i \gamma_m(y_i, \varphi_i) T(y_i, \varphi_i) \right).$$

Since $\gamma_m(y_i, \varphi_i)$ is bounded between 0 and 1, H is indeed a Borel function on $\mathbb{R}^d$ in the second variable taking its values in an open subset $\mathcal{H}$ of $\mathbb{R}^{2nM}$.

(C2) Define $CL : \mathcal{H} \times \Theta \to \mathbb{R}$ as:

$$CL(h; \theta) \triangleq \langle h, \phi(\theta) \rangle - \psi(\theta).$$

The functions ϕ and ψ are twice continuously differentiable on Θ.

(C3) The function $\bar{h} : \Theta \to \mathcal{H}$ defined as

$$\bar{h}(\theta) \triangleq \mathbb{E}(H(y, \varphi) | y; \theta)$$

is continuously differentiable on Θ.

(C4) The function $l : \Theta \to \mathbb{R}$ defined as the observed-data log-likelihood

$$l(\theta) \triangleq \log \int_{\mathbb{R}^d} f(y, \varphi, \theta) \, d\varphi$$

is continuously differentiable on Θ and

$$\partial_\theta \int_{\mathbb{R}^d} f(y, \varphi, \theta) \, d\varphi = \int_{\mathbb{R}^d} \partial_\theta f(y, \varphi, \theta) \, d\varphi.$$

(C5) There exists a function $\hat{\theta} : \mathcal{H} \to \Theta$, such that:

$$\forall h \in \mathcal{H}, \forall \theta \in \Theta, CL\left(h, \hat{\theta}(h)\right) \geq CL(h; \theta).$$

Furthermore, $\hat{\theta} \in C^1(\mathcal{H})$.

Let us define

$$Q\left(\theta|\theta'\right)=\int_{\mathbb{R}^d\times\{1,\ldots,M\}}\mathcal{L}\left(y,\varphi,z;\theta\right)p\left(\varphi,z|y;\theta'\right)\mu\left(d\varphi,dz\right). \tag{3.48}$$

According to (3.47), in order to compute (3.48), it is sufficient to compute $\mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta'\right)$ which is same as computing $\mathbb{E}\left(H\left(y,\varphi\right)|y;\theta'\right)$.

At iteration k of the MSAEM algorithm, the S-step consists in generating a realization of the missing data vector $\varphi^{(k)}$ instead of $\left(z^{(k)},\varphi^{(k)}\right)$ under the conditional distribution $p\left(\cdot|y;\theta^{(k-1)}\right)$ by approaching it with a transition probability $p_{\theta^{(k-1)}}\left(\varphi,\varphi^{(k-1)}\right)$ and the integration step in a stochastic averaging procedure

$$h_k=h_{k-1}+\delta_k\left(H\left(y,\varphi^{(k)}\right)-h_{k-1}\right).$$

The maximization step is then given by $\theta^{(k)}=\hat{\theta}\left(h_k\right)$. We consider here several additional conditions on the stochastic approximation procedure, which are similar to those displayed in (Kuhn and Lavielle, 2004) : It is assumed that the random variables $h_0,\varphi^{(1)},\varphi^{(2)},\ldots$ are defined on the same probability space $(\Omega,\mathcal{A},P)$. We denote $\mathcal{F}=\{\mathcal{F}\}_{k\geq 0}$ the increasing family of σ-algebras generated by the random variables $h_0,\varphi^{(1)},\varphi^{(2)},\ldots,\varphi^{(k)}$. We also assume that:

(SAEM1) For all $k\in\mathbb{N}$, $\gamma_k\in[0,1]$, $\sum_{k=1}^{\infty}\gamma_k=\infty$ and $\sum_{k=1}^{\infty}\gamma_k^2<\infty$.

(SAEM2) $l:\Theta\to\mathbb{R}$ and $\hat{\theta}:\mathcal{H}\to\Theta$ are v times differentiable, where v is the integer such that $\mathcal{H}$ is an open subset of $\mathbb{R}^v$.

(SAEM3) $\star$ The chain $\left(\varphi^{(k)}\right)_{k\geq 0}$ takes its value in a compact subset $\mathfrak{E}$ of $\mathbb{R}^d$.

$\star$ For any compact subset U of Θ, there exists a real constant L such that for any (θ,θ') in U^2

$$\sup_{(\alpha,\beta)\in\mathfrak{E}^2}\left|p_\theta\left(\alpha,\beta\right)-p_{\theta'}\left(\alpha,\beta\right)\right|\leq L|\theta-\theta'|.$$

$\star$ The transition probability p_θ generates a uniformly ergodic chain whose invariant probability is the conditional distribution $p\left(\cdot|y;\theta\right)$:

$$\exists K_\theta\in\mathbb{R}^+\ \exists\rho_\theta\in]0,1[\ \forall\varphi\in\mathfrak{E}\ \forall k\in\mathbb{N}\left\|p_\theta^k\left(\varphi,\cdot\right)-p\left(\cdot|y;\theta\right)\right\|_{TV}\leq K_\theta\rho_\theta^k,$$

where $\|\|_{TV}$ denotes the total variation norm. We suppose also that:

$$K\triangleq\sup_\theta K_\theta<\infty\text{ and }\rho\triangleq\sup_\theta\rho_\theta<1.$$

$\star$ The function H is bounded on $\mathfrak{E}$.

We obtain the following convergence result:

Theorem 3..1. *Assume that assumptions (C1 - C5), (SAEM1 - SAEM3) hold. Assume in addition the assumption (D): The sequence $(h_k)_{k\geq 0}$ takes its values in a compact subset of $\mathcal{A}$. Then, we have with probability 1 that ,*

$$\lim_{k\to\infty}d\left(\theta^{(k)},\mathcal{L}\right)=0$$

where $d(v,B)$ denotes the distance of v to the closed subset B and $\mathcal{L}=\{\theta\in\Theta,\partial_\theta l\left(y;\theta\right)=0\}$ is the set of stationary points of the observed likelihood l.

Remark 3..2. *We have considered θ as fixed in $\gamma_m(y_i, \varphi_i)$ to be able to consider $H(y, \varphi)$ as a statistics. In fact, during the E-step of the algorithm, $\gamma_m(y_i, \varphi_i)$ depends on the previous value of the parameter which is known and indeed doesn't affect the current maximization step.*

Proof. First of all, we have to check that under assumptions (C1 - C5), assumptions (M1 - M5) of (Kuhn and Lavielle, 2004) are fulfilled by considering that the model involves the completed data (y, φ, z). According to (C1), the complete data likelihood belongs to the curved exponential family and the complete data log-likelihood is given by

$$\mathcal{L}(y, \varphi, z; \theta) = \langle S(y, \varphi, z), \Phi(\theta) \rangle - \Psi(\theta)$$

where

$$S(y, \varphi, z) = (S_m(y, \varphi, z))_{m=1,\dots,M}$$

with

$$S_m(y, \varphi, z) = \left(\sum_{i=1}^{n} \mathbb{1}_{z_i=m}, \sum_{i=1}^{n} \mathbb{1}_{z_i=m} T(y_i, \varphi_i) \right),$$

$\Phi(\theta) = (\log p_m - \psi(\theta_m); \phi(\theta_m))_{m=1,\dots,M}$ and $\Psi(\theta) \equiv 0$. The assumption (M1) is therefore fulfilled.
By construction and according to assumption (C2), the functions ψ and ϕ are twice continuously differentiable on Θ. The assumption (M2) is now fulfilled.
According to the relation

$$\mathbb{E}(S(y, \varphi, z) | y; \theta) = \mathbb{E}(H(y, \varphi) | y; \theta),$$

assumption (M3) is fulfilled under assumption (C3).
Assumption (C4) is similar as assumption (M4).
Assumption (M5) is fulfilled under assumption (C5) by construction. Assumptions (M1-M5) are now fulfilled. According to (Kuhn and Lavielle, 2004), under additional assumptions (SAEM1 - SAEM3), the convergence result is straightforward. The proof of Theorem 3..1 is now complete.

■

Remark 3..3. *The convergence of this algorithm has been proved in the particular case of non observed individual parameters living in a compact subset. However, as we use mixtures of Gaussian distributions, we cannot assume that their support is compact. In order to provide an algorithm whose convergence can be proved in this more general framework, we can use a more general setting originally introduced in (Andrieu et al., 2005) which involves truncation on random boundaries. An aplication of this technique for SAEM was proposed in (Allassonnière et al., 2010) and can easily be extended for MSAEM.*

3..3 Asymptotic properties of the MLE in Mixture of NLMEM

The Maximum Likelihood, a popular method proposed as a general estimation method by Fisher(1912) has become a standard tool for making inference on unknown parameters in mixed effects models. Nevertheless, obtaining MLE's involves tremendous computational difficulty because of the integrated likelihood, which does not have a close form. The asymptotic properties of MLE have been widely investigated. (Chanda, 1954) generalizes a result by (Cramer, 1946) and proves, under some regularity conditions, that there exists a unique solution of the likelihood equations which is consistent and asymptotically normally distributed. Using the same conditions (Peters and Walker, 1978) show that there is a unique strongly consistent solution of the likelihood equations, which locally maximizes the log-likelihood functions. (Redner, 1981) gives an extension of a previous work by (Wald, 1949)) on the strong consistency of MLE in the non-identifiable case, using some integrability conditions. (Bradley and Gart, 1962) and (Hoadley, 1971) proved weak consistency when the observations are sampled from independent associated populations, i.e., random samples are independent but not identically distributed (it's generally the case in mixed-effects models). A compendium of all these previous results are given in (Redner and Walker, 1984). We can also refer to Kaufmann and Fahrmeir () for standard tools on proof on MLE consistency. However, it is usually assumed that the MLE is consistent in mixed-effects models, without providing a proof. A rigorous proof of the consistency by verifying conditions from existing results can be very difficult due to the integrated likelihood which doesn't have a close form. There are very few works dedicated for that purpose in literature. (Nie and Yang, 2005) has established the strong consistency of MLE in nonlinear mixed-effects models with large cluster size. The same author (Nie, 2006) presents some easily verifiable conditions for the strong consistency of the MLE in generalized linear and nonlinear mixed-effects models. (Nie, 2007) once again, investigates in another paper the rate of convergence depending of the size of the population and the number of observation per subject. This can be summarized as three different type of asymptotic as follow:

- ◇ The number of observation tends to infinity while the number of observation per individual is bounded ($N \to \infty,\ \sup n_i < \infty,\ i = 1, \ldots, N$).
- ◇ The number of observation is finite while the number of observation per individual is allowed to grow ($N < \infty,\ \inf n_i \to \infty,\ i = 1, \ldots, N$).
- ◇ Both the number of observation and the number of observation per individual are allowed to grow ($N \to \infty,\ \inf n_i \to \infty,\ i = 1, \ldots, N$).

In this section, we will focus on the first type of asymptotic and will investigated the asymptotic properties of the MLE in the special case of mixture of NLMEM.

Consistency of the MLE

The study of the consistency of the MLE is important for two reasons. The first one is obvious – in order to apply the MLE, we need to specify conditions under which its asymptotic properties hold. The second reason is more subtle in the sense that for models like mixture in NLMEM, the likelihood is obtained only as an integral over the random effects. Hence, computation of the MLE is generally difficult. As a result, parameters are estimated using alternative estimators such as those provided by the MSAEM algorithm, which has been proved to provide an estimator close to the MLE in the previous section 3..2. First of all, we have to recall the main result of (Nie, 2006) on the consistency of the MLE in generalized and nonlinear mixed-effects models.

Consider the data partitioned into N clusters (or subjects), where the ith cluster consists of n_i observations, $y_i = (y_{i1}, \ldots, y_{in_i})'$. Let $\pi_i(y_i, \varphi_i, \theta_y)$ denote the conditional probability density function of y_i given a random variable $\varphi_i = (\varphi_{i1}, \ldots, \varphi_{ip})$. We assume that $y_1, \ldots, y_N$ are conditionally independent. The unobservable random effects vectors $\varphi_1, \ldots, \varphi_N$ are assumed to be a random sample from a distribution with probability density function $\Phi(\varphi_i, \theta_\varphi)$. Let Θ denote the parameter space for $\theta = (\theta_y, \theta_\varphi)$. The marginal likelihood $l(y;\theta)$ of θ based on the observations $y_1, \ldots, y_N$ is a product of the individual marginal likelihood,

$$l(y;\theta) = \prod_{i=1}^{N} l_i(y_i;\theta),$$

where $l_i(y_i;\theta)$ is the marginal likelihood based on y_i,

$$l_i(y_i;\theta) = \int \pi_i(y_i, \varphi_i, \theta_y)\, \Phi(\varphi_i, \theta_\varphi)\, d\varphi_i. \tag{3.49}$$

Theorem 3..2. *The maximum likelihood estimating equation*

$$\frac{\partial}{\partial\theta} l(y;\theta) = 0$$

has a root $\hat{\theta}_n$ *such that,*

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{\theta}_n = \theta^0\right) = 1,$$

if there is $C > 0$ *and an open subset* ω *of* Θ *which contains the true parameter* θ^0 *such that the following conditions are true for all* $i = 1, \ldots, N$.

N1- For almost all y_i, *the density* $\pi_i(y_i, b_i, \theta_y)$ *and* $\Phi(\varphi_i, \theta_\varphi)$ *admits all first, second and third derivatives on* $\theta \in \omega$.

N2- There are functions $K_\gamma(\varphi_i)$ *for* $|\gamma| \leq 3$ *and* $\Psi(\varphi_i)$, *such that*

$$\begin{aligned}
\sup_{\theta\in\omega} |D^\gamma \Phi(\varphi_i, \theta_\varphi)| &\leq K_\gamma(\varphi_i) \\
\inf_{\theta\in\omega} |\Phi(\varphi_i, \theta_\varphi)| &\geq \Psi(\varphi_i) > 0 \\
\mathbb{E}_{\varphi_i|\theta_\varphi^0}\left[\left(\Phi^{-1}\left(\varphi_i, \theta_\varphi^0\right) K_\gamma(\varphi_i)\right)^{32}\right] &\leq C \\
\mathbb{E}_{\varphi_i|\theta_\varphi^0}\left[\left(\Psi^{-1}(\varphi_i)\, \Phi\left(\varphi_i, \theta_\varphi^0\right)\right)^{32}\right] &\leq C.
\end{aligned}$$

N3- There are functions F_α *for* $|\alpha| \leq 3$ *and* R_i*, such that*

$$\begin{aligned}
\sup_{\theta\in\omega} |D^\alpha \pi_i (y_i, \varphi_i, \theta_y)| &\leq F_\alpha (y_i, \varphi_i) \\
\inf_{\theta\in\omega} |\pi_i (y_i, \varphi_i, \theta_y)| &\geq R_i (y_i, \varphi_i) > 0 \\
\mathbb{E}_{\varphi_i|\theta^0_\varphi} \mathbb{E}_{y_i|\varphi_i,\theta^0_y} \left[\left(\pi_i^{-1} \left(y_i, \varphi_i, \theta^0_y \right) F_\alpha (y_i, \varphi_i) \right)^{32} \right] &\leq C \\
\mathbb{E}_{\varphi_i|\theta^0_\varphi} \mathbb{E}_{y_i|\varphi_i,\tau^0} \left[\left(R_i^{-1} (y_i, \varphi_i) \pi_i \left(y_i, \varphi_i, \theta^0_y \right) \right)^{32} \right] &\leq C.
\end{aligned}$$

N4- $\liminf_{N\to\infty} \lambda_N = \lambda > 0$*, where* λ_N *is the smallest eigenvalue of*

$$F_N \left(\theta^0 \right) = -\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{y_i|\theta^0} \left(\frac{\partial^2 \ln l_i (y_i; \theta)}{\partial\theta\partial\theta'} \right).$$

In order to proof this result, Nie shows that under assumptions (N1-N4), assumptions of the following lemma are fulfilled. Those assumptions constitute conditions in a typical study of asymptotic properties of MLE. We can refer to (Chanda, 1954; Bradley and Gart, 1962; Hoadley, 1971; Lehmann, 1983) for similar conditions.

Lemma 3..1. *The maximum likelihood estimating equation*

$$\frac{\partial}{\partial\theta} l (y; \theta) = 0$$

has a root $\hat{\theta}_n$ *such that,*

$$\mathbb{P} \left(\lim_{n\to\infty} \hat{\theta}_n = \theta^0 \right) = 1,$$

if there exists an open convex subset ω *of* Θ *containing the true parameter point* θ^0 *and* $L > 0$ *such that*

- ◇ *for almost all* y_i*, the density* $l_i (y_i; \theta)$ *admits all first, second and third derivatives.*
- ◇
$$\mathbb{E}_{y_i|\theta^0} \left[\frac{\partial \ln l_i (y_i; \theta)}{\theta_k} |_{\theta=\theta^0} \right]^2 \leq L, \ \forall 1 \leq k \leq r = u + w.$$
- ◇
$$\mathbb{E}_{y_i|\theta^0} \left[\frac{\partial^2 \ln l_i (y_i; \theta)}{\theta_k \theta_l} |_{\theta=\theta^0} \right]^2 \leq L, \ \forall 1 \leq k, l \leq r.$$
- ◇ *There is a sequence of functions* $\{G_1 (y_1), \ldots, G_N (y_N)\}$ *such that,*
$$\left| \frac{\partial^3 \ln l_i (y_i; \theta)}{\theta_k \theta_l \theta_h} \right| \leq G_i (y_i), \ \forall\theta \in \omega, \ \forall 1 \leq k, l, h \leq r$$

and

$$\mathbb{E}_{y_i|\theta^0} \left[G_i^2 (y_i) \right] \leq L, \ \forall 1 \leq i \leq N.$$

◇

$$\liminf_{N\to\infty} \lambda_N = \lambda > 0.$$

Theorem 3..2 is thus obtained by weakening assumptions of Lemma 3..1 which is quite general, but difficult to check in mixed-effects models framework.

In order to prove the consistency result on mixture in nonlinear mixed-effects models, we have to check that assumptions (N1-N4) of Theorem3..2 are fulfilled. Since the theorem involves the factorization of the probability density function of the complete data as the conditional probability density function of observations given the random effects and the probability density function of the random effects, we can not apply the Theorem on our global mixture of NLMEM. There is indeed a need to use particular models of our global mixture of NLMEM.
Let us consider first of all a mixture of distribution model defined as:

$$\begin{cases} y_{ij}|\varphi_i, \sigma^2 \sim \mathcal{N}\left(f_{ij}\left(\varphi_i\right), \sigma^2 f_{ij}^2\left(\varphi_i\right)\right) \\ \varphi_i|\theta_\varphi \sim \sum_{m=1}^{M} \pi_m \mathcal{N}\left(\mu_m, \Sigma_m\right) \\ \varphi_i \in \mathbb{R}^P \text{ and } \sum_{m=1}^{M} p_m = 1 \end{cases} \tag{3.50}$$

According to notations in Theorem 3..2, the marginal likelihood based on y_i is given by (3.49) with

$$\pi\left(y_i, \varphi_i, \sigma^2\right) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{\left(y_{ij} - f_{ij}\left(\varphi_i\right)\right)^2}{\sigma^2 f_{ij}^2\left(\varphi_i\right)} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma^2\right)\right]\right\}, \tag{3.51}$$

and

$$\Phi\left(\varphi_i, \mu, \Sigma, p\right) = \sum_{m=1}^{M} p_m \exp\left\{-\frac{1}{2}\left[\left(\varphi_i - \mu_m\right)' \Sigma_m^{-1}\left(\varphi_i - \mu_m\right) + \ln\left|2\pi\Sigma_m\right|\right]\right\}.$$

By identification, the parameters are defined as follow:

$$\begin{cases} \theta_y = \sigma^2 \\ \theta_\varphi = \left(p, \mu_1, \ldots, \mu_p, \theta_1, \ldots, \theta_{\frac{p(p+1)}{2}}\right) \\ \mu_k = \left(\mu_{k,m}\right)_{m=1,\ldots,M}, \ \forall k = 1, \ldots, p \text{ and } \theta_j = \left(\theta_{j,m}\right)_{m=1,\ldots,M} \ \forall j = 1, \ldots, \frac{p(p+1)}{2} \\ p = \left(p_1, \ldots, p_M\right) \end{cases}$$

Let $w = \prod_{i=1}^{\frac{p(p+3)+4}{2}} w_i$ containing the true value of the parameters, in which

$$\begin{aligned} w_1 &= \left]\sigma_1^2, \sigma_2^2\right[\\ w_2 &= \left]\kappa_1, \kappa_2\right[^M, \ \kappa_1 > 0, \ \kappa_2 < 1 \\ w_i &= \left]-t, t\right[^M, \ i = 3, \ldots, p+2, \ 0 < t < \infty \\ w_i &= \prod_{m=1}^{M} \left]\theta_{i-p-2,1}^m, \theta_{i-p-2,2}^m\right[\ i = p+3, \ldots, \frac{p\left(p+3\right)+4}{2} \end{aligned}$$

Assumption (N1) is fulfilled since $\pi\left(y_i, \varphi_i, \sigma^2\right)$ and $\Phi\left(\varphi_i, \mu, \Sigma, p\right)$ are smooth functions.
Let $u_{ij} = \frac{y_{ij} - f_{ij}(\varphi_i)}{f_{ij}(\varphi_i)}$. (3.51) becomes

$$\pi\left(y_i, \varphi_i, \sigma^2\right) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{u_{ij}^2}{\sigma^2} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma^2\right)\right]\right\}.$$

We have the following inequality:

$$\pi\left(y_i, \varphi_i, \sigma^2\right) \leq \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{u_{ij}^2}{\sigma_2^2} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma_1^2\right)\right]\right\}.$$

The first derivative of $\pi\left(y_i, \varphi_i, \sigma^2\right)$ w.r.t σ^2 is given by

$$\frac{\partial\pi}{\partial\sigma^2} = \pi\left(y_i, \varphi_i, \sigma^2\right)\left(\frac{1}{2}\sum_{j=1}^{n_i}\frac{u_{ij}^2}{\sigma^4} - \frac{n_i\pi}{\sigma^2}\right).$$

By taking the absolute value, we have the following majoration

$$\left|\frac{\partial\pi}{\partial\sigma^2}\right| \leq \pi\left(y_i, \varphi_i, \sigma^2\right)\left(\frac{1}{2}\sum_{j=1}^{n_i}\frac{u_{ij}^2}{\sigma^4} + \frac{n_i\pi}{\sigma^2}\right).$$

Similarly, we obtain the following majoration for the absolute value of the second derivative of $\pi\left(y_i, \varphi_i, \sigma^2\right)$ w.r.t σ^2

$$\left|\frac{\partial^2\pi}{\partial\left(\sigma^2\right)^2}\right| \leq \pi\left(y_i, \varphi_i, \sigma^2\right)\left(\frac{1}{4\sigma^6}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^2 + \frac{n_i\pi}{2\sigma^6}\sum_{j=1}^{n_i}u_{ij}^2 + \frac{n_i\pi}{2\sigma^4}\sum_{j=1}^{n_i}u_{ij}^2 + \left(\frac{n_i\pi}{\sigma^2}\right)^2\right).$$

The third derivative w.r.t σ^2 is given by

$$\begin{aligned}\frac{\partial^3\pi}{\partial\left(\sigma^2\right)^3} &= \pi\left(y_i, \varphi_i, \sigma^2\right)\left(\frac{1}{8\sigma^{10}}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^3 - \frac{n_i\pi}{4\sigma^{10}}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^2 - \frac{n_i\pi}{4\sigma^8}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^2\right.\\
&+\frac{\left(n_i\pi\right)^2}{\sigma^8}\sum_{j=1}^{n_i}u_{ij}^2 - \frac{n_i\pi}{4\sigma^8}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^2 + \frac{\left(n_i\pi\right)^2}{2\sigma^6}\sum_{j=1}^{n_i}u_{ij}^2 - \frac{\left(n_i\pi\right)^3}{\sigma^6}\\
&\left.-\frac{3}{4\sigma^6}\left(\sum_{j=1}^{n_i}u_{ij}^2\right)^2 + \frac{3\pi n_i}{2\sigma^8}\sum_{j=1}^{n_i}u_{ij}^2 + \frac{n_i\pi}{\sigma^6}\sum_{j=1}^{n_i}u_{ij}^2 - \frac{2\left(n_i\pi\right)^2}{\sigma^6}\right).\end{aligned}$$

Finally, we obtained the following result:

$$\forall\alpha/\ \ |\alpha| \leq 3,\ \left|D^\alpha\pi\left(y_i, \varphi_i, \sigma^2\right)\right| \leq \pi\left(y_i, \varphi_i, \sigma^2\right)Q\left(u_i, \sigma^2\right)$$

with

$$u_i = (u_{i,j})_{j=1,\ldots,n_i} \text{ and } Q\left(u_i, \sigma^2\right) = \sum_{l=1}^{3}\sum_{k=1}^{5} c_{lk} \frac{\left(\sum_{j=1}^{n_i} u_{ij}^2\right)^l}{\sigma^{2k}},$$

where c_{lk} are positive constants. We thus have

$$\sup_{\sigma^2\in\omega_1} \left|D^{\alpha}\pi\left(y_i, \varphi_i, \sigma^2\right)\right| \leq \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{u_{ij}^2}{\sigma_2^2} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma_1^2\right)\right]\right\} Q\left(u_i, \sigma_1^2\right) \tag{3.52}$$

Let

$$\boxed{F_{\alpha_i}\left(y_i, \varphi_i\right) = \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{n_i}\left[\tfrac{u_{ij}^2}{\sigma_2^2} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma_1^2\right)\right]\right\} Q\left(u_i, \sigma_1^2\right)}$$

Then,

$$\left(\pi^{-1}\left(y_i, \varphi_i, \sigma_0^2\right) F_{\alpha_i}\left(y_i, \varphi_i\right)\right)^{32} = \exp\left[-\frac{1}{2}\left(\frac{32}{\sigma_2^2} - \frac{32}{\sigma_0^2}\right)\sum_{j=1}^{n_i} u_{ij}^2 - 16 n_i \ln\frac{\sigma_1^2}{\sigma_0^2}\right] Q^{32}\left(u_i, \sigma_1^2\right)$$

Thus, the 32nd moment is given by

$$\begin{aligned}\mathbb{E}_{y_i|\varphi_i,\sigma_0^2}\left(\pi^{-1}\left(y_i, \varphi_i, \sigma_0^2\right) F_{\alpha_i}\left(y_i, \varphi_i\right)\right)^{32} &= \mathbb{E}_{u_i|\varphi_i,\sigma_0^2}\left\{\exp\left[-\frac{1}{2}\left(\frac{32}{\sigma_2^2} - \frac{32}{\sigma_0^2}\right)\sum_{j=1}^{n_i} u_{ij}^2 - 16 n_i \ln\frac{\sigma_1^2}{\sigma_0^2}\right]\right. \\ &\qquad \left. Q^{32}\left(u_i, \sigma_1^2\right)\right\} \end{aligned} \tag{3.53}$$

We have to notice that $u_{ij}|\varphi_i, \sigma_0^2 \sim \mathcal{N}\left(0, \sigma_0^2\right)$; thus,Given the random variables φ_i, u_{ij} are i.i.d and an immediate consequence is that (3.53) doesn't depends on φ_i. Hence,

$$\mathbb{E}_{\varphi_i|\theta^0}\mathbb{E}_{y_i|\varphi_i,\sigma_0^2}\left(\pi^{-1}\left(y_i, \varphi_i, \sigma_0^2\right) F_{\alpha_i}\left(y_i, \varphi_i\right)\right)^{32} = \mathbb{E}_{y_i|\varphi_i,\sigma_0^2}\left(\pi^{-1}\left(y_i, \varphi_i, \sigma_0^2\right) F_{\alpha_i}\left(y_i, \varphi_i\right)\right)^{32}.$$

The u_{ij} being centered, the evaluation of the previous operation involves an integration of the form $\int_0^{\infty} e^{-ax^2}\, dx$ with $a = \frac{1}{2}\left(\frac{1}{\sigma_0^2} - 32\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_2^2}\right)\right)$. This type of integral is finite if and only if $a > 0$. We thus obtained an constraint on σ_2 for the definition of the open subset ω_1. The condition is as follow:

$$\frac{1}{\sigma_0^2} - 32\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_2^2}\right) > 0. \tag{3.54}$$

Over these condition, (3.53) is finite.
In the other hand, we have that

$$\inf_{\sigma^2\in\omega_1} \pi\left(y_i, \varphi_i, \sigma^2\right) \geq \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{u_{ij}^2}{\sigma_1^2} + \ln f_{ij}^2\left(\varphi_i\right) + \ln\left(2\pi\sigma_2^2\right)\right]\right\}.$$

Let

$$\boxed{R_i\left(y_i,\varphi_i\right)=\exp\left\{-\tfrac{1}{2}\sum_{j=1}^{n_i}\left[\tfrac{u_{ij}^2}{\sigma_1^2}+\ln f_{ij}^2\left(\varphi_i\right)+\ln\left(2\pi\sigma_2^2\right)\right]\right\}}$$

By using a similar procedure as before, we obtain that

$$\mathbb{E}_{y_i|\varphi_i,\sigma_0^2}\left(\pi\left(y_i,\varphi_i,\sigma_0^2\right)R_i^{-1}\left(y_i,\varphi_i\right)\right)^{32}$$

doesn't depends on φ_i and is finite if

$$\frac{1}{\sigma_0^2}-32\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)>0. \tag{3.55}$$

We therefore verified Condition (N3) in Theorem 3..2.

Assuming that φ_i is a mixture of M gaussian distributions, we obtained the following result derived from the partial derivative of order 1, 2 and 3 of $\Phi\left(\varphi_i,\theta_\varphi\right)$ w.r.t θ_φ:

$$\forall\alpha/\ |\alpha|\leq 3,\ D^\alpha\Phi\left(\varphi_i,\theta_\varphi\right)<\sum_{m=1}^{M}\mathcal{K}_1^m\left(\varphi_i\right)\exp\left\{-\frac{1}{2}\left(\varphi_i-\mu_m\right)'\Sigma_m^{-1}\left(\varphi_i-\mu_m\right)\right\},\ \theta\in\omega,$$

where $\mathcal{K}_1^m\left(\varphi_i\right)$ is a polynomial function of φ_i with order less than 6. Coefficients of this polynomial are finite values depending on the boundaries of ω ; more precisely t and $\theta_{k,j}^m,\ k=1,\ldots,p(p+1)/2;\ j=1,2$. With the aim of providing a majoration of $\exp\left\{-\frac{1}{2}\left(\varphi_i-\mu_m\right)'\Sigma_m^{-1}\left(\varphi_i-\mu_m\right)\right\}$ which doesn't depends on θ_φ, we suppose that there exists a set of real constant $\gamma_1=\left(\gamma_{1,m}\right)_{m=1,\ldots,M}$ such that

$$\inf_{\theta\in\omega}\Sigma_m^{-1}-\gamma_{1,m}\Sigma_{m0}^{-1}\text{ is positive definite }\forall m=1,\ldots,M \tag{3.56}$$

Σ_{m0} is the true value of the parameter Σ_m. Let $\varphi_{im}=\Sigma_{m0}^{-1}\varphi_i$, $\mu_m^*=\Sigma_{m0}^{-1}\mu_m$. In addition with (3.58), we have the following inequality

$$\exp\left\{-\frac{1}{2}\left(\varphi_i-\mu_m\right)'\Sigma_m^{-1}\left(\varphi_i-\mu_m\right)\right\}\leq\exp\left\{-\frac{1}{2}\gamma_{1m}\varphi_{im}'\varphi_{im}+\gamma_{1m}\sum_{j=1}^{p}|\varphi_{im}^j||\mu_m^{*j}|\right\},$$

where φ_{im}^j is the jth component of the vector φ_{im} and μ_m^{*j} the jth component of the vector μ_m^*. Define

$$\mathcal{K}_2^m\left(\varphi_i\right)=\exp\left\{-\frac{1}{2}\gamma_{1m}\varphi_{im}'\varphi_{im}+\gamma_{1m}\sum_{j=1}^{p}|\varphi_{im}^j|\sup_{\theta\in\omega}|\mu_m^{*j}|\right\}.$$

Thus, we have

$$\sup_{\theta\in\omega}D^\alpha\Phi\left(\varphi_i,\theta_\varphi\right)\leq\sum_{m=1}^{M}\mathcal{K}_1^m\left(\varphi_i\right)\mathcal{K}_2^m\left(\varphi_i\right).$$

Let

$$\boxed{\Phi_\alpha(\varphi_i) = \sum_{m=1}^{M} \mathcal{K}_1^m(\varphi_i) \exp\left\{-\tfrac{1}{2}\gamma_{1m}\varphi'_{im}\varphi_{im} + \gamma_{1m}\textstyle\sum_{j=1}^{p}|\varphi_{im}^j| \sup_{\theta\in\omega}|\mu_m^{*j}|\right\}.}$$

After using the relation

$$\left(\sum_{m=1}^{M} a_m\right)^n \leq M^{n-1}\sum_{m=1}^{M} a_m^n,\ a_m \geq 0 \forall m = 1,\ldots,M;\ n \in \mathbb{N}^*,$$

obtained after using the Hölder inequality, we obtain the following result:

$$\left(\Phi^{-1}\left(\varphi_i, \theta_\varphi^0\right)\Phi_\alpha(\varphi_i)\right)^{32} \leq \frac{\text{cste}}{\kappa_1^{32}}\sum_{m=1}^{M} H_m(\varphi_i)\exp\left(-16\left(\gamma_{1m}-1\right)\varphi_i'\Sigma_{l0}^{-1}\varphi_i\right),$$

where $l \in \{1,\ldots,M\}$ and $H_m(\varphi_i)$ is a product of a polynomial function of φ_i and an exponential of a polynomial function of $|\varphi_{im}^j|$ with order 1. Using similar arguments as before, the 32nd moment of $\Phi^{-1}\left(\varphi_i, \theta_\varphi^0\right)\Phi_\alpha(\varphi_i)$ is finite if the following condition holds:

$$1 - 32\left(1-\gamma_{1m}\right) > 0, \quad \text{for all } m = 1,\ldots,M. \tag{3.57}$$

By using an analogue procedure, we suppose that there exists a set of real constant $\gamma_2 = (\gamma_{2,m})_{m=1,\ldots,M}$ such that the following condition holds:

$$-\sup_{\theta\in\omega}\Sigma_m^{-1} + \gamma_{2,m}\Sigma_{m0}^{-1} \text{ is positive definite} \quad \text{for all } m = 1,\ldots,M. \tag{3.58}$$

Then, Let

$$\boxed{\Psi(\varphi_i) = \sum_{m=1}^{M}\sqrt{\frac{\gamma_{1,m}^p}{|2\pi\Sigma_{m0}|}}\exp\left(-\tfrac{1}{2}\gamma_{2,m}\left[\varphi'_{im}\varphi_{im} + \left(\gamma_{1,m}\sum_{j=1}^{p}\sup_{\theta\in\omega}|\mu_m^{*j}|\right)^2\right] + \gamma_{2,m}\sum_{j=1}^{p}|\varphi_{im}^j|\inf_{\theta\in\omega}|\mu_m^{*j}|\right)}$$

By the same manner, we can show that the 32nd moment of $\Psi^{-1}(\varphi_i)\,\Phi\left(\varphi_i, \theta_\varphi^0\right)$ is finite if the following condition holds:

$$1 - 32\left(\gamma_{2,m}-1\right) > 0, \quad \text{for all } m = 1,\ldots,M. \tag{3.59}$$

We therefore verified Condition (N2) in Theorem 3..2.

In order to show Condition (N4), it suffices to show that each $F_i(\theta^0)$ is positive definite, where,

$$F_i\left(\theta^0\right) = \mathbb{E}_{y_i|\theta^0}\left(\frac{\partial^2 \ln l_i(y_i;\theta)}{\partial\theta\partial\theta'}\right).$$

$F_i(\theta^0)$ is positive definite if and only if the components of the score vector obtained at θ^0 are linearly independent, that is, there do not exist constants $(\lambda_1, \ldots, \lambda_r)$, such that $\sum_{k=1}^{r} \lambda_k \frac{\partial \ln l_i(y_i;\theta)}{\partial \theta_k}|_{\theta=\theta^0} = 0$ with $y_i|\theta^0$-probability one ; r being an integer such that Θ is an open subset of $\mathbb{R}^r$. We have the following relation:

$$\forall k \in \{1, \ldots, r\}, \frac{\partial \ln l_i(y_i;\theta)}{\partial \theta_k}|_{\theta=\theta^0} = \mathbb{E}_{\varphi_i|y_i,\theta}\left(\frac{\partial\left(\ln \pi(y_i;\theta_y) + \ln \Phi(\varphi_i;\theta_\varphi)\right)}{\partial \theta_k}\right)|_{\theta=\theta^0}.$$

Thus,

$$\frac{\partial \ln l_i(y_i;\theta)}{\partial \sigma^2}|_{\theta=\theta^0} = \mathbb{E}_{\varphi_i|y_i,\theta}\left(\frac{\partial\left(\ln \pi(y_i;\sigma^2)\right)}{\partial \sigma^2}\right)|_{\theta=\theta^0}$$

and

$$\frac{\partial \ln l_i(y_i;\theta)}{\partial \theta_\varphi}|_{\theta=\theta^0} = \mathbb{E}_{\varphi_i|y_i,\theta}\left(\frac{\partial\left(\ln \Phi(\varphi_i;\theta_\varphi)\right)}{\partial \theta_\varphi}\right)|_{\theta=\theta^0}.$$

In addition, we have that

$$\forall k \in \{1, \ldots, r\} \text{ s.t. } \theta_k \neq \sigma^2, \frac{\partial \ln l_i(y_i;\theta)}{\partial \theta_k} \neq b\frac{\partial \ln l_i(y_i;\theta)}{\partial \sigma^2} \text{ for any constant } b.$$

Using the regularity conditions on $\Phi(\varphi_i, \theta_\varphi)$ it is clear that, there do not exist constants $\{a_2, \ldots, a_r\}$, such that, $\sum_{k=2}^{r} a_k \frac{\partial \ln l_i(y_i;\theta)}{\partial \theta_k}|_{\theta=\theta^0} = 0$, with $\theta_1 = \sigma^2$.

We have therefore verified condition (N4) of Theorem3..2. As a consequence, the MLE of θ is consistent. Thus, the $\sqrt{N}$-consistency is straightforward here and is obtained by considering a Taylor expansion of the score function in a neighborhood of the true parameter.

Chapter 4

Between-subject and within-subject model mixtures for classifying HIV treatment response

Contents

Abstract

We present a method for using longitudinal data to classify individuals into clinically-relevant population subgroups. This is achieved by treating "subgroup" as a categorical covariate whose value is unknown for each individual, and predicting its value using mixtures of models that represent "typical" longitudinal data from each subgroup. Under a nonlinear mixed effects model framework, two types of model mixtures are presented, both of which have their advantages. Following illustrative simulations, longitudinal viral load data for HIV-positive patients is used to predict whether they are responding – completely, partially or not at all – to a new drug treatment.

Keywords: *Mixture models, HIV, SAEM, Classification.*

4.1 Introduction

For a variety of reasons – some known, some not – different patients respond differently to the same drug treatment. For certain patients, a drug does what it was prescribed to do: kill bacteria, reduce blood pressure, decrease viral load, etc., but for others, the drug may be toxic or ineffective. When we collect response data on patients undergoing a treatment, it is useful to try to find patients for which the treatment is ineffective, and thus suggest modifications. We are particularly interested here in longitudinal response data in a population; the methods we present are generally applicable to this type of data.

The real-world example that motivates the approach is longitudinal HIV viral load data. For HIV-positive patients on a given drug regime, the evolution of the viral load in the blood can be measured over time. For some patients, the drug regime is ineffective and the viral load does not consistently drop; we call these *non-responders*, and it is of interest to detect them and provide alternative – hopefully more effective – treatments. Other patients react favourably to the treatment and the viral load drops to undetectable levels and stays there for a long period of time; these are called *responders*. Yet another group – *rebounders* – show an initial drop in viral load followed by an increase back towards the initial high viral load. They too will eventually require an alternative treatment.

Our goal is to use longitudinal data to infer the efficacy of the treatment, i.e., infer whether each patient is a non-responder, responder, rebounder or, as we will explain, some mixture of the above. In order to do this, we first model longitudinal HIV viral load data using recent additions to the nonlinear mixed-effects model (NLMEM) framework. Then, we extract relevant posterior probabilities or individual parameters to infer patient status. We now briefly introduce the NLMEM framework and model mixtures. To avoid potential confusion, note that *mixed-effects models* and *model mixtures* are not the same thing.

NLMEM – a special case of mixed-effects models – are statistical models which use both fixed and random effects in their construction (see (Sheiner and Beal, 1985; Lindstrom and Bates, 1990; Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997) for more details). The model structure is hierarchical. At a first level, each individual has their own parametric regression model, known as the structural model, each identically defined up to a set of unknown individual parameters. At a second level, each set of individual parameters is assumed to be randomly drawn from some unknown population distribution. These models are particularly useful in population studies (e.g., population pharmacology – see (Wakefield et al., 1998)) where data is available for many individuals. Two types of variability are involved: *intra-* and *inter*-subject. We attempt to explain the latter using known *fixed-effects* (covariates) such as weight, blood type, etc. The non-explained part of the inter-subject variability is then modeled using *random effects*.

The introduction of a categorical covariate (e.g., sex, blood type, etc.) into

such a model supposes that the population can be divided into subpopulations with respect to that covariate. However, there may be a categorical covariate which interests us but whose value is unknown for all individuals, such as the covariate "patient status" which interests us here. In this case, part of the goal becomes to infer the value of this covariate for each individual as part of the modeling process. One way to do this is to introduce *model mixtures*. There exist several types of model mixture which are useful in the context of mixed effects models; we will focus on two here:

- ◇ *Between-Subject Model Mixtures* (BSMM) assume that each individual's longitudinal data follows one of M "base" models, but we do not necessarily know *a priori* which one. Individual i thus has a label $z_i = m \in \{1, \ldots, M\}$ referring to the model that is supposed to have generated it. If the z_i are known, they can be treated as categorical covariates. We will show how to deal with the more challenging case of when they are unknown. Furthermore, for this z_i unknown case, we will show how to extract *a posteriori* estimates of the probability that each individual was generated by each of the base models; this will be used to predict which type of patient we have: non-responder, responder or rebounder. We note that BSMM were introduced as an example of a more general framework in (Mbogning and Lavielle,) but were not developed further there.

- ◇ *Within-Subject Model Mixtures* (WSMM) make the hypothesis that the model mixture occurs *within each individual*. In the HIV example, this means that we consider that each patient is partially a non-responder, partially a responder and partially a rebounder. This is perhaps more biologically plausible than BSMMs in the sense that each individual's response may be due to their own particular combination of virus strains, cell populations, etc. Within the NLMEM framework, this means including individual "model proportion" parameters into the model and having to estimate them along with the other parameters of the NLMEM. It turns out that this does not require any mathematical extensions to a typical NLMEM. But as will be seen in the HIV example, we can use the estimated proportions to help categorize patients, especially those who do not naturally fall into one of the three "typical" categories.

BSMM and WSMM are new approaches in the context of NLMEM. We refer the reader to (Redner and Walker, 1984; McLachland and Basford, 1988; Bryant, 1991; Roeder and Wasserman, 1997; McLachland and Peel, 2000) for general details on mixture models in a standard context. We note also that there are fully Bayesian approaches to similar types of problems, in particular Bayesian nonparametric ones, see (Hjort et al., 2010) for more details.

The paper is structured as follows. We introduce BSMMs in the NLMEM framework and calculate their log-likelihood, before briefly presenting WSMMs; they require no new mathematical framework. We then describe how to per-

form maximum likelihood estimation (MLE) for BSMMs using the Stochastic Approximation Expectation Maximization (SAEM) algorithm (Delyon et al., 1999). SAEM is implemented in the MONOLIX software and can be widely applied to various data types and real-life scenarios (Samson et al., 2007; Snoeck et al., 2010; Chan et al., 2011; Dubois et al., 2011; Lavielle et al., 2011). Next, we present an example for a simple BSMM case, and then a simulated example for mixtures of two models in the both the BSMM and WSMM cases. The simulations illustrate the quality of the parameter estimation of both methods and also their classification performance, i.e., how well they "predict" which model was used to generate each individual's data (in the BSMM case) and which model represented the biggest proportion (in the WSMM case). This is followed by a comprehensive modeling of HIV treatment response longitudinal data from a cohort of 578 patients using both BSMM and WSMM and a comparison of the quality and practical usefulness of each method. A discussion follows.

4.2 Models and methods

4.2.1 Between-subject model mixtures

Between-subject model mixtures (BSMMs) are a special case of NLMEMs that assume that the structural model is a mixture of M different structural models and can be written, in the case of a continuous response, as:

$$y_{ij} = \sum_{m=1}^{M} \mathbb{1}_{\{z_i=m\}} \Big(f_m\left(x_{ij};\psi_i\right) + g_m\left(x_{ij};\psi_i,\theta_y\right)\varepsilon_{ij} \Big), \tag{4.1}$$

where

- $y_{ij} \in \mathbb{R}$ denotes the jth observation of the ith individual, $1 \leq i \leq N$ and $1 \leq j \leq n_i$.
- N is the number of individuals and n_i the number of observations of the ith individual.
- $x_{ij}$is a vector of regression variables (for longitudinal data, x_{ij} will generally be time t_{ij}).
- ψ_i is the d-vector of individual parameters of individual i. We assume that all the ψ_i are drawn from the same population distribution and are defined as Gaussian transformations:

  $$\psi_i = h(\mu, c_i, \eta_i), \tag{4.2}$$

  where h is a function which describes the covariate model, μ a vector of fixed-effects, c_i a vector of known covariates, $\eta_i \sim_{i.i.d} \mathcal{N}\left(0,\Sigma\right)$ a vector of random effects and Σ the inter-individual variance-covariance matrix.
- $z_i \in \{1,\ldots,M\}$ represents the (un)known group to which belongs the individual i. The proportion of individuals in group m is given by $\pi_m = \mathbb{P}\left(z_i = m\right)$ with $\sum_{m=1}^{M} \pi_m = 1$.

- $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ are the residual errors, and are independent of individual parameters ψ_i.
- f_m for $m = 1, \ldots, M$ are functions defining structural models in each group.
- g_m for $m = 1, \ldots, M$ are functions defining the (possibly heteroscedastic) residual error model. We will consider here error models of the form $g_m = a + b\, f_m$.

Furthermore, let $\theta = (\mu, \Sigma, \theta_y, \pi_1, \ldots, \pi_M)$ represent the complete set of population parameters.

BSMMs are particularly relevant in the domain of population pharmacology if we are aiming to distinguish between different classes of response to the same treatment. For example, if we can effectively model each class of longitudinal response data (specific mathematical function with parameters, etc.), *a posteriori* estimation of the label z_i for each individual leads to being able to assign each to a given "typical" response. We shall see later in the real HIV data that classes such as "non-responder", "responder", and "rebounder" could be used to categorize individuals' responses in this application.

4.2.2 Log-likelihood of between-subject model mixtures

In this section, we briefly recall the initial exposition and notation in Chapter 3 for the log-likelihood of mixture models in general; it will be useful in the following. The complete data is noted (y, ψ, z) with y the observed data and (ψ, z) the unobserved data. For subject i, the log-likelihood of the complete data is

$$\mathcal{L}(y_i, \psi_i, z_i; \theta) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \left(\mathcal{L}_m(y_i, \psi_i; \theta) + \log \mathbb{P}(z_i = m) \right), \tag{4.3}$$

where $\mathcal{L}_m(y_i, \psi_i; \theta)$, the log-likelihood of pairs of variables (y_i, ψ_i) in group $G_m := \{i, 1 \leq i \leq N \text{ such that } z_i = m\}$, is given by $\mathcal{L}_m(y_i, \psi_i; \theta) = \mathcal{L}_{Y,m}(y_i|\psi_i; \xi) + \mathcal{L}_\psi(\psi_i; \mu, \Sigma)$. The right-hand side terms are simple to calculate. $\mathcal{L}_m$ is assumed to belong to the exponential family, i.e., there exists a function ψ of θ and a minimal sufficient statistic $T(y_i, \psi_i)$ such that $\mathcal{L}_m(y_i, \psi_i; \theta) = \langle T(y_i, \psi_i), \theta \rangle - \psi(\theta)$. In what follows, we will note $\mathbb{P}(z_i = m)$ as π_m or π_{im} for "proportion" for respectively BSMM or WSMM. We have that

$$\mathcal{L}(y_i, \psi_i, z_i; \theta) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \left(\langle T(y_i, \psi_i), \theta \rangle + \log \pi_m - \psi(\theta) \right).$$

The likelihood of the complete data also belongs to the exponential family as it can be written $\mathcal{L}(y, \psi, z; \theta) = \langle S(y, \psi, z), \theta \rangle - \psi(\theta)$, where the mth row of S is given by

$$\left(\sum_{i=1}^{n} \mathbb{1}_{z_i=m} \;,\; \sum_{i=1}^{n} \mathbb{1}_{z_i=m} T(y_i, \psi_i) \right).$$

We will show later that this representation of the log-likelihood is helpful for implementing stochastic EM-like algorithms for the BSMM case.

4.2.3 Within-subject model mixtures

It may be too simplistic to assume that each individual is represented by only one well-defined model from the mixture. For instance, in a pharmacological setting there may be subpopulations of cells, viruses (etc.) *within each patient* that react differently to a drug treatment. In this case, it makes sense to consider that the mixture of models happens *within* each individual. Such within-subject model mixtures (WSMMs) therefore require additional vectors of individual parameters $\pi_i = (\pi_{i1}, \dots \pi_{iM})$ representing proportions of the M models within each individual i. These vectors are supposed independent of ψ, and naturally sum to 1 for each individual. Using the same notation as Section 4.2.1, observations are modeled by:

$$y_{ij} = \sum_{m=1}^{M} \pi_{im} \left(f_m \left(x_{ij}; \psi_i \right) + g_m \left(x_{ij}; \psi_i, \theta_y \right) \varepsilon_{ij} \right). \tag{4.4}$$

Since there are no latent categorical covariates, WSMMs actually fall under the framework of classical NLMEMs. Thus, no further specific methodology needs to be developed, and we refer to (Delyon et al., 1999; Kuhn and Lavielle, 2005) for the standard treatment.

4.3 Maximum likelihood estimation algorithms for between-subject model mixtures

A method such as the Stochastic Approximation EM (SAEM) algorithm (Delyon et al., 1999) needs to be used to replace the E-step of the EM algorithm in the NLMEM framework. The stochastic step of SAEM is performed in practice using an MCMC procedure. SAEM has been successfully applied to a wide number of data types and real-life situations including bioequivalence crossover trials (Dubois et al., 2011), estimation of population pharmacokinetic-pharmacodynamic viral dynamics parameters (Chan et al., 2011), longitudinal ordered categorical data (Savic et al., 2011), population models for count data (Savic and Lavielle, 2009) and group comparison tests in longitudinal data analysis (Samson et al., 2007).

We now develop this technique in the case of BSMMs. Iteration k consists of a number of MCMC iterations with $p(\psi, z|y; \theta^{(k)})$ as the stationary distribution. More precisely, the Gibbs algorithm is combined with the Metropolis-Hastings algorithm, with various proposal kernels. Here, the N subjects are assumed to be independent and the same procedure is used for each of the N subjects. For subject i, draw $z_i^{(k)} \in \{1, \dots, M\}$ from the multinomial distribution

$$\mathcal{M}_M \left(\frac{\pi_m^{(k-1)} p_m \left(y_i, \psi_i^{(k-1)}; \theta^{(k-1)} \right)}{\sum_{r=1}^{M} \pi_r^{(k-1)} p_r \left(y_i, \psi_i^{(k-1)}; \theta^{(k-1)} \right)} \right)_{m=1,\dots,M}.$$

A first possible kernel uses the marginal distribution $p(\psi_i; c_i, \theta^{(k)})$ for generating a candidate ψ_i^c; more precisely, $\eta_i \sim \mathcal{N}\left(0, \Sigma^{(k)}\right)$ and ψ_i^c is as in (4.2). The probability of acceptance, i.e., the probability to move from ψ_i to ψ_i^c, becomes

$$\alpha\left(\psi_i, \psi_i^c\right) = \min\left(1, \frac{p\left(y|\psi_i^c, z_i^{(k)}; \theta^{(k)}\right)}{p\left(y|\psi_i, z_i^{(k)}; \theta^{(k)}\right)}\right).$$

A random walk can also be used as a possible kernel: $\psi_i^c \sim \mathcal{N}\left(\psi^{(k-1)}, \Omega\right)$. The diagonal matrix Ω can be adaptively adjusted to get a chosen acceptance rate. Setting different elements of the diagonal of Ω to 0 during iterations can be done to use different directions. The probability of acceptance is

$$\alpha\left(\psi_i, \psi_i^c\right) = \min\left(1, \frac{p\left(y, \psi_i^c, z_i^{(k)}; \theta^{(k)}\right)}{p\left(y, \psi_i, z_i^{(k)}; \theta^{(k)}\right)}\right).$$

For details on how to choose parameters such as the number of iterations of the MCMC procedure during the simulation step, the step-size sequence (δ_k), etc., we refer the reader to (Mbogning and Lavielle,). We remark that this version of SAEM for BSMMs is now implemented in the MONOLIX software.

BSMM Example. In order to illustrate MLE for the BSMM model, let us consider a simple example. Suppose first that we have a constant residual model $g_m\left(x_{ij}; \psi_i, \theta_y\right) = a$. Then, the conditional log-likelihood of the observations in the group G_m is:

$$\mathcal{L}_{Y,m}\left(y_i|\psi_i; \theta_y\right) = -\frac{1}{2a^2}\sum_{j=1}^{n_i}\left(y_{ij} - f_m\left(x_{ij}, \psi_i\right)\right)^2 - n_i \log\left(a\right) - \frac{n_i}{2}\log\left(2\pi\right).$$

Furthermore, assuming a Gaussian distribution without covariates (i.e., $\psi_i = \mu + \eta_i$) for individual parameters, the likelihood of the individual parameters is

$$\mathcal{L}_{\psi}\left(\psi_i; \mu, \Sigma\right) = -\frac{1}{2}\left(\psi_i - \mu\right)' \Sigma^{-1}\left(\psi_i - \mu\right) - \frac{d}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(|\Sigma|\right).$$

In the first step of iteration k, we must approximate the following minimal sufficient statistics:

$$\begin{aligned}
s_{k,1,m} &= s_{k-1,1,m} + \delta_k\left(\sum_{i=1}^{N} \mathbb{1}_{z_i^{(k)}=m} - s_{k-1,1,m}\right) \\
s_{k,2} &= s_{k-1,2} + \delta_k\left(\sum_{i=1}^{N} \psi_i^{(k)} - s_{k-1,2}\right) \\
s_{k,3} &= s_{k-1,3} + \delta_k\left(\sum_{i=1}^{N} \psi_i^{(k)}\psi_i^{(k)'} - s_{k-1,3}\right) \\
s_{k,4} &= s_{k-1,4} + \delta_k\left(\sum_{i,j,m} \mathbb{1}_{z_i^{(k)}=m}\left(y_{ij} - f_m\left(x_{ij}, \psi_i^{(k)}\right)\right)^2 - s_{k-1,4}\right).
\end{aligned}$$

Then, in the M-step, we update parameters according to:

$$\begin{aligned}
\pi_m^{(k)} &= \frac{s_{k,1,m}}{N} \\
\mu^{(k)} &= \frac{s_{k,2}}{N} \\
\Sigma^{(k)} &= \frac{s_{k,3}}{N} - \left(\frac{s_{k,2}}{N}\right)\left(\frac{s_{k,2}}{N}\right)' \\
a^{(k)} &= \sqrt{\frac{s_{k,4}}{\sum_{i=1}^{N} n_i}}.
\end{aligned}$$

4.3.1 Estimation of individual parameters

The overall goal of this paper is not to provide complicated viral dynamics models (ordinary differential equations, etc.) but rather to show how the model mixture framework can be used to predict classes of individuals. For our real-life application, this means being able to decide whether HIV patients are non-responders, responders, rebounders or some mixture of the above. For BSMM, the latent categorical covariate z contains the unknown "class" labels that need to be estimated. For a given set of population parameters θ, we can use each individual's conditional distribution $p(z_i, \psi_i|y_i, \theta)$ to estimate the latent variable z_i and the vector of individual parameters ψ_i. A first estimate is the Maximum a Posteriori (MAP) which is obtained by maximizing this joint conditional distribution with respect to (z_i, ψ_i):

$$\left(\hat{z}_i, \hat{\psi}_i\right) = \arg\max_{(z_i,\psi_i)} p\left(z_i, \psi_i|y_i, \theta\right).$$

This maximization is not straightforward and we refer the reader to Chapter 3 for a complete methodology. Another way to estimate the latent covariate z_i is to maximize the marginal conditional distribution:

$$\hat{z}_i = \arg\max_m \mathbb{P}\left(z_i = m|y_i; \theta\right). \tag{4.5}$$

The value of (4.5) can be estimated using a stochastic approximation during the SAEM iterations.

As for WSMM, since there are no latent categorical covariates z, estimation of individual proportions π_i as well as individual parameters ψ_i is straightforwardly obtained by maximizing the joint conditional distribution $p\left(\pi_i, \psi_i|y_i, \theta\right)$ with respect to (π_i, ψ_i):

$$\left(\hat{\pi}_i, \hat{\psi}_i\right) = \arg\max_{(\pi_i,\psi_i)} p\left(\pi_i, \psi_i|y_i, \theta\right).$$

Once we have estimates of individual parameters, individual predictions for BSMM are obtained using $\hat{y}_{ij} = f_{\hat{z}_i}\left(x_{ij}, \hat{\psi}_i\right)$. Similarly, for WSMM, the individual pre-

dictions are calculated as

$$\hat{y}_{ij} = \sum_{m=1}^{M} \hat{\pi}_{im} f_m \left(x_{ij}, \hat{\psi}_i \right).$$

4.4 Simulated Data Example

4.4.1 Modeling with between-subject model mixtures

We performed a simulation study to evaluate the performance of the proposed BSMM algorithm for estimating parameters and classifying individuals using a mixture of two simple models defined as follows:

$$\begin{array}{ll} \text{if } z_i = 1, & y_{ij} = f_1(\psi_i, t_{ij}) + a\varepsilon_{ij} \\ \text{if } z_i = 2, & y_{ij} = f_2(\psi_i, t_{ij}) + a\varepsilon_{ij}, \end{array}$$

where

$$f_1(\psi_i, t_{ij}) = A_i\,, \qquad f_2(\psi_i, t_{ij}) = A_i e^{-L_i t_{ij}}\,,$$

and the vectors of individual parameters $\psi_i = (\log(A_i), \log(L_i))$ are such that A_i and L_i are log-normally distributed, $\log(A_i) \sim \mathcal{N}(\log(A), \sigma_A^2)$ and $\log(L_i) \sim \mathcal{N}(\log(L), \sigma_L^2)$. Note that this model can also be seen as a parameter mixture for L_i: $L_i = 0$ in group 1 and L_i log-normally distributed in group 2. For the experiments, we set $A = 10$, $L = 0.2$, $\sigma_A^2 = 0.5$ and $\sigma_L^2 = 0.5$. Furthermore, we fixed $\mathbb{P}(z_i = 1) = \pi_1 = 1/3$ and $a = 1$. t_{ij} is the jth measurement time for subject i, and we used the same set of times $t = 0, 1, \ldots, 8$ for all N subjects. $K = 1000$ datasets were simulated and the parameters were estimated using the proposed algorithm. Figure 4.1 shows examples of longitudinal data for 10 subjects with (a) $z_i = 1$ and (b) $z_i = 2$.

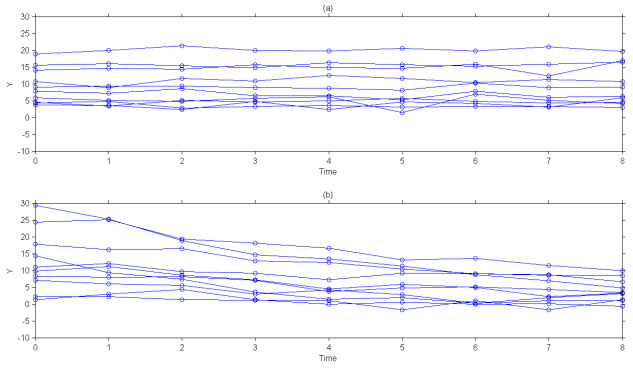

Figure 4.1: Spaghetti plots of data for ten individuals belonging to group 1 (a) and 10 others belonging to group 2 (b).

Let θ^* be the true value of any parameter and $\hat{\theta}_k$ the estimated value obtained with the kth simulated dataset. The relative estimation error (in %) REE_k was used as a quality criteria:

$$\text{REE}_k = \frac{\hat{\theta}_k - \theta^\star}{\theta^\star} \times 100.$$

Figure 5.1 shows the distribution of the REE_k for each parameter when (a-d) $N = 100$ and (e-h) $N = 1000$ with various experimental situations, i.e., when ψ and/or z are supposed known/unknown. It suggests that most parameters are

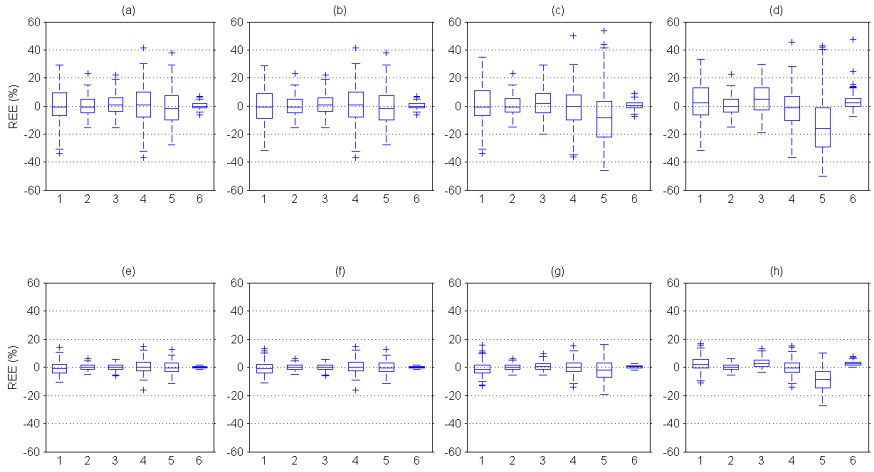

Figure 4.2: Empirical distribution of the relative estimation error (REE_k (%)) in the BSMM scenario with different sample sizes: (a-d) $N = 100$, (e-h) $N = 1000$ and various cases: (a)&(e) z and ψ are known, (b)&(f) z is unknown and ψ is known, (c)&(g) z is known and ψ is unknown, (d)&(h) z and ψ are unknown. The estimated parameters are 1: π_1; 2: A; 3: L; 4: σ_A^2; 5: σ_L^2; 6: a.

estimated with little or no bias, except perhaps the variance of the individual parameter L_i. Cases (a-b) and (e-f) are quite similar, suggesting that the EM algorithm is efficient with respect to bias (mixture parameters are estimated in cases (b-f) with the EM algorithm). Furthermore, cases (c-d) and (g-h) are quite similar and we see that there is little degradation compared with (a-b) and (e-f) respectively. Thus, the SAEM algorithm for mixtures appears to be efficient with respect to bias. As expected, we see that parameters are better estimated with $N = 1000$, but even with $N = 100$ the results are acceptable.

Quantitative results are presented in Table 4.1, which gives mean as well as the standard errors for each of the estimated parameters when $N = 100$ and $N = 1000$.

The parameter estimates, overall, match the population values and the standard errors seems reasonable and are quite low, especially with large sample size. In addition, Figure 4.3, representing the relative difference between the estimated

		N=100		N=1000	
θ	θ^*	Mean of estimates	SE of $\hat{\theta}$	Mean of estimates	SE of $\hat{\theta}$
π_1	0.33	0.334	0.049	0.331	0.015
A	10	10.03	0.713	10.01	0.226
L	0.20	0.202	0.02	0.20	0.006
σ_A^2	0.50	0.497	0.075	0.50	0.023
σ_L^2	0.50	0.49	0.099	0.50	0.033
a	1	1.004	0.026	1.002	0.008

Table 4.1: Mean of parameter estimates and standard errors for the BSMM scenario with $N = 100$ or $N = 1000$.

standard error and the empirical standard error, shows that the standard errors are well estimated and match pretty well the empirical ones.

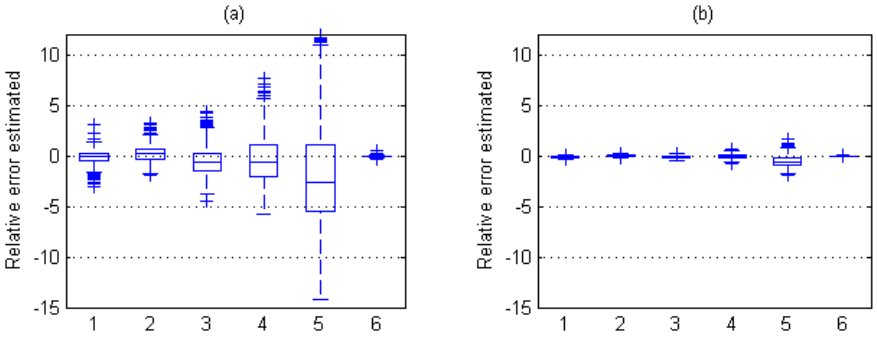

Figure 4.3: Relative difference (in %) between estimated and empirical standard errors for the BSMM scenario with (a): $N = 100$, (b): $N = 1000$. The estimated parameters are 1: π_1; 2: A; 3: L; 4: σ_A^2; 5: σ_L^2; 6: a.

Figure 4.4 provides a graphical illustration of the probability of correct classification in both groups for $N = 100$ and $N = 1000$ subjects. Note that the number of individuals in each group is considered as fixed here ($N_1 = N\pi_1$) during the Monte Carlo simulation. For each of the $K = 1000$ runs, the probabilities of correct classification for the N subjects were computed and ranked in increasing order. Then, the empirical median sequence of these 1000 sequences was computed in each group. This median is displayed in solid line in Figure 4.4. This graphics is more informative than the distribution of the number of subject misclassified over the simulations. Indeed, we see for instance that, with $N = 100$, less than 3 (resp. 4) subjects among 33 (resp. 67) of group 1 (resp. 2) have a probability smaller than 0.8 to be correctly classified in half of the cases.

As expected, the probability of correct classification is greater when (ψ_i) is known, but it is interesting to note that the difference is relatively small.

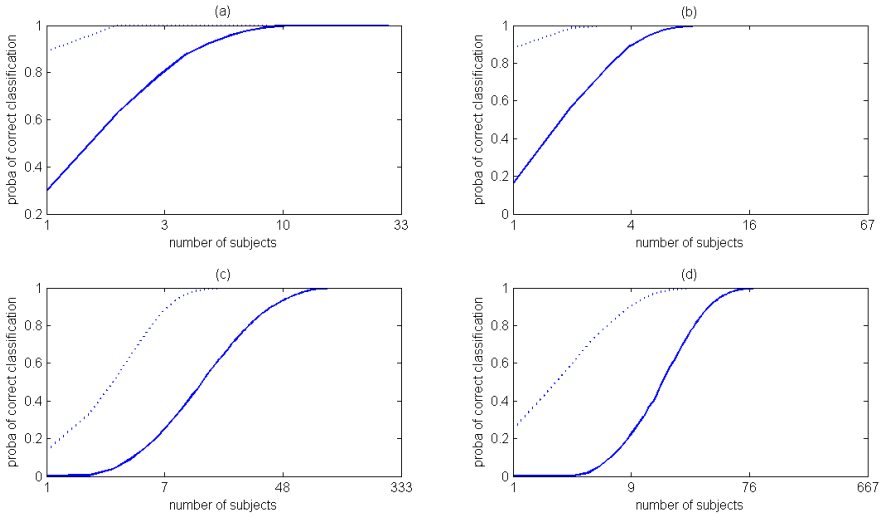

Figure 4.4: Medians of the probabilities of correct classification ranked in increasing order in both groups with (a& b): $N = 100$ ($N_1 = 33,\ N_2 = 67$) and (c& d): $N = 1000$ ($N_1 = 333,\ N_2 = 667$) subjects ; (a& c): group 1 and (b& d): group 2. Solid line: the individual parameters (ψ_i) are unknown; dotted line: the individual parameters (ψ_i) are known.

4.4.2 Modeling with within-subject mixture models

We used the same model but now assumed that individual proportions π_{i1} were model parameters:

$$f\left(\psi_i, t_{ij}\right) = \pi_{i1} f_1\left(\psi_i, t_{ij}\right) + \left(1 - \pi_{i1}\right) f_2\left(\psi_i, t_{ij}\right),$$

where the structural models f_1 and f_2 were the same as for those for the BSMM, and individual proportions modeled as:

$$\pi_{i1} = \frac{1}{1 + s\, e^{\eta_i}}, \qquad \eta_i \sim \mathcal{N}\left(0, \sigma_s^2\right),$$

with $s = 2$ and $\sigma_s^2 = 0.2$. Population parameters were fixed at $A = 10, L = 0.2, \sigma_A^2 = 0.5$ and $\sigma_L^2 = 0.5$. We used the same measurement times for the N subjects as in the BSMM scenario.

Figure 4.5 shows the distribution of the REE_k for each parameter (a)& (b) and the relative difference between the estimated standard error and the empirical standard error (c) & (d) when (a)& (c) $N = 100$ and (b)& (d) $N = 1000$ subjects. It suggests according to (a) and (b), that parameters are estimated with very little bias, and with high precision when the sample size increases to $N = 1000$. In other hand, according to (c) and (d), the standard errors are well estimated and match pretty well the empirical ones when the sample size is large.

Quantitative results are presented in Table 4.2, which gives mean as well as the standard errors for each of the estimated parameters when $N = 100$ and $N = 1000$.

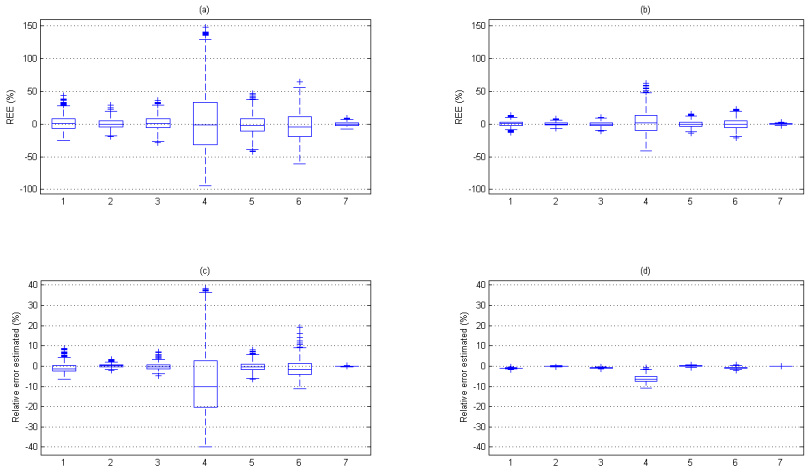

Figure 4.5: Empirical distribution of the relative estimation error (REE_k (%)) in the WSMM scenario (a)& (b) and the relative difference between the estimated standard error and the empirical standard error (c) & (d) whit two sample sizes: (a)& (c) $N = 100$ and (b)& (d) $N = 1000$. The estimated parameters are 1: s; 2: A; 3: L; 4: σ_s^2; 5: σ_A^2; 6: σ_L^2; 7: a.

The parameter estimates, overall, match the population values and the standard

		N=100		N=1000	
θ	θ^*	Mean of estimates	SE of $\hat{\theta}$	Mean of estimates	SE of $\hat{\theta}$
s	2	2.02	0.2183	2.00	0.053
A	10	10.04	0.714	10.00	0.226
L	0.20	0.202	0.021	0.199	0.006
σ_s^2	0.20	0.212	0.11	0.205	0.022
σ_A^2	0.50	0.495	0.073	0.50	0.023
σ_L^2	0.50	0.481	0.112	0.50	0.032
a	1	1.00	0.027	1.00	0.008

Table 4.2: Mean of parameter estimates and standard errors for the WSMM scenario with $N = 100$ or $N = 1000$.

errors seems reasonable and are quite low with large sample size.

4.5 Application to real data

4.5.1 Description of the data

The randomized, controlled and partially blinded POWER project conducted by TIBOTEC comprises 3 studies performed in highly treatment-experienced HIV-

infected patients using Darunavir/Ritonavir or an investigator-selected control protease inhibitor, combined with an optimized background regimen of nucleotide reverse transcriptase inhibitors with or without the fusion inhibitor enfuvirtide. The output data is the viral load evolution for 578 patients. Figure 4.6 gives examples of patients with one of three "characteristic" viral load progressions:

- ◇ *Non-responders* (1) show no decline in viral load.
- ◇ *Responders* (2) exhibit a sustained viral load decline.
- ◇ *Rebounders* (3 and 4) exhibit an initial drop in viral load, then a rebound to higher viral load levels.

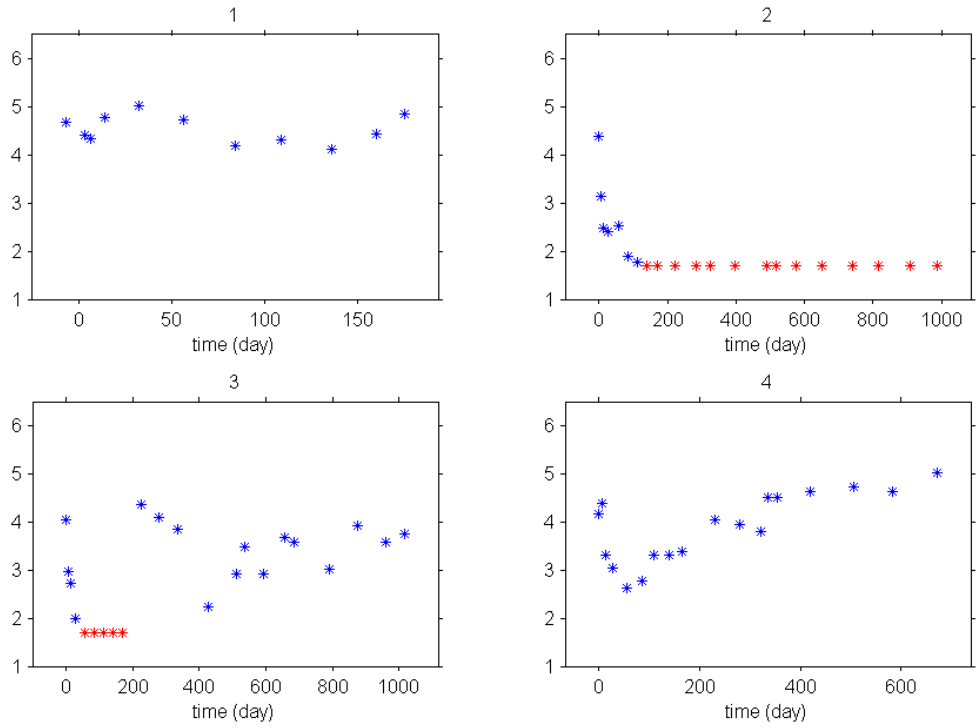

Figure 4.6: Viral load progression for 4 HIV-infected patients. (1) non-responder; (2) responder; (3) and (4) are rebounders. Red points indicate below level of quantification data.

Remark. There is a detection limit at 50 HIV RNA copies/ml, corresponding to a log-viral load of 1.7; i.e., data are left censored. These points are shown in red in Figure 4.6. Censoring is taken into account in the following analysis as described in (Samson et al., 2006).

4.5.2 Class prediction using between-subject model mixtures

Within a few months of HIV infection, patients typically enter a steady state of chronic infection and show their concentration of HIV-1 in blood plasma stabilized. When the anti-retroviral treatment starts, the viral load of patients who respond shows an initial rapid exponential decay, usually followed by a slower second phase of exponential decay, see (Perelson et al., 1997). It is shown in (Ding and Wu, 2001) that the biphasic decay in viral load can be approximated by a bi-exponential model $A_1 e^{-\lambda_1 t} + A_2 e^{-\lambda_2 t}$.

After the decrease in viral load levels, some subjects show a rebound which can be due to several factors (non adherence to the therapy, emergence of drug-resistant virus strains, ...). We propose to extend the bi-exponential model for these patients by adding a third phase described by a logistic growth process $A_3/(1 + e^{-\lambda_3(t-\tau)})$ where τ is the inflection point of this growth process.

We then propose to describe the log-transformed viral load with a BSMM with three simple models, corresponding to each of three characteristic viral load progressions:

- ◇ Non-responder-like data can be described using a simple horizontal line. The structural model is given by:

$$z_i = 1, \quad f_1(\psi_i, t_{ij}) = A_{1i} + A_{2i}\,.$$

- ◇ As described above, the drop in viral load in responder-like data can be described using a bi-exponential mixed-effects model:

$$z_i = 2, \quad f_2(\psi_i, t_{ij}) = A_{1i} e^{-\lambda_{1i} t_{ij}} + A_{2i} e^{-\lambda_{2i} t_{ij}},$$

  where λ_{1i} and λ_{2i} describe the rate of exponential decay and A_{1i} and A_{2i} are intercept parameters for individual i. As in (Davidian and Giltinan, 1995), these parameters are considered to be strictly positive.
- ◇ Rebounder-like data show a rebound after a biphasic decrease in viral load levels:

$$z_i = 3, \quad f_3(\psi_i, t_{ij}) = A_{1i} e^{-\lambda_{1i} t_{ij}} + A_{2i} e^{-\lambda_{2i} t_{ij}} + \frac{A_{3i}}{1 + e^{-\lambda_{3i}(t_{ij} - \tau_i)}}\,.$$

The log-transform viral load is then modeled by:

$$\log(y_{ij}) = \sum_{m=1}^{3} \mathbb{1}_{z_i = m} \log\left(f_m(\psi_i, t_{ij})\right) + \varepsilon_{ij}, \tag{4.6}$$

where y_{ij} is the viral load for subject i at time t_{ij} and $\psi_i = (A_{1i}, A_{2i}, A_{3i}, \lambda_{1i}, \lambda_{2i}, \lambda_{3i}, \tau_i)$ the vector of individual parameters.

These parameters are positive and distributed according to log-normal distributions. Thus,

$$\begin{aligned} \log A_{li} &= \log A_l + \eta_{li}, & \eta_{li} &\sim \mathcal{N}\left(0, \sigma^2_{A_l}\right), & l &= 1,2,3 \\ \log \lambda_{mi} &= \log \lambda_m + \eta_{(m+3)\,i}, & \eta_{(m+3)\,i} &\sim \mathcal{N}\left(0, \sigma^2_{\lambda_m}\right), & m &= 1,2,3 \\ \log \tau_i &= \log \tau + \eta_{7i}, & \eta_{7i} &\sim \mathcal{N}\left(0, \sigma^2_{\tau}\right). \end{aligned}$$

By setting $\pi_1 = \mathbb{P}(z_i = 1), \pi_2 = \mathbb{P}(z_i = 2)$ and $\pi_3 = \mathbb{P}(z_i = 3)$ with $\sum_{m=1}^{3} \pi_m = 1$, the complete set of population parameters to be estimated is given by $\theta = (A_1, A_2, A_3, \lambda_1, \lambda_2, \lambda_3, \tau, \pi_1, \pi_2, \pi_3)$.

Parameter estimation was performed using the SAEM algorithm for BSMM implemented in MONOLIX . This algorithm combined with a previous version (Samson et al., 2006)(which is an extension of the SAEM algorithm (Kuhn and Lavielle, 2005) to left censored data) takes properly into account the censored viral load data below the limit of quantification. Figure 4.7 shows the individual fits for the 4 patients, the vector of estimated posterior probabilities $\hat{p}_i$ where $p_{im} = \mathbb{P}\left(z_i = m | y_i; \hat{\theta}\right)$, i.e., the probabilities for each subject to belong to each of the three classes, and the class z_i to which they are assigned (1 = non-responder, 2 = responder, 3 = rebounder) corresponding to the maximum of the estimated posterior probabilities. We see that in all four cases, there is little ambiguity in the results, i.e., the correct class has a posterior probability very close to 1.

4.5.3 Class prediction using within-subject mixture models

Not all observed viral load progressions fall so easily into one of the three classes, as for example the patients shown in Figure 4.8. In these cases, it does not seem quite so reasonable to model the data under the BSMM assumption that each patient must belong uniquely to one class; instead, it is perhaps more natural to suppose that each patient is partially responding, partially non-responding and partially rebounding with respect to the given drug treatment. The goal becomes to find the relative strength of each process in each patient, and a WSMM is an ideal tool to do this. The proportions π_i are now individual parameters in the model and the problem is transformed into a standard NLMEM. Since these proportions are assumed to be positive and summing to 1 for each patient, we assumed a logit-normal distribution on the $\pi_{im}, m = 1,2,3$ as follows:

$$\pi_{i1} = \frac{\gamma_{1i}}{1 + \gamma_{1i} + \gamma_{2i}} \qquad \pi_{i2} = \frac{\gamma_{2i}}{1 + \gamma_{1i} + \gamma_{2i}} \qquad \pi_{i3} = \frac{1}{1 + \gamma_{1i} + \gamma_{2i}},$$

with

$$\log \gamma_{ki} = \log \gamma_k + \eta_{ki}, \qquad \eta_{ki} \sim \mathcal{N}\left(0, \sigma^2_{\gamma_k}\right) \quad \text{for } k = 1,2.$$

We performed parameter estimation for WSMM using the SAEM algorithm in MONOLIX , taking properly into account the censored viral load data below

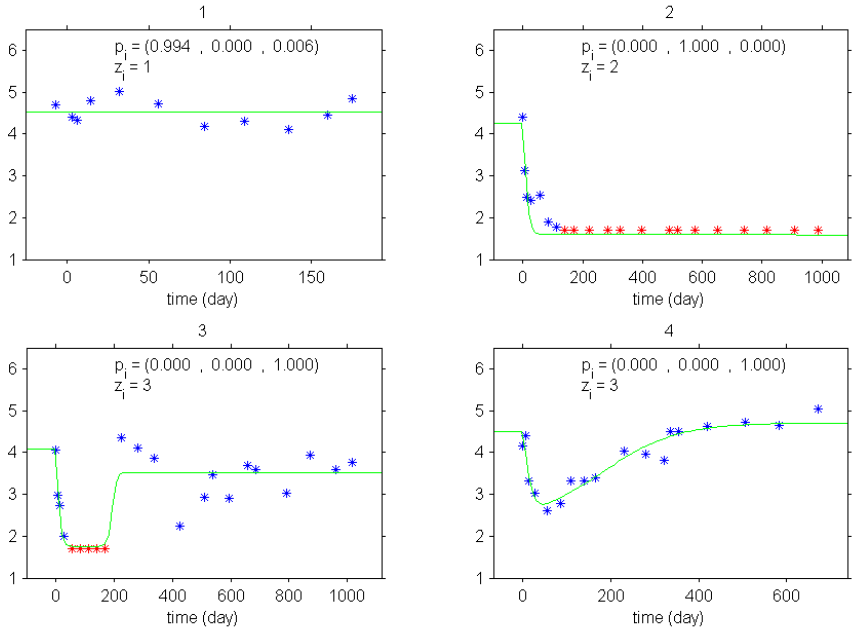

Figure 4.7: Viral load time series and individual fits for the four patients in Figure 4.6. z_i is the predicted class of the patient (1 = non-responder, 2 = responder, 3 = rebounder) and p_i the posterior probability that the patient is in classes 1 to 3; the index of its maximum component is used to predict z_i.

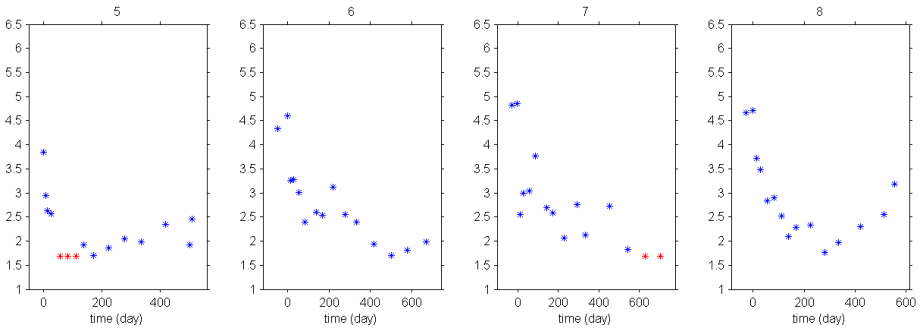

Figure 4.8: Viral load data for 4 patients with ambiguous progressions. Red points indicate below level of quantification data.

the limit of quantification as before. Results for the four patients are presented in Figure 4.9; the first row gives the results which would be obtained using BSMM for these four patients, the second row the WSMM results.

We see that for all four patients, BSMM predicts that the patient was a responder. The visually poor individual fits in all four cases give rise to suspicion in the validity of the result. In particular, in plot 8 – top row – what appears to be a "late" rebounder is hardly "seen" by the algorithm; the posterior probability that the patient is a rebounder is only 0.085. Clearly, forcing each patient to belong to one class in the interior of the algorithm is a disadvantage of the method for

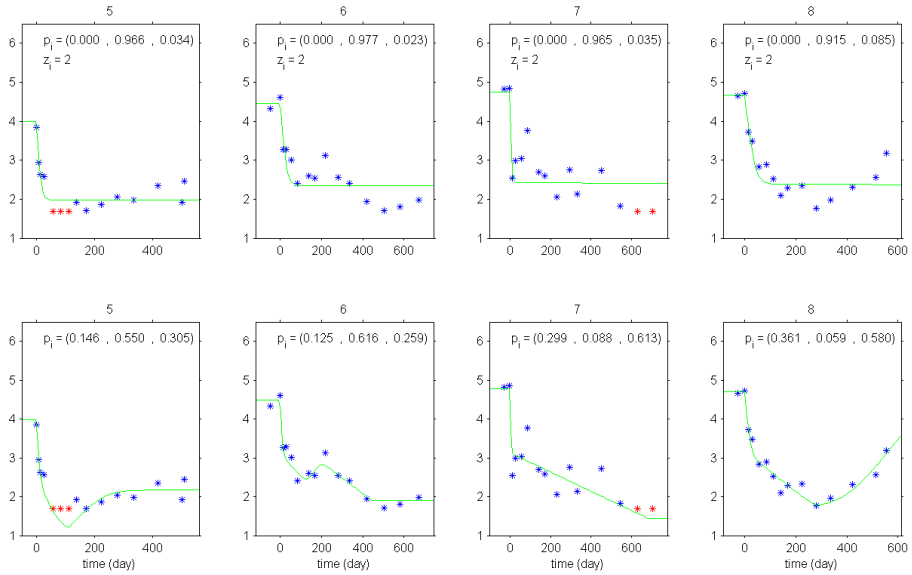

Figure 4.9: Viral load time series and individual fits for patients 5–8 from Figure 4.8 when using BSMM (row 1) and WSMM (row 2). For BSMM, z_i is the predicted class of the patient (1 = non-responder, 2 = responder, 3 = rebounder) and p_i the posterior probability that the patient is in classes 1 to 3. For WSMM, p_i are the mixture parameters of the classes estimated as part of the model.

real-world data.

In contrast to this, the first thing that is immediately obvious with the WSMM modeling (Figure 4.9 – bottom row) is that the individual fits are significantly better than the BSMM ones. This can be confirmed using the BIC criteria which clearly selects the WSMM model: BIC(WSMM)=14 668, whereas BIC(BSMM)=15029. The estimated parameters p_i are generally consistent with what we "see" in the graphs, and when they are not, it helps us to look closer. For instance, in plots 5 and 6 – bottom row – we see evidence of responding and a hint of rebound; this is confirmed in the estimated parameters p_i. In plot 7 – bottom row – the patient would seem to be a very slowly-but-surely responder. However, modeling happens population-wise, and this patient's reponder curve is far from the "normal" (steep drop followed by flatlining) responder; consequently the responder proportion in the mixture is small (0.088). Another reason may be that the 4th, 5th and 6th data points indicate a steep rebound from the 3rd data point (even though subsequent points drop) possibly influencing the weight of the rebounder model in the mixture (0.613). This is probably also accentuated by the fact that the rebounder model "includes" the responder model's biexponential decay terms. This remark also goes some way to explaining plot 8 – bottom row; the rebounder mixture coefficient (0.580) dominates and the responder coefficient (0.059) is unintuitively small. Again this is probably because the steep drop that we associate with the responder model is already included as the first two terms

in the rebounder model; hence the small responder coefficient.

4.6 Discussion

We have presented a classification methodology to interpret longitudinal data in a population context using model mixtures in the NLMEM framework. Two classes of model mixtures were introduced: between-subject and within-subject (BSMM and WSMM), and it was shown how to perform maximum likelihood estimation in the BSMM case. These algorithms are now available in the software MONOLIX . In simulations with mixtures of two models, we saw generally good parameter estimation performance and prediction performance, i.e., prediction of which model had been used to generate each individual's longitudinal data. In real longitudinal HIV viral load data, we found that BSMMs were very efficient both in modeling and predicting the type of patient (non-responder, responder, rebounder) whenever the patient's viral load evolution was a "model case" of one of the three classes. However, when the viral load evolution was not so "typical", the fact that BSMMs force the allocation of once class to each patient lead to poor individual model fits and what's more, dubious class prediction.

On the other hand, WSMMs allow for more flexibility in modeling and are consistent with the biologically plausible hypothesis that each individual may be only partially responding/non-responding/rebounding due to their own unique virus strains, cell populations, etc. Allowing a mixture of biologically relevant models *internally* to each patient clearly improved the individual model fits and permitted a more nuanced reply to the classification question of whether the patient was responding adequately to the treatment or not. Based on these results, we would have more confidence in suggesting a modification to HIV drug regimes based on WSMM modeling than BSMM.

Chapter 5

Joint modeling of longitudinal and repeated time-to-event data with maximum likelihood estimation via the SAEM algorithm.

Contents

Abstract

We propose a nonlinear mixed-effects framework to jointly model longitudinal and repeated time-to-event data. A parametric nonlinear mixed-effects model is used for the longitudinal observations and a parametric mixed-effects hazard model for repeated event times. We show the importance for parameter estimation of properly calculating the conditional density of the observations (given the individual parameters) in the presence of interval and/or right censoring. Parameters are estimated by maximizing the exact joint likelihood with the Stochastic Approximation Expectation-Maximization algorithm. The simulation study demonstrates that our methodology yields satisfactory results in this complex setting. As an illustration, such an approach is applied on two real world data sets. The first one is a primary biliary cirrhosis data, with longitudinal serum bilirubin measurements, jointly modelled with death. The second one is an epileptic seizure counts data, with seizures considered as interval censored. This workflow for joint models is now implemented in the MONOLIX software.

Keywords: *Interval censoring; Joint models; Maximum likelihood; Mixed-effects models; SAEM algorithm; Repeated time-to-events.*

5.1 Introduction

Joint models are a class of statistical methods for bringing together longitudinal data and time-to-event data into a unified framework. In the medical setting (the most common application of joint models), we often have, for a set of patients, time-to-event data of interest, e.g. tumor recurrences, epileptic seizures, asthma attacks, migraines, infectious episodes, heart attacks, injuries, hospital admissions, or even death. One may be interested in modeling the process inducing the event(s), using for example a suitable chosen (time-dependent or not) hazard function to describe the instantaneous chance of an event occurrence.

Simultaneously, for each patient we may be able to measure a longitudinal outcome (called biomarker in the following) and model its progression. Joint models come into the picture when there is a distinct possibility that a given longitudinal biomarker has a real influence on the time-to-event process. In such cases and in the most general way possible, the *joint model strategy* is to suggest a relationship between the biomarker and the hazard function, i.e., have its predicted value influence the instantaneous probability of the event of interest.

Early attempts to create joint models and apply them to biological settings were introduced by (Self and Pawitan, 1992) and (DeGruttola and Tu, 1994) with applications in AIDS research. What goes today as the standard joint model was introduced by (Faucett and Thomas, 1996) and (Wulfsohn and Tsiatis, 1997) and since that time, developments in the field have continued apace. We now briefly present the joint modeling framework as it currently stands, then explain the contribution of the present article to the state of the art. For a more thorough introduction, we point the reader to the monograph by (Rizopoulos, 2012b).

Joint modeling tries to characterize the relationship between a longitudinal biomarker's evolution and the risk of a given event, while also providing an acceptable model of the biomarker's evolution itself. First, let us concentrate on the longitudinal biomarker. Its evolution is often modeled under a *linear mixed-effects* framework (Laird and Ware, 1982; Harville, 1977; Verbeke and Molenberghs, 2000) using for instance splines (Ruppert et al., 2003) or B-splines with random effects (Rizopoulos et al., 2009; Brown et al., 2005). This framework takes into account the correlated nature of the measures for a given individual, while also allowing inter-individual random variability in key model parameters (e.g. slope, intercept). We can thus estimate the mean values of these parameters, as well as model/plot the evolution of the biomarker for each individual using their own estimated parameter values. Parameter estimation is often performed using a maximum likelihood strategy. However, linearity and the associated supposition of normally distributed parameters are strong hypotheses which are not necessarily representative of what is seen in real-life situations. For instance, in pharmacometrics and in particular pharmacokinetic-pharmacodynamic (PKPD) applications, linear models are usually not sufficient to satisfactorily model data. Consequently, nonlinear mixed-effects models have been largely adopted (Sheiner

and Beal, 1985; Lindstrom and Bates, 1990; Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997) even though they involve computationally taxing calculations when performing maximum likelihood, a stumbling block until recently. However, strategies such as the Stochastic Approximation EM (SAEM) algorithm (Kuhn and Lavielle, 2005), implemented in the MONOLIX software and R (R Development Core Team, 2008), have recently led to significantly faster methods for not only linear mixed-effects but also nonlinear mixed-effects models.

Next let us consider the event risk itself, modeled by a hazard function λ, which characterizes the distribution of the time-to-event process. The hazard function may be constant or vary as a function of time. Joint modeling is achieved by allowing the hazard function λ at time t to also depend on the value of the longitudinal biomarker predicted at t. In this framework, the baseline value of λ can be supposed constant or randomly varying across the population. Joint modeling then involves the simultaneous estimation of all the parameters from both parts of the model, again often by a maximum-likelihood strategy.

Due to significant complexity in the calculation of likelihoods for these models, initial approaches to fit them focused on two-stage methods (Self and Pawitan, 1992; Tsiatis et al., 1995), with the downside of often producing biased results (Dafny and Tsiatsis, 1998) in simulation studies. Full likelihood approaches have therefore been introduced to try to eliminate this bias (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Hsieh et al., 2006). Maximization of the log-likelihood function is then often attempted using the EM algorithm (Dempster et al., 1977), treating random effects as missing data. However, the integrals that must be approximated for joint models are a big computational problem, especially the integral with respect to the random effects. Indeed, this is perhaps the main bottleneck blocking the more frequent everyday use of joint models. The R package **JM** (Rizopoulos, 2012b) provides a set of procedures for solving these problems in the linear mixed-models framework. We will show in this article that the SAEM algorithm (Kuhn and Lavielle, 2005) can be extended to efficiently perform joint modeling in the more general nonlinear framework and in the presence of censored data, as described in the following.

Censoring is a defining characteristic of time-to-event data and poses problems for using many typical statistical tools (sample average and standard deviation, t-tests, linear regression) because these suppose we have complete information. Censoring takes several forms: left censoring (event time is before the individual enters the study), right censoring (event happens after the last recorded time) and interval censoring (event is known to happen between two time-points, but the exact event time is not known). If censoring is not correctly taken into account when calculating likelihoods, bias will be introduced into parameter estimation.

The current article therefore advances the state of the art in several ways. First, it presents time-to-events models for repeated events, including for censored events (interval-censored, right-censored) where there may be 0, 1 or many events in each interval. Second, it develops a framework for joint models combining *non-*

linear mixed effects models for continuous covariates/biomarkers with (perhaps repeated) time-to-events data. Third, for likelihood calculations it presents a rigorous calculation of the conditional density of the observations given the individual parameters in a wide variety of situations, and shows that bias is introduced if we approximate by replacing a censoring interval by its mean, or do not take into account when it is known that there is a maximum number of events. Fifth, it shows that the Stochastic Approximation Expectation Maximization (SAEM) algorithm (Kuhn and Lavielle, 2005) is not only capable, but also extremely fast, when it comes to performing maximum likelihood estimation for joint models. And last, it shows that we can also estimate the Fisher information matrix, the observed likelihood and the individual parameters under the same framework.

The article is structured as follows. In section 5.2 we first recall nonlinear mixed-effects models for the population approach and then repeated time-to-events models, and show how they can be combined as joint models. In section 5.3 we introduce the SAEM algorithm (Kuhn and Lavielle, 2005), proven to be efficient for maximum likelihood estimation of parameters in mixed-effects models in general, and nonlinear models in particular. We then describe how to estimate standard errors and individual parameters, and how to approximate the log-likelihood using importance sampling in order to calculate model selection criteria such as BIC.

For all of these calculations, we require an expression for the conditional density of the observations given the individual parameters, and furthermore, this expression depends on the situation we are interested in, whether it be repeated events with interval censoring, right-censored time-to-events, multiple events per interval, joint models, etc. Therefore, in section 5.4 we derive these expressions across a wide range of situations. In section 5.5, a series of simulated examples are provided to illustrate the methods in action and to show that the SAEM algorithm is accurate and extremely fast for the joint modeling of longitudinal and time-to-events data. Finally, in section 5.5.2, the methodology is applied on two real word data sets. The first one is a primary biliary cirrhosis data, and the second one is an epileptic seizure counts data, with seizures considered as interval censored.

5.2 Models

5.2.1 Nonlinear mixed-effects models for the population approach

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p(y_i|\psi_i)$ be the probability distribution of y_i, where ψ_i is a vector of parameters.

In the population framework, the vector of parameters ψ_i is assumed to be

drawn from a population distribution $p(\psi_i;\theta)$. Then, the probabilistic model is the joint probability distribution

$$p(y_i,\psi_i;\theta) = p(y_i|\psi_i)p(\psi_i;\theta). \tag{5.1}$$

To define a model for the data thus consists in defining precisely these two terms.

First, let us present ψ_i in its most general form:

$$\psi_i = H(\psi_{pop},\beta,c_i,\eta_i),$$

where ψ_{pop} is a "typical" value of the parameters in the population, β a set of coefficients (usually called fixed effects), c_i a vector of individual covariates and η_i the random component (usually called random effects). For example, in a linear model we assume that, up to some transformation, ψ_i is a linear function of the covariates and the normally distributed random effects:

$$h(\psi_i) = h(\psi_{pop}) + \beta c_i + \eta_i, \tag{5.2}$$

where h is some monotonic function (log, logit, probit, etc.) and $\eta_i \sim \mathcal{N}(0,\Omega)$. The set of population parameters that define the population distribution $p(\psi_i;\theta)$ of the individual parameters ψ_i is thus $\theta = (\psi_{pop},\beta,\Omega)$.

The conditional distribution $p(y_i|\psi_i)$ of the observations depends on the type of observations (continuous, categorical, count, time-to-event, etc.). We consider here two situations:

- ◇ observations are time-to-events, perhaps repeated (several events per individual are observed) and perhaps interval or right censored (times of events are not precisely known).
- ◇ observations are a combination of continuous values (some biomarker) and time-to-events. They are thus characterized by a joint model which describes the relationship between the two types of data.

5.2.2 Repeated time-to-event model

For several reasons, time-to-event data are not amenable to standard statistical procedures. One is that they are generally not symmetrically distributed. Typically, a histogram constructed from the times-to-events of a group of similar individuals will tend to be positively skewed. As a consequence, it is not reasonable to assume that data of this type have a normal distribution. This difficulty could be resolved by first transforming the data to give a more symmetric distribution, for example by taking logarithms, which results in the well know Accelerated Failure Time (AFT) model (Klein and Moeschberger, 1997). However, a more satisfactory approach is to adopt an alternative model for the distribution of the original data.

The main feature of time-to-event data that renders standard methods inappropriate is that times-to-event are frequently censored (Klein and Moeschberger,

1997), i.e., the time-points of interest may not be observed for some individuals. This may be because the study data are analyzed at a point in time where some individuals have not yet experienced the event of interest. Alternatively, the status of an individual at the time of the analysis might not be known because that individual has been lost to follow-up.

More precisely, suppose an individual who enters a study at time t_0 experiences a relevant event at time $t_0 + t$. However, t may be unknown, either because the individual has not experienced the event yet, or because they have been lost to follow-up. If they were last known to be alive at time $t_0 + c$, then c is called a censored event time. This censoring occurs to the right of (i.e., after) the last recorded time, and is therefore known as right censoring. Similarly, left censoring is encountered when the actual event time is early than the entry of the individual into the study. Left censoring occurs less than right censoring.

Another type of censoring is interval censoring. Here, individuals are known to have experienced the event within a given interval of time, but the exact event time is not known. Interval-censored data commonly arise in studies where there is a non-lethal end-point, such as the recurrence of a disease or condition. Several authors have focused on a single interval-censored event. For instance, (Kongerud and Samuelsen, 1991) and (Samuelsen and Kongerud, 1993) studied respiratory and asthmatic symptoms among Norwegian aluminum workers in which the time to the development of symptoms was only known to be between consecutive health examinations.

In summarizing time-to-event data, there are two main functions of interest, namely the survival function and the hazard function. The actual event time t can be regarded as the value taken by a non-negative random variable T. For the case of a single event process, the survival function $S(t)$ is defined as

$$S(t) = \mathbb{P}(T \geq t) = 1 - F(t),$$

where F is the cumulative distribution function of T. The hazard function $\lambda(t)$ is the instantaneous risk of experiencing an event at some time t; therefore S can also be written:

$$S(t) = e^{-\int_0^t \lambda(u)du}.$$

In the case of a repeated events process we have instead a sequence of event times (T_j) and are now interested in the probability of an event after t_j given the previous event at t_{j-1}:

$$\mathbb{P}(T_j > t_j | T_{j-1} = t_{j-1}) = e^{-\int_{t_{j-1}}^{t_j} \lambda(u)du}.$$

Under a population framework, we suppose a hazard function λ_i for each individual i:

$$\lambda_i(t) = \lambda(\psi_i, t).$$

As an example, consider the model with constant hazard (Karlsson et al., 2011) given by $\lambda_i(t) = \lambda_i$. Then, the duration between successive events has an exponential distribution with parameter λ_i, and the number of events in any interval

of length Δ has a Poisson distribution with parameter $\Delta\lambda_i$. Here, the vector of individual parameters reduces to $\psi_i = \lambda_i$.

In the most simple case, y_i is a vector of known event times: $y_i = (t_{i1}, t_{i2}, \ldots, t_{i\,n_i})$. But if we only know that events occur within certain intervals, then observations are the *number of events per interval*. Let $(I_{i1}, \ldots, I_{i\,n_i})$ be a set of disjoint time intervals for individual i relevant to the experimental design. We then can write $y_i = (k_{i1}, \ldots, k_{i\,n_i})$, where $k_{i\ell}$ is the number of events for individual i that have occurred in interval $I_{i\ell}$. Note that this includes the interval censored case with finite intervals $I_{i\ell}$, as well as the right censored case with $I_{i\,n_i} = [t_{\text{end}}\,, \infty)$.

5.2.3 Joint models

Besides the parametric form of the model, an essential point of joint modeling is the type of dependency between the longitudinal data model and the events. Suppose that we have a continuous biomarker of the form

$$b_{ij} = f\left(t_{ij}, \psi_i^{(1)}\right) + g\left(t_{ij}, \psi_i^{(1)}\right)\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i}. \tag{5.3}$$

Next, we connect this with RTTE via the hazard function given in general form:

$$\lambda_i(t) = \lambda\left(f(t, \psi_i^{(1)}), \psi_i^{(2)}\right).$$

Observations are therefore a combination of the $n_{1,i}$ continuous-valued biomarker measurements with the $n_{2,i}$ event times (if observed):

$$y_i = ((b_{ij}, 1 \leq j \leq n_{1,i}), (t_{i\ell}, 1 \leq \ell \leq n_{2,i}),$$

or with the number of events per interval in the case of censoring:

$$y_i = ((b_{ij}, 1 \leq j \leq n_{1,i}), (k_{i\ell}, 1 \leq \ell \leq n_{2,i}).$$

The vector of individual parameters $\psi_i = (\psi_i^{(1)}, \psi_i^{(2)})$ combines the individual parameters from the two parts of the joint model.

Example. Suppose that the biomarker measurements can be modeled by

$$b_{ij} = \gamma_i + \delta_i t_{ij} + a_i\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i}, \tag{5.4}$$

and are related to an event process with hazard function

$$\lambda_i(t) = \lambda_{0,i}e^{\alpha_i(\gamma_i + \delta_i t)}. \tag{5.5}$$

Here, $\psi_i = (\gamma_i, \delta_i, a_i, \lambda_{0,i}, \alpha_i)$. Formulas 5.4 and 5.5 thus characterize the probability model $p(y_i|\psi_i)$ of the observations. In the following, we will suppose certain parameters to be constants (i.e., without inter-individual variability), essentially

for reasons of identifiability. This is not to do with the distribution of the observations, but rather with the distributions of the individual parameters. In effect, building the statistical model for the individual parameters relies in part on deciding which components of ψ_i vary or not within the population. In the present example, we are unable to estimate all the parameters and their variability. In numerical trials in Section 5.5, we will therefore investigate this example further when supposing $\lambda_{0,i} = \lambda_0$, $\alpha_i = \alpha$ and $a_i = a$ to be population constants.

5.3 Tasks and methods

There are a variety of tasks that we are interested in performing here, whether it be estimating population parameters and their variation, estimating individual parameters or estimating the likelihood, the latter useful for performing likelihood ratio tests and calculating information criteria such as BIC.

In the following sections, we propose methodology for each of these tasks. We show that each requires calculation of the joint pdf (5.1) and in particular the conditional density $p(y_i|\psi_i)$ since under (5.2), $p(\psi_i;\theta)$ is straightforward to compute since it is derived from a Gaussian density.

5.3.1 Maximum likelihood estimation of the population parameters

Estimation in mixed-effects models consists of estimating the probability distribution of the ψ_i's in the population from the observations of the N subjects, i.e., in evaluating both the typical values in the population and the variability between subjects. More precisely, we aim to compute the maximum likelihood estimate of θ in (nonlinear) mixed-effects models by maximizing the observed likelihood $p(y;\theta)$. Estimation is complex because the N random vectors of parameters ψ_i are not observed and because there is a nonlinear relationship between the observations and the random effects defined in (5.2). For these reasons the likelihood function can not be explicitly given and its maximization is far from straightforward.

In a general way, linear and nonlinear mixed-effects models, including mixed-effects diffusion models, can be seen as incomplete data models in which the individual parameters $\pi = (\psi_1, \ldots, \psi_N)$ 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 $y = (y_1, \ldots, y_N)$. The EM algorithm (Dempster et al., 1977) iteratively performs parameter estimation in such models. The algorithm requires computing at each iteration the conditional expectation $E\left(\log p(y,\psi;\theta)|y,\theta^{(k-1)}\right)$, 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. For instance, in the SAEM algorithm (Delyon et al., 1999), the E-step is evaluated by a stochastic approximation procedure.

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 (Delyon et al., 1999; Kuhn and Lavielle, 2005):

- ◇ In the simulation step, $\theta^{(k-1)}$ is used to simulate the missing data $\psi_i^{(k)}$ under the conditional distribution $p(\psi_i|y_i,\theta^{(k-1)})$, $i=1,\ldots,N$.
- ◇ In the stochastic approximation step, the simulated data $\psi^{(k)}$ and the observations y are used together to update the stochastic approximation $Q_k(\theta)$ of the conditional expectation $E\left(\log p(y,\psi;\theta)|y,\theta^{(k-1)}\right)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \nu_k\left(\log p(y,\phi^{(k)};\theta) - Q_{k-1}(\theta)\right), \tag{5.6}$$

  where $(\nu_k)_{k>0}$ is a sequence of positive step sizes decreasing to 0 and starting with $\nu_1 = 1$.
- ◇ 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 (Delyon et al., 1999).

When an estimate $\hat{\theta}$ has been obtained, estimates of the standard errors of its components can be derived by estimating the Fisher information matrix $I(\hat{\theta}) = -\partial^2 \log(p(y;\theta))/\partial\theta\partial\theta'|_{\theta=\hat{\theta}}$ following the stochastic approximation procedure suggested in (Kuhn and Lavielle, 2005), which requires simulation of the ψ_i's under $p(\cdot|y,\hat{\theta})$ via a Metropolis-Hastings algorithm.

Estimates of the ψ_i's can also be derived from the conditional distribution $p(\psi_i|y_i,\hat{\theta})$ such as the conditional mode or the conditional mean. Whatever the estimate chosen, simulating this conditional distribution via Metropolis-Hastings, or maximizing it, requires computing the conditional distribution of the observations $p(y_i|\psi_i)$.

Standard model selection criteria such as BIC require calculation of the observed log-likelihood $\log(p(y;\hat{\theta}))$. As the log-likelihood cannot be computed in a closed form here, it is approximated using an importance sampling procedure. This consists of drawing $\psi^{(1)},\psi^{(2)},\ldots,\psi^{(M)}$ under a given sampling distribution $\tilde{\pi}$, and approximating the likelihood with:

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

Here also, we see that computation of $p(y_i|\psi_i)$ is required. In summary, calculation of $p(y_i|\psi_i)$ for the various cases (repeated events with interval censoring, right-censored time-to-events, joint models, etc.) is a critical step in the modeling process. In the following section, we therefore explicitly calculate this pdf for a wide range of cases.

5.4 Computing the probability distribution for repeated time-to-events

The aim of this section is to compute precisely the conditional distribution $p(y_i|\psi_i)$ for any subject i, when the vector of observations y_i only consists of (possibly repeated and possibly censored) time-to-events. For the sake of simplicity we only consider a single subject, and therefore omit the subscript i in notation. Also for simplicity we denote $\lambda(t)$ the hazard function at time t and omit the dependence with respect to the parameter ψ. We assume that the trial starts at time t_0 and ends at time t_{end}. Both t_0 and t_{end} are known. Let $T = (T_1, T_2, \ldots)$ be the (random) event times after t_0, and Λ be the cumulative hazard function:

$$\begin{aligned}\Lambda(a,b) &:= \int_a^b \lambda(t)\,dt \\ &= \Lambda(t_0,b) - \Lambda(t_0,a).\end{aligned}$$

By definition, recall that

$$\mathbb{P}(T_j > t_j | T_{j-1} = t_{j-1}) = e^{-\Lambda(t_{j-1},t_j)}. \tag{5.7}$$

We now distinguish between the three following situations:

- ◇ *exactly observed events:* the sequence of event times (T_j) is exactly known;
- ◇ *single interval-censored events:* we only know that the j-th event occurred between a_j and b_j, where $([a_j, b_j])$ is a sequence of intervals such that $a_j \geq b_{j-1}$;
- ◇ *multiple interval-censored events:* a sequence $(K_j;\, j \geq 1)$ is observed, where K_j is the number of events in interval $[b_{j-1}, b_j]$. Here, $([b_{j-1}, b_j])$ is a sequence of successive intervals with $b_0 = t_0 < b_1 < b_2 < \ldots$

We remark that the exactly observed event situation is the limiting case of the single interval-censored events one as $a_j \to b_j$. Similarly, the single interval-censored case is a special case of the multiple interval-censored case, where the number of events in each interval is 0 or 1.

Now suppose n is the number of observed events between t_0 and t_{end}. We further distinguish between two situations:

(i) the sequence of event times (T_j) is finite and all events are known to occur between t_0 and t_{end}, i.e., $t_0 < T_1 < T_2 < \ldots < T_n \leq t_{end}$;

(ii) the sequence of event times (T_j) is not bounded above by t_{end} and *at least one event* is known to occur after t_{end}. It is very important to realise that even if the $(n+1)$-th event is not observed, the information that $T_{n+1} > t_{end}$ is pertinent and contributes to the probability distribution of the observations. It should therefore always be taken into account.

5.4.1 Exactly observed events

i) the last event is observed. Assume that we observe n events at times t_1, t_2, ..., t_n and that no event occurs after t_{end}. The vector of observations is $y = (t_1, t_2, \ldots, t_n)$ and

$$\begin{aligned} p(y|\psi) &= p(t_1, t_2, \ldots, t_n) \\ &= p(t_1|t_0)p(t_2|t_1)p(t_3|t_2)\ldots p(t_n|t_{n-1}). \end{aligned}$$

By definition, $p(t_j|t_{j-1}) = \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}$. Thus,

$$p(y|\psi) = \prod_{j=1}^{n} p(t_j|t_{j-1}) \tag{5.8}$$

$$= \prod_{j=1}^{n} \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}\,. \tag{5.9}$$

ii) the last event is not observed. Assume that we observe n events at times t_1, t_2, ..., t_n and that an event is known to occur at time $T_{n+1} > t_{end}$. Here, the vector of observations is $y = (t_1, t_2, \ldots, t_n, t_{n+1} > t_{end})$ and

$$\begin{aligned} p(y|\psi) &= p(t_1, t_2, \ldots, t_n)\mathbb{P}(T_{n+1} > t_{end}|T_n = t_n) \\ &= p(t_1|t_0)p(t_2|t_1)p(t_3|t_2)\ldots p(t_n|t_{n-1})\mathbb{P}(T_{n+1} > t_{end}|T_n = t_n) \\ &= \left(\prod_{j=1}^{n} \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}\right) e^{-\Lambda(t_n,t_{end})}\,. \end{aligned}$$

5.4.2 Single interval-censored events

Assume that n events occur between t_0 and t_{end} but that we only know that $t_1 \in [a_1, b_1]$, $t_2 \in [a_2, b_2]$, ..., $t_n \in [a_n, b_n]$.

i) the last event is observed. Here, no event occurs after t_{end}. The vector of observations is $y = (t_1 \in [a_1, b_1], t_2 \in [a_2, b_2], \ldots, t_n \in [a_n, b_n])$ and its joint

probability distribution is:

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(T_1 \in [a_1, b_1], T_2 \in [a_2, b_2], \ldots, T_n \in [a_n, b_n]) \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \ldots \int_{a_n}^{b_n} p(t_1, t_2, \ldots, t_n)\, dt_1\, dt_2, \ldots, dt_n. \end{aligned}$$

Using equations 5.8-5.9,

$$\begin{aligned} p(t_1, t_2, \ldots, t_n) &= \prod_{j=1}^{n} \lambda(t_j) e^{-\Lambda(t_{j-1}, t_j)} \\ &= \left(\prod_{j=1}^{n} \lambda(t_j) \right) e^{-\sum_{j=1}^{n} \Lambda(t_{j-1}, t_j)} \\ &= \left(\prod_{j=1}^{n-1} \lambda(t_j) \right) \lambda(t_n) e^{-\Lambda(t_0, t_n)} \\ &= \left(\prod_{j=1}^{n-1} \lambda(t_j) \right) p(t_n|t_0). \end{aligned}$$

Thus, the multiple integral can be computed:

$$\begin{aligned} p(y|\psi) &= \left(\prod_{j=1}^{n-1} \Lambda(a_j, b_j) \right) \mathbb{P}(T_n \in [a_n, b_n] | T_{n-1} = t_0) \\ &= \left(\prod_{j=1}^{n-1} \Lambda(a_j, b_j) \right) \left(e^{-\Lambda(t_0, a_n)} - e^{-\Lambda(t_0, b_n)} \right). \end{aligned} \tag{5.10}$$

ii) the last event is not observed. Here, at least one event is known to occur after t_{end}. Thus, $y = (t_1 \in [a_1, b_1], t_2 \in [a_2, b_2], \ldots, t_n \in [a_n, b_n], t_{n+1} > t_{end})$. The previous result (see equation 5.10) holds with $a_{n+1} = t_{end}$ and $b_{n+1} = +\infty$:

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(T_1 \in [a_1, b_1], T_2 \in [a_2, b_2], \ldots, T_n \in [a_n, b_n], T_{n+1} > t_{end}) \\ &= \left(\prod_{j=1}^{n} \Lambda(a_j, b_j) \right) e^{-\Lambda(t_0, t_{end})}. \end{aligned} \tag{5.11}$$

5.4.3 Multiple events per interval

Consider first a single interval $[0, b]$ and let $k_{\max}$ ($k_{\max} \leq +\infty$) be the maximum number of events. Let K be the number of events in $[0, b]$.

For any $k < k_{\max}$, $K = k$ implies that the $(k+1)$-th event occurs after time b.

Then, for any $k < k_{\max}$,

$$\begin{aligned}
\mathbb{P}(K=k) &= \mathbb{P}(T_1 \in [0,b], \ldots, T_k \in [0,b], T_{k+1} > b;\ T_1 < \ldots < T_k < T_{k+1}) \\
&= \int_0^b \int_{t_1}^b \cdots \int_{t_{k-1}}^b \int_b^{+\infty} p(t_1, t_2, \ldots, t_k, t_{k+1})\, dt_1\, dt_2 \ldots dt_k\, dt_{k+1} \\
&= \int_0^b \int_{t_1}^b \cdots \int_{t_{k-1}}^b \int_b^{+\infty} \left(\prod_{j=1}^{k} \lambda(t_j) \right) p(t_{k+1}|t_0)\, dt_1\, dt_2 \ldots dt_k\, dt_{k+1} \\
&= \frac{\Lambda(0,b)^k}{k!} e^{-\Lambda(0,b)}. \qquad (5.12)
\end{aligned}$$

Remark. In the case of a constant hazard function $\lambda(t) = \lambda$, the inter-event times follow the exponential distribution with parameter λ. Then, the number of events in any interval of length b follows a Poisson distribution with parameter $\Lambda(0,b) = \lambda b$. For any $k < k_{\max}$,

$$\mathbb{P}(K=k) = \frac{(\lambda b)^k}{k!} e^{-\lambda b}. \qquad (5.13)$$

Equation 5.12 thus shows that this type of property still holds for non-constant hazard functions $\lambda(t)$.

So, for a bounded number of events ($k_{\max} < +\infty$),

$$\begin{aligned}
\mathbb{P}(K=k_{\max}) &= 1 - \sum_{k=0}^{k_{\max}-1} \mathbb{P}(K=k) \\
&= 1 - \sum_{k=0}^{k_{\max}-1} \frac{\Lambda(0,b)^k}{k!} e^{-\Lambda(0,b)}. \qquad (5.14)
\end{aligned}$$

Consider now n contiguous intervals $([b_{j-1}, b_j]; 1 \leq j \leq n)$, where $b_0 = t_0$ and $b_n = t_{end}$. Let K_j be the number of events in interval $[b_{j-1}, b_j]$.

i) the last event is observed.. Let $s_{n-1} = \sum_{j=1}^{n-1} k_j$. Using equations 5.12 and 5.14, we can show that

$$\begin{aligned}
p(y|\psi) &= \mathbb{P}(K_1 = k_1, K_2 = k_2, \ldots, K_n = k_{\max} - s_{n-1}) \\
&= \left(\prod_{j=1}^{n-1} \mathbb{P}(K_j = k_j) \right) \left(1 - \sum_{k=0}^{k_{\max}-s_{n-1}} \mathbb{P}(K_n = k) \right) \qquad (5.15) \\
&= \left(\prod_{j=1}^{n-1} \frac{\Lambda(b_{j-1}, b_j)^{k_j}}{k_j!} e^{-\Lambda(b_{j-1}, b_j)} \right) \qquad (5.16) \\
&\quad \times \left(1 - \sum_{k=0}^{k_{\max}-s_{n-1}} \frac{\Lambda(b_{n-1}, b_n)^k}{k!} e^{-\Lambda(b_{n-1}, b_n)} \right). \qquad (5.17)
\end{aligned}$$

ii) the last event is not observed. This implies that the first non-observed event occurs after t_{end}. Using equation 5.17, it is straightforward to show that if $\sum_{j=1}^{n} k_j < k_{\max}$, then

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(K_1 = k_1, K_2 = k_2, \ldots, K_n = k_n) \\ &= \prod_{j=1}^{n} \left(\frac{\Lambda(b_{j-1}, b_j)^{k_j}}{k_j!} e^{-\Lambda(b_{j-1}, b_j)} \right). \end{aligned} \tag{5.18}$$

5.5 Numerical experiments

5.5.1 Simulations

A series of simulation studies were conducted to evaluate the proposed methodology for calculating the maximum likelihood estimate of the population parameters. For each scenario, the SAEM algorithm was used with $M = 100$ simulated datasets for computing the parameter estimates $(\hat{\theta}_m, 1 \leq m \leq M)$. To assess statistical properties of the proposed estimators for each parameter, percentage-wise relative estimation errors $(REE_m, 1 \leq m \leq M)$ were computed:

$$REE_m = \frac{\hat{\theta}_m - \theta^\star}{\mid \theta^\star \mid} \times 100. \tag{5.19}$$

Using the REEs, the relative bias (RB) and relative root mean square errors (RRMSE) were computed for each parameter in each scenario:

$$RB = \frac{1}{M} \sum_{m=1}^{M} REE_m \tag{5.20}$$

$$RRMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} REE_m^2}\,. \tag{5.21}$$

Also, for each scenario the Fisher information matrix was estimated and standard errors $(\hat{se}_m, 1 \leq m \leq M)$ of the estimated parameters derived. Of course, the true standard errors $se^\star$ are unknown, but they can be empirically estimated by the root mean square errors (RMSE) of the estimated parameters:

$$RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \theta^\star)^2}. \tag{5.22}$$

To assess statistical properties of the proposed estimator of the standard errors, we can then compare them with the RMSE by computing relative estimation errors (in %) for each replicate:

$$REEse_m = \frac{\hat{se}_m - RMSE}{\mid \theta^\star \mid} \times 100. \tag{5.23}$$

Example 1

We first consider a basic RTTE model with constant hazard (See (Karlsson et al., 2011)) for each $i = 1, 2, \ldots N$, expressed as

$$\lambda_i(t) = \lambda_i \tag{5.24}$$
$$\log(\lambda_i) \sim \mathcal{N}\left(\log(\lambda), \omega^2\right). \tag{5.25}$$

We assume that events are observed between time $t_0 = 0$ and time $t_{\text{end}} = 12$. Furthermore the event times are assumed to be exactly known. We are thus in the situation described Section 5.4.1 with observed and right censored events. The conditional distribution of the observations is given in equation 5.9. The 100 datasets with 120 individuals in each were simulated under nine different scenarios with $\lambda \in \{0.01,\ 0.1,\ 1\}$ and $\omega \in \{0.1,\ 0.5,\ 1\}$. The distributions of the REE_m and $REEse_m$ are displayed in Figures 5.1 and 5.2.

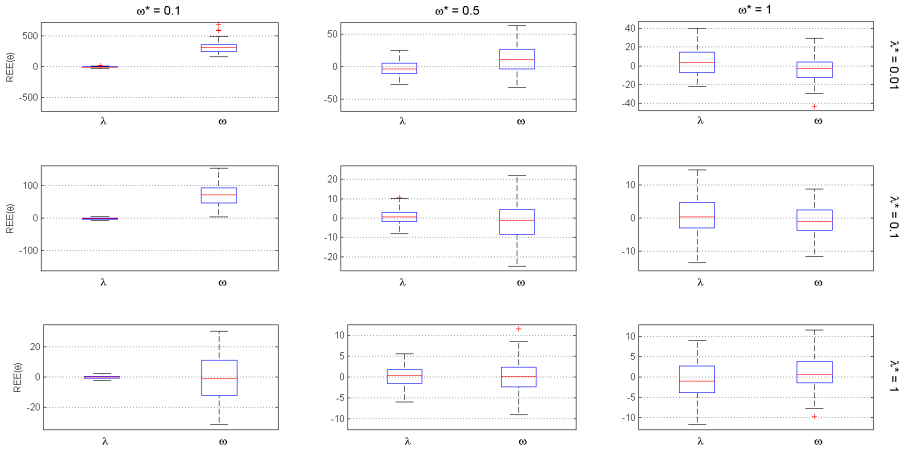

Figure 5.1: Relative estimation errors (in %) for λ and ω obtained with 9 different scenarios.

Figures 5.1 and 5.2 display the relative estimation errors for λ and ω and for their respective standard errors, obtained with 9 different scenarios ($\lambda = 0.01, 0.1, 1$ and $omega = 0.1, 0.5, 1$). This figures show that λ and its standard error are well-estimated generally and that the estimator is essentially unbiased. We see also that ω and its standard error are poorly estimated when both λ and ω are small, but as the true value of λ increases (i.e., more events happen), estimation of ω significantly improves and becomes unbiased.

Figure 5.3 shows the log-likelihood function for the parameter combinations $(\lambda^\star, \omega^\star) = (0.01, 0.5)$ on the left and $(\lambda^\star, \omega^\star) = (0.1, 0.5)$ on the right, and shows the disparity between the true parameters and the maximum likelihood estimates from two simulation runs. We see that the log-likelihood is much more concentrated

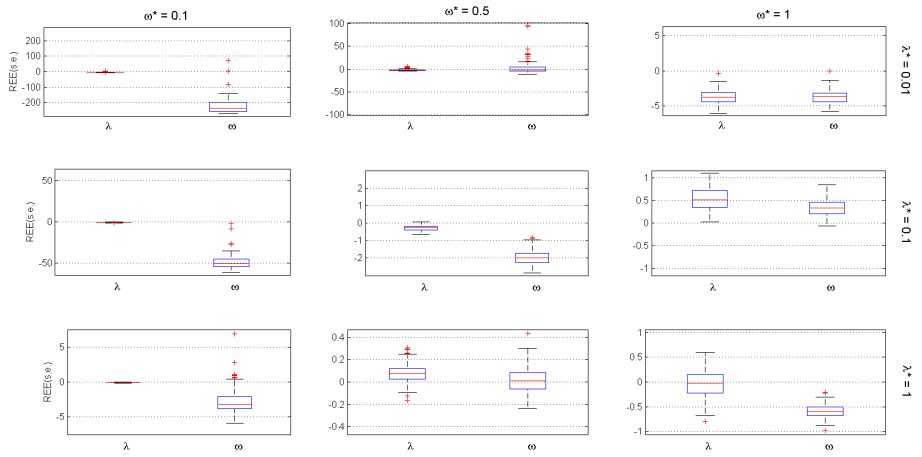

Figure 5.2: Relative estimation errors (in %) for the standard errors of $\hat{\lambda}$ and $\hat{\omega}$ obtained with 9 different scenarios.

around its global maximum when $\lambda = 0.1$. On the other hand, maximization of the log-likelihood cannot provide an accurate estimation of ω if $\lambda = 0.1$, i.e. if the number of events is too small. Note that Gaussian quadrature, efficient

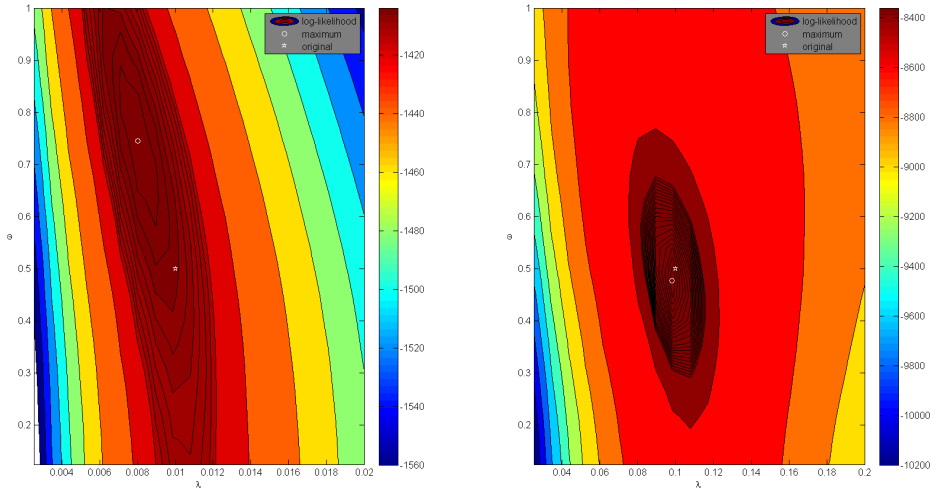

Figure 5.3: Observed log-likelihood as a function of λ and ω obtained with 2 different scenarios. Left: $(\lambda^\star, \omega^\star) = (0.01, 0.5)$, right: $(\lambda^\star, \omega^\star) = (0.1, 0.5)$.

for low-dimensional numerical integration, was used to compute the log-likelihood colormap. We can see why parameter estimation improves as we increase the true value of λ from 0.01 to 0.1 (and thus increase number of events per subject), keeping $\omega = 0.5$ fixed. In effect, the region of likelihood with a value close to the maximum is much more concentrated around the true value when λ increases

from 0.01 to 0.1, and consequently, maximum likelihood estimation will have better statistical properties.

Example 2

Using the same basic model as the previous example, here we look at the quality of the maximum likelihood estimation for interval-censored events. The simulation scheme is the following:

- ◇ $M = 100$ datasets with $N = 1000$ subjects in each.
- ◇ The true parameter values are $(\lambda^\star, \omega^\star) = (0.5, 0.5)$.
- ◇ The event process is single-event (e.g., death).
- ◇ Events are interval or right-censored. The intervals are contiguous and of length Δ.
- ◇ Observations occur between $t_0 = 0$ and $t_{\text{end}} = 24$.

As the events are interval or right-censored, the correct formula is (5.10). If we incorrectly take this information into account, for instance by considering that the event has happened at the interval midpoint, decreasing estimation quality and bias are introduced as the length Δ of the intervals increases. Figure 5.4 shows that very little information is lost (with respect to the case where we *do* know the exact times) if the correct formula is applied, whereas if the model is misspecified, the RRMSE and relative bias increase considerably as Δ increases.

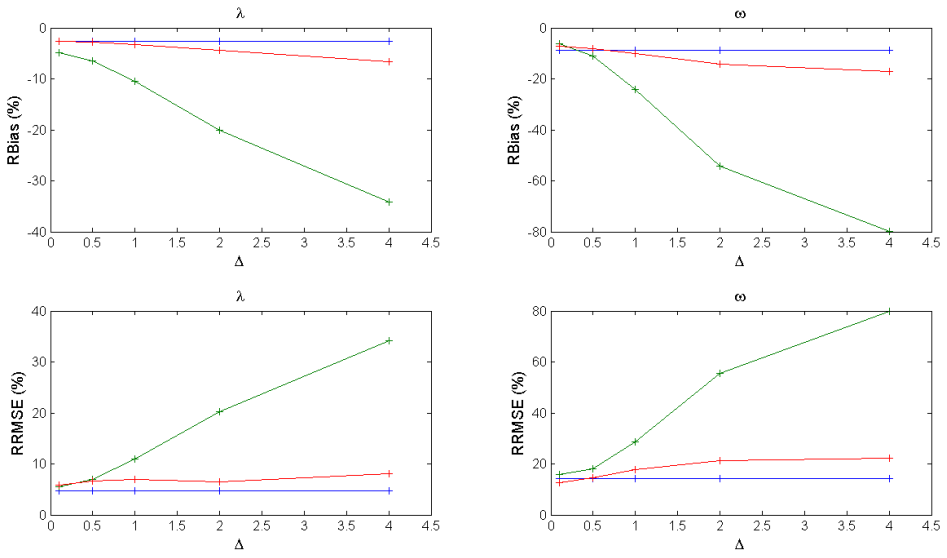

Figure 5.4: Relative bias and relative root mean square errors for λ and ω as a function of the width Δ of the censoring interval. Blue: using the exact event times; Red: taking correctly into account that the event is interval or right-censored; Green: taking incorrectly into account that the event is interval or right-censored.

Example 3

Here, we take the same interval or right-censored model as the previous examples but this time, we suppose that there are a maximum of $k_{\max} = 5$ events per subject. Note that the $k_{\max} = 5$ events are not necessarily observed during the trial period.

If for subject i we *have* observed $k_{\max} = 5$ events, the correct formula is (5.17), because it takes into account the fact that there are a maximum of 5 events and they have all been observed. If on the other hand the last event or events have not been observed, then equation (5.18) should be used when performing maximum likelihood estimation. Figure 5.5 shows what happens when this is not correctly taken into account. Indeed, as the width Δ of the intervals increases, the RRMSE

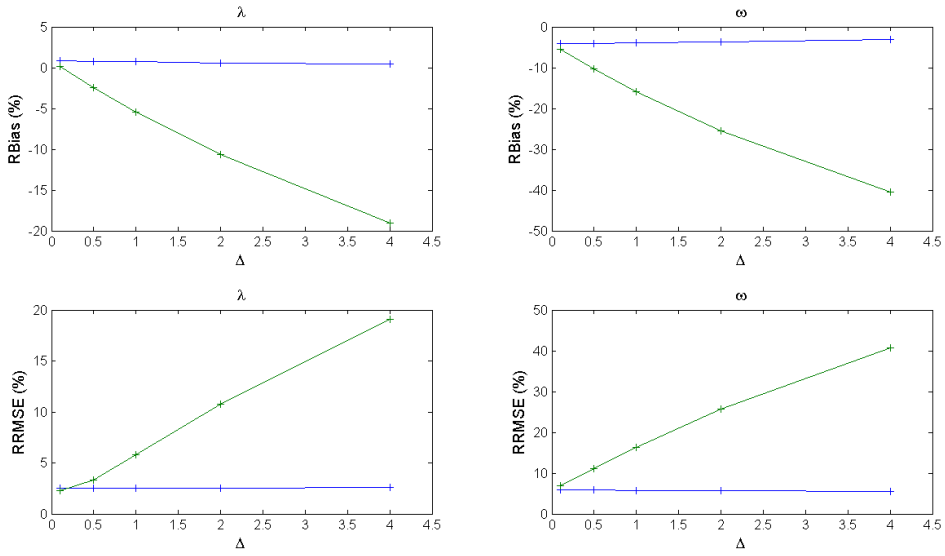

Figure 5.5: Relative bias and relative root mean square errors for λ and ω as a function of the width Δ of the censoring interval. Blue: taking into account the fact that the number of events is bounded ($k_{\max} = 5$); Green: ignoring the fact that the number of events is bounded.

and the (absolute) relative bias increase markedly in the misspecified case with respect to the correct one.

Example 4

This example, first introduced in Section 5.2.3, can be seen as an extension of the previous ones to joint modeling, or analogously as an RTTE model with a time-dependent covariate, taken here as a biomarker. We consider thus a joint model with a biomarker representing disease progress and an event which can occur

several times during the study. The model is:

$$\begin{aligned} b_{ij} &= \gamma_i + \delta_i t_{ij} + a\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i} \\ \lambda_i(t) &= \lambda e^{\alpha(\gamma_i + \delta_i t)}, \end{aligned}$$

where

$$\begin{aligned} \log(\gamma_i) &\sim \mathcal{N}(\log \gamma_{\text{pop}} + \beta \mathrm{C}_i,\ \omega_\gamma^2) \\ \log(\delta_i) &\sim \mathcal{N}(\log \delta_{\text{pop}},\ \omega_\delta^2) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, 1). \end{aligned}$$

Here, the hazard increases exponentially as the disease progresses linearly. C_i represents the treatment covariate, which takes values 0 (untreated) and 1 (treated). Treatment is associated with an effect which produces an immediate reduction of the slope of the disease progress.

Remark. The model is linear with respect to the parameters γ_i and δ_i. Nevertheless, methods developed for linear models (e.g. (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Hsieh et al., 2006)) cannot be used here since the model is not linear with respect to the random effects: the continuous observations are not normally distributed since γ_i and δ_i are not normally distributed. A more advanced method, such as SAEM for example, is therefore required for estimating the population parameters of the model.

For the model in question, we consider the following design:

- ◇ the biomarker (b_{ij}) is observed at times 0, 25 and 50 weeks: ($n_{i,1} = 3$).
- ◇ possibility of repeated interval-censored events until the end of the experiment at $t_{\text{end}} = 50$ weeks. Observations are the number of events in each 5 week period between $t_0 = 0$ and $t_{\text{end}} = 50$ (so $k_{i,1}$ is the number of events between 0 and 5,..., $k_{i,10}$ the number of events between 45 and 50). Thus, $n_{i,2} = 10$.
- ◇ parameter values are $\gamma_{\text{pop}} = 1$, $\delta_{\text{pop}} = 100$, $\beta = -0.3$, $\alpha = 0.02$, $\lambda = 0.01$, $\omega_\gamma = 0.1$, $\omega_\lambda = 0.1$, $a = 1$.

We suppose the total number of subjects in the trial is $N = 1000$. Note that we take a large N here because our fundamental goal is not to show in detail the performance of the maximum likelihood estimation. Rather, it is to show that the SAEM algorithm is effective in this framework: non-linear model, repeated interval-censored events, that it is fast and that it leads to little or no bias as well as small REEs.

We found that the SAEM algorithm performed well with the given model and experimental design. First, it was fast, taking 82 seconds on an Intel(R) Core(TM) i7-2760QM laptop with a 2.4 GHz processor. Figure 5.6 shows the convergence of the parameter estimates in a typical run, requiring less than 100 iterations for all parameters. Figure 5.7 shows that there is little or no bias in the parameter estimation, and consistently small REEs across the trials.

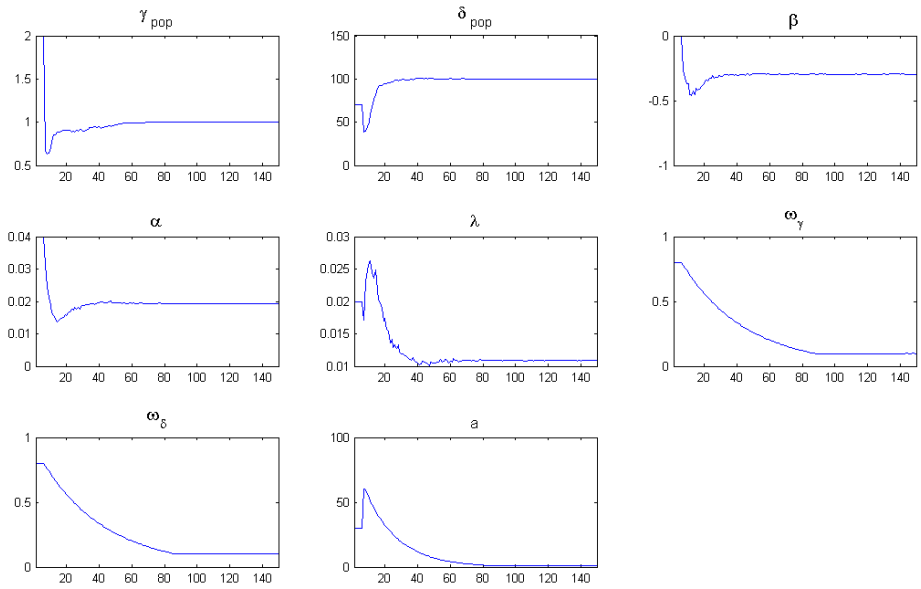

Figure 5.6: Convergence of the SAEM algorithm

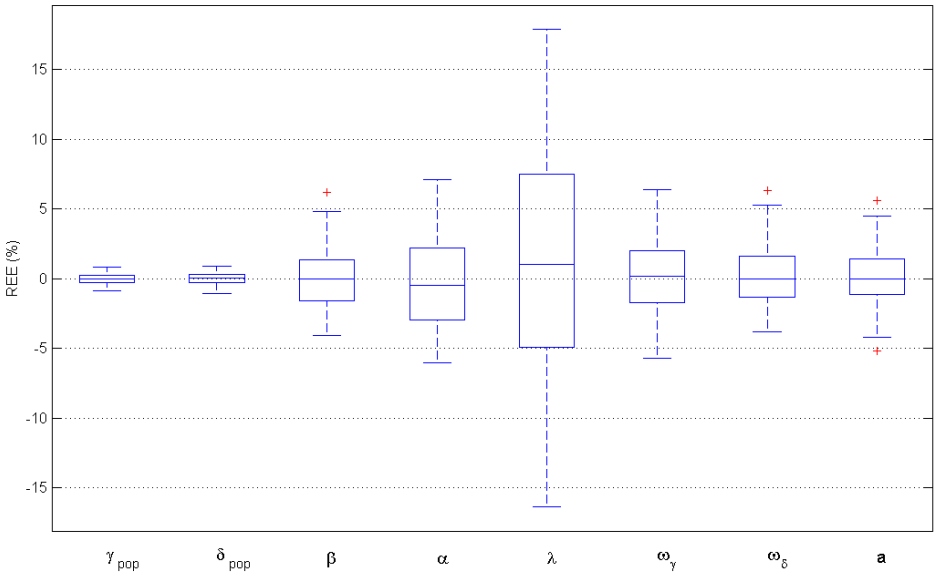

Figure 5.7: Relative estimation errors (in %) for the joint model.

5.5.2 Applications

Primary Biliary Cirrhosis Data

This well known dataset comes from a study conducted by the Mayo Clinic from 1974 to 1984. The study includes 158 patients who received D-penicillamine and 154 who received a placebo. Patient survival is the outcome of main interest. By the end of the study, 140 patients had died and 172 were still alive. Several biomarkers, including serum bilirubin, were measured during the study. A total of 1945 measurements of serum bilirubin is available.

Different joint models for this data are proposed in (Rizopoulos, 2012b). All of these joint models assume a linear mixed effects model for the longitudinal data. We will show that our approach provides straightforward extensions to non linear mixed effect models.

Following (Rizopoulos, 2012b), we use the following model for the serum bilirubin:

$$m_i(t) = c_{0,i} + c_{1,i}t + c_{2,i}t^2 \tag{5.26}$$
$$\log b_{ij} = m_i(t_{ij}) + a\varepsilon_{ij} \tag{5.27}$$

Here, $m_i(t)$ is the predicted concentration of bilirubin for patient i at time t and b_{ij} its measured concentration at time t_{ij}.

We use a simple proportional hazard model for the survival data

$$\lambda_i(t) = \lambda_{i,0}e^{\alpha_i\, m_i(t)} \tag{5.28}$$

The vector of individual parameters for patient i is $\psi_i = (c_{0,i}, c_{1,i}, c_{2,i}, \lambda_{0,i}, \alpha_i)$ Different statistical models for the ψ_i's were compared assuming first normal distributions for the $c_{\ell,i}$ and fixed parameters $\lambda_{0,i} = \lambda_0$ $\alpha_i = \alpha$.

We then considered a latent class model for the longitudinal data, assuming that the population is heterogeneous and constituted of two subpopulations that cannot be clearly identified by any of the available covariates. In other words, we assumed a mixture of 2 normal distributions for the $c_{\ell,i}$. Let (z_i) be a sequence of latent variables such that $z_i = 0$ if patient i belongs to subpopulation 1 and $z_i = 1$ if patient i belongs to subpopulation 2.

We also introduced the treatment (D-penicillamine/placebo) as a categorical covariate: let (d_i) be a sequence of observed variables such that $d_i = 0$ if patient i receives the placebo and $d_i = 1$ if patient receives the active treatment.

The statistical model for the individual parameters can therefore be described

as follows:

$$
\begin{aligned}
c_{0,i} &= c_0 + \beta_{0,z} z_i + \beta_{0,d} d_i + \eta_{0,i} \quad ; \quad \eta_{0,i} \sim \mathcal{N}(0, \omega_0^2) && (5.29)\\
c_{1,i} &= c_1 + \beta_{1,z} z_i + \beta_{1,d} d_i + \eta_{1,i} \quad ; \quad \eta_{1,i} \sim \mathcal{N}(0, \omega_1^2) && (5.30)\\
c_{2,i} &= c_2 + \beta_{2,z} z_i + \beta_{2,d} d_i + \eta_{2,i} \quad ; \quad \eta_{2,i} \sim \mathcal{N}(0, \omega_2^2) && (5.31)\\
\lambda_{0,i} &= h_0 + \beta_{\lambda,d} d_i && (5.32)\\
\alpha_i &= \alpha + \beta_{\alpha,d} d_i && (5.33)
\end{aligned}
$$

A diagonal variance-covariance matrix is assumed for the random effects.

Extension of the SAEM algorithm for mixture of mixed effects models have been developed and implemented in Monolix. We combined this method for mixture models with the proposed methods for joint models in order to simultaneously fit the longitudinal and survival data.

Table 5.1 provides the estimations of the parameters of the model.

parameter	estimation	standard error	$P(\|\beta\| > \|\beta^{\text{obs}}\|)$
c_0	0.846	0.042	
$\beta_{0,d}$	-0.092	0.056	0.10
$\beta_{0,z}$	1.26	0.067	$<10^{-4}$
c_1	0.068	0.021	
$\beta_{1,d}$	0.002	0.027	0.94
$\beta_{1,z}$	0.009	0.040	0.81
c_2	0.0054	0.0022	
$\beta_{2,d}$	-0.0009	0.003	0.77
$\beta_{2,z}$	0.069	0.007	$<10^{-4}$
λ_0	0.0039	0.0011	
$\beta_{\lambda,d}$	0.002	0.42	0.999
α	1.64	0.11	
$\beta_{\alpha,d}$	-0.004	0.15	0.999
ω_0	0.431	0.022	
ω_1	0.196	0.011	
ω_2	0.0135	0.0015	
a	0.209	0.004	

Table 5.1: Primary Biliary Cirrhosis Data: estimation of the population parameters.

These results show that there is no any significant effect of the treatment neither on the serum bilirubin nor on the survival probability. On the other hand, two different typical profiles describe the serum bilirubin kinetics since the distribution of (c_0, c_2) is pretty well described by a mixture of two normal distributions.

Epileptic seizure counts

The data base consisted double blind placebo controlled parallel group multicenter studies. All recruited patients were on standard anti-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. Different count data models have been proposed, including a mixture of two Poisson models (Miller and al., 2003) or a hidden Markov model (Delattre et al., 2012). Any count data model assumes that the probability function of the number of seizures is piecewise constant, *i.e.* constant during a unit of time. We propose to extend this approach considering the seizures as interval censored events. Then, following 5.4.3, this approach is equivalent to consider the seizures count as a non homogenous Poisson process which intensity is a continuous function of time. The hazard function was modeled assuming a constant hazard in both phases and an smooth transition between the two phases:

$$\lambda_i(t) = \begin{cases} a_i & \text{if } t \leq t_0 \\ b_i + (a_i - b_i)e^{-c_i(t-t_0)} & \text{if } t > t_0 \end{cases} \tag{5.34}$$

where t_0 is the time when the active treatment starts. We used the following statistical model for describing the inter patient variability of the individual parameters a_i, b_i and c_i:

$$\log(a_i) = \log(a) + \eta_{a,i} \quad ; \quad \eta_{a,i} \sim \mathcal{N}(0, \omega_a^2) \tag{5.35}$$

$$\log(b_i) = \log(b) + \beta_b \log(1 + D_i) + \eta_{b,i} \quad ; \quad \eta_{b,i} \sim \mathcal{N}(0, \omega_b^2) \tag{5.36}$$

$$\log(c_i) = \log(c) + \beta_c \log(1 + D_i) \tag{5.37}$$

where D_i is the amount of gabapentin administrated to patient i.

The estimated parameters are displayed Table 5.2 and the distribution of the hazard functions associated to different doses of gabapentin are displayed Figure 5.8. Even if the inter patient variability of the hazard function is very large, we can see a slight placebo effect and a mild effect of gabapentin on the seizures rate.

5.6 Discussion

Joint modeling of longitudinal biomarkers and time-to-events data is an important step in the improvement in understanding of the connection between biological changes in time and the arrival of a (perhaps critical) event to the patient. In recent years, linear mixed-effects models have been coupled with time-to-single event processes and parameter estimation performed, often using maximum likelihood

parameter	estimation	standard error	$P(\|\beta\| > \|\beta^{\text{obs}}\|)$
a	0.491	0.019	
b	0.463	0.022	
β_b	-0.239	0.054	$<$0.0001
c	0.097	0.013	
β_c	0.605	0.230	0.0076
ω_a	1.05	0.027	
ω_b	1.1	0.029	
$\rho_{a,b}$	0.889	0.009	

Table 5.2: Epileptic daily seizures count: estimation of the population parameters.

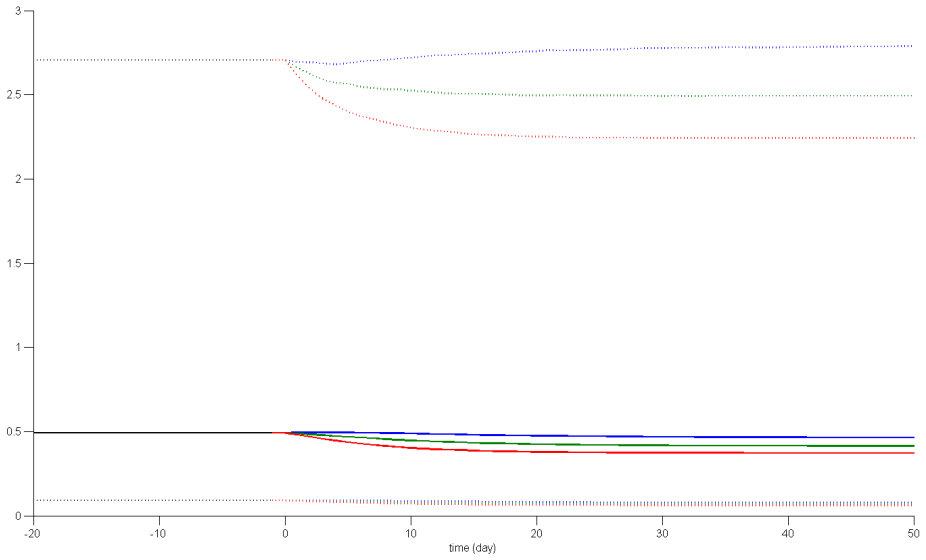

Figure 5.8: Hazard function for the epileptic seizures count data. Different hazard functions associated to different doses of gabapentin are displayed, blue: 0g (placebo) ; green: 0.6g ; red : 1.5g. The median hazard functions are displayed with solid lines, the 90% prediction intervals with dotted lines.

coupled with the EM algorithm. When there are more than a negligible number of random effects in the model, likelihood calculations are a huge bottleneck, discouraging use of these methods.

Here, we have shown that the SAEM algorithm is more than capable of performing parameter estimation for joint models where the mixed-effects can be nonlinear, the events can be repeated, all in the presence of right and/or interval censoring. To be able to implement SAEM for joint models in the afore-mentioned range of cases, we have for each derived precise expressions for the conditional likelihood of the observations given the individual parameters. In a series of simulation studies, we have shown that the SAEM algorithm converges for joint models in a matter of minutes rather than hours or days. As a consequence, we can also

quickly estimate the Fisher information matrix, the observed likelihood and the individual parameters.

SAEM for joint models is intuitively implemented in the MONOLIX software: in order to pass from nonlinear mixed effects modeling to joint modeling, *all that is required* of the modeler is to provide the parametric form of the hazard function. Note that several diagnostic tools are also implemented in MONOLIX based on Kaplan-Meier plots; further details are beyond the scope of the article.

In conclusion, now that there exists a simple, fast and high-performance tool for joint modeling, we believe that there is no reason why these methods should not be more used in everyday statistical practice.

5.7 Appendix: Several examples on the computation of the likelihood of a RTTE model

This section provides some practical technics to compute the likelihood of a RTTE model under interval censoring that could help the readers for a better understanding of the conditional likelihoods computed in Section 5.4.
We assume that $t_0 = 0$.

Id	Time	Y	log-likelihood	Comment
1	5	1	$\log(\lambda(5)) - \Lambda(0,5)$	Event at $T = 5$
2	5	0	$-\lambda(0,5)$	Event after $T = 5$
3	10	0	$-\lambda(0,10)$	Event after $T = 10$
4	7	1	$\log(\lambda(7)) - \Lambda(0,7)$	Event at $T = 7$

Table 5.3: One single event: Exact time of event.

Id	Time	Y	log-likelihood	Comment
1	5	1	$\log(1 - \exp[-\Lambda(0,5)])$	Event in [0, 5]
2	5	0	$-\Lambda(0,5)$	Event after $T = 5$
3	10	0	$-\Lambda(0,10)$	Event after $T = 10$
4	5	0	$-\Lambda(0,5)$	Event in [5, 7]
4	7	1	$\log(1 - \exp[-\Lambda(5,7)])$	

Table 5.4: One single event: Interval censored event.

Id	Time	Y	Flag	log-likelihood	Comment
1	5	1	1	$\log\left(1-\exp[-\Lambda\left(0,5\right)]\right)$	Event in [0, 5]
2	5	0	0	$-\Lambda\left(0,5\right)$	Event after $T=5$
3	10	0	0	$-\Lambda\left(0,10\right)$	Event after $T=10$
4	5	0	0	$-\Lambda\left(0,5\right)$	Event in [5, 7]
4	7	1	2	$\log\left(1-\exp[-\Lambda\left(5,7\right)]\right)$	

Table 5.5: One single event: Combined event, with Flag $=1$ if exact and Flag $=2$ if interval.

Id	Time	Y	log-likelihood	Comment
1	5	1	$\log\left(\lambda(5)\right)-\Lambda\left(0,5\right)$	Event 1 at $T=5$
1	10	1	$\log\left(\lambda(10)\right)-\Lambda\left(5,10\right)$	Event 2 at $T=10$
1	15	0	$-\Lambda\left(10,15\right)$	Event3 after $T=15$
2	5	0	$-\Lambda\left(0,5\right)$	Event1 after $T=5$
3	5	1	$\log\left(\lambda(5)\right)-\Lambda\left(0,5\right)$	Event 1 at $T=5$
3	7	1	$\log\left(\lambda(7)\right)-\Lambda\left(5,7\right)$	Event 2 at $T=7$
4	9	1	$\log\left(\lambda(9)\right)-\Lambda\left(0,9\right)$	Event 1 at $T=9$

Table 5.6: Succession of isolated events: Exact time of events.

Id	Time	Y	log-likelihood	Comment
1	5	1	$\log\left(\Lambda\left(0,5\right)\right)-\Lambda\left(0,5\right)$	Event 1 in [0, 5]
1	10	1	$\log\left(\Lambda\left(5,10\right)\right)-\Lambda\left(5,10\right)$	Event 2 in [5, 10]
1	15	0	$-\Lambda\left(10,15\right)$	0 event in [10, 15]: event3 after $T=15$
2	5	0	$-\Lambda\left(0,5\right)$	Event1 after $T=5$
3	5	1	$\log\left(\Lambda\left(0,5\right)\right)-\Lambda\left(0,5\right)$	Event 1 in [0, 5]
3	7	1	$\log\left(1-\exp[-\Lambda\left(5,7\right)]\right)$	Event 2 in [5, 7] (The last)
4	5	0	$-\Lambda\left(0,5\right)$	Event 1 after $T=5$
4	9	1	$\log\left(\Lambda\left(5,9\right)\right)-\Lambda\left(5,9\right)$	Event 1 in [5, 9] (Not the last one: Y=0 after)
4	9	0	$\log\left(1\right)$	Event 2 after $T=9$

Table 5.7: Succession of isolated events: Interval censored events.

Id	Time	Y	Flag	log-likelihood	Comment
1	5	1	1	$\log(\lambda(5)) - \Lambda(0,5)$	Event 1 at $T = 5$
1	10	1	2	$\log(\Lambda(5,10)) - \Lambda(5,10)$	Event 2 in $[5,\ 10]$
1	15	0	0	$-\Lambda(10,15)$	event3 after $T = 15$
2	5	0	0	$-\Lambda(0,5)$	Event1 after $T = 5$
3	5	1	2	$\log(\Lambda(0,5)) - \Lambda(0,5)$	Event 1 in $[0,\ 5]$
3	7	1	1	$\log(\lambda(7)) - \Lambda(5,7)$	Event 2 at $T = 7$
4	5	0	0	$-\Lambda(0,5)$	Event 1 after $T = 5$
4	9	1	2	$\log(\Lambda(5,9)) - \Lambda(5,9)$	Event 1 in $[5,\ 9]$ (Not the last one: Y=0 after
4	9	0	0	$\log(1)$	Event 2 after $T = 9$

Table 5.8: Succession of isolated events: Combined events, with Flag $= 1$ if exact, Flag $= 2$ if interval and Flag $= 0$ if right censored.

Chapter 6

Conclusion et perspectives

Nos travaux de Thèse portent sur des développements méthodologiques dans le cadre de l'estimation des paramètres des MNLEM. Ils permettent de répondre à l'une des préoccupations de l'industrie pharmacologique, qui est le développement et la mise en oeuvre de méthodologies de plus en plus performantes pour la modélisation de phénomènes complexes.

Il existe à ce jour très peu de méthodes disponibles pour l'estimation par maximum de vraisemblance dans les modèles de mélange à effets mixtes. Les méthodes mises en oeuvre dans le package R nlme et dans le logiciel NONMEM sont basées sur une linéarisation de la vraisemblance. Il est néanmoins reconnu que ces méthodes posent des problèmes pratiques réels dans plusieurs situations (le biais, forte influence des valeurs initiales, mauvaise convergence...). De plus les propriétés théoriques des estimateurs obtenus par ces méthodes restent inconnues dans la plupart des situations. Des méthodes de type EM ont été développées dans le cadre des modèles mixtes incluant des mélanges. Ces méthodes effectue généralement une approximation de l'étape E par des intégrations Monte-Carlo. (De la Cruz-Mesia et al., 2008) proposent d'intégrer la vraisemblance $p(y, \varphi; \theta)$ des données complètes, pour déterminer la vraisemblance marginale $p(y; \theta)$ des observations dans chaque cluster. Les algorithmes du type MCEM utilisent les intégration Monte-Carlo pour déterminer la distribution conditionnelle $p(\varphi|y; \theta)$. Ces méthodes peuvent nécessiter un temps d'exécution rédhibitoire en pratique, notamment lorsque le modèle structurel est complexe, ce qui est généralement le cas dans plusieurs applications aux modèles PKPD.

Nous proposons une extension de l'algorithme SAEM pour les mélanges de modèles non linéaires à effets mixtes. Le modèle considéré est très général et inclut des mélanges de distributions, des mélanges de modèles structurels et des mélanges de modèles résiduels. La convergence de l'algorithme MSAEM vers un maximum (local) de la vraisemblance des observations est obtenue sous des hypothèses générales. L'algorithme s'avère être extrêmement rapide, principalement par le fait que les intégrations Monte-Carlo sont remplacées par des approximations stochastiques. Il a en effet besoin de la simulation d'une unique Chaîne de Markov à chaque itération. De plus, l'algorithme MSAEM reste très peu sensible aux valeurs initiales, constituant de ce fait une très bonne propriété pour des applications pratiques. Cette méthode d'estimation est donc extrêmement prometteuse, tant au niveau des résultats théoriques de convergence qu'au niveau algorithmique. Cet algorithme, destiné aux mélanges de MNLEM est désormais disponible sous le logiciel MONOLIX , un des leaders dans l'industrie pharmacologique. MONOLIX est gratuit pour les étudiants et pour des recherches académiques, et contient quelques examples de démos sur les mélanges de MNLEM.

Nous proposons, dans le but de stratifier les patients ayant le VIH sur la base des mesures réelles longitudinales de charges virales, deux types de modèles. Un mélange de 3 modèles structurels a été utilisé pour classifier les patients en 3 groupes distincts : le groupe des "répondant", caractérisé par une décroissance continue de la charge virale ; le groupe des "rebondissant", caractérisé par une

décroissance de la charge virale suivie d'une phase de rebond ; le groupe des "non-répondant" caractérisé par une non-décroissance de la charge virale. Les résultats obtenus par ce modèle montrent une bonne précision aussi bien sur l'estimation des paramètre du mélange que sur la classification des patients, du moment que l'évolution des charges virales des individus considérés est effectivement décrite par l'un des modèles structurels considérés. Néanmoins, quand l'évolution de la charge virale ne peut-être décrite convenablement par l'un des trois modèles structurels, les fits individuels deviennent mauvais et les prédictions des classes douteuses. Nous avons ensuite considéré que chaque patient contenait des virus qui répondent, ceux qui répondent partiellement et ceux qui ne réponde pas du tout au traitement. Chaque patient est désormais partiellement "répondant", partiellement "rebondissant" et partiellement "non-répondant". Cette approche améliore clairement les fits individuels et permet une réponse plus nuancée à la question de savoir si un individu répond de manière adéquate au traitement ou pas. Sur la base de ces résultats, la deuxième approche semble la plus indiquée pour des éventuels changements de régime thérapeutique chez les patients.

Nous envisageons des extensions de l'algorithme MSAEM pour s'affranchir de plusieurs problèmes d'ordre pratique. Une stratégie optimale de construction du modèle pourrait s'avérer utile à cet effet, pour sélectionner simultanément la forme du modèle adéquat ainsi que le nombre de classes. Nous avons considéré dans cette Thèse uniquement la méthode du maximum de vraisemblance pour les mélanges considérés. Les méthodes Bayésiennes peuvent être considérées en utilisant les simulations de la loi à posteriori via des méthodes MCMC telles que celles proposées par (Frühwirth-Schnatter, 2006; De la Cruz-Mesia et al., 2008).

Nous avons également proposé dans cette Thèse de modéliser de manière conjointe une réponse longitudinale en utilisant un modèle non linéaire à effets mixtes, et une suite de délais successifs jusqu'à un évènement récurrent (censuré à droite ou par intervalle) en utilisant un modèle de risque mixte. L'inférence statistique s'est faite via un algorithme de type SAEM convergent rapidement vers la cible. Il existe plusieurs logiciels traitant de l'estimation dans les modèles conjoints d'une variable longitudinale et d'un unique évènement terminal. On peut citer, entre autres, le plus récent qui est un package R dénommé JM, proposé par (Rizopoulos, 2010). Ce package ne considère néanmoins que des modèles linéaires mixtes pour la variable longitudinale. Ceci pourrait être considéré comme une importante restriction, spécialement dans le cadre des modèles PKPD, où les modèles structurels sont généralement des solutions d'équations différentielles ordinaires ou stochastiques. Notre méthode offre ainsi une grande flexibilité au niveau de la variable longitudinale, permettant ainsi d'expliquer une très grande variété de phénomènes biologiques. La plupart des méthodes utilisées dans le cadre des modèles conjoints classiques (modèle linéaire mixte pour la variable longitudinale et modèle de risque pour l'unique évènement terminal) utilisent des approximations du modèle conjoints. JM par exemple, utilise une méthode de Gauss-Hermite ou une méthode pseudo-adaptative de Gauss-Hermite d'intégration (Rizopoulos, 2012a) pour approcher la vraisemblance du modèle conjoint. Nous proposons de maxi-

miser la vraisemblance conjointe exacte, sans approximations. Au vu de toutes ces observations, on peut considérer que le modèle que nous proposons dans cette Thèse est une avancée importante dans l'état de l'art des modèles conjoints car en plus de la flexibilité utilisée pour le modèle de la variable longitudinale, nous proposons aussi un modèle permettant de considérer les évènements répétés et éventuellement censurés à droite ou par intervalles lors de l'étude. La méthodologie générale SAEM proposée pour les modèles conjoints s'appliquent aisément aux modèles conjoints "classiques" rencontrés dans diverses publications, et aux modèles d'évènements récurrents. Cette méthodologie est désormais disponible sous MONOLIX. Son application à des données réelles de patients atteints d'une cirrhose biliaire primitive met en exergue la relation entre l'évolution du sérum bilirubin et la survie des patients.

Nous envisageons, de considérer dans de futurs travaux, des situations où on aurait plusieurs évènements de type distinct lors de l'étude. Une idée serait alors de stratifier la fonction de risque (en fonction des différents types d'évènements) via la fonction de risque de base ou des covariables, constituant ainsi une extension des risques compétitifs aux évènements récurrents.

Table des figures

Liste des tableaux

Références

Aitkin, M. and Rubin, D. B. (1985). Estimation and Hypothesis testing in Finite Mixture Models. *Journal of the Royal Statistical Society*, 47 :67–75.

Allassonnière, S., Kuhn, E., and Trouvé, A. (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16 :641–678.

Andrieu, C., Moulines, E., and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1) :283–312.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non Gaussian clustering. *Biometrics*, 49 :803–821.

Barndoff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley., Chichester.

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

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725.

Bradley, R. and Gart, J. (1962). The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1-2) :205–214.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, 88 :9–25.

Brown, E., Ibrahim, J., and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61 :64–73.

Bryant, P. G. (1991). Large Sample Results for Optimization Based Clustering Methods. *Journal of Classification*, 8 :31–44.

Celeux, G. (1998). *Bayesian inference for mixtures : The label switching problem.* Compstat 98-Proc. in Computational Statistics, Physica, Heidelberg, in : payne, r., green, p.j. edition.

Celeux, G. and Diebolt, J. (1986). L'algorithme SEM : Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. (The SEM algorithm : An algoritm of probabilistic learning for the determination of mixtures of densities). *Rev. Stat. Appl.*, 34 :35–52.

Celeux, G. and Diebolt, J. (1990). Une version de type recuit simulé de l'algorithme EM. *C. R. Acad. Sci. Paris Sér. I Math*, 310 :119–124.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics Stochastics Rep*, 41 :119–134.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm and two Stochastic Versions. *Computational Statistics and Data Analysis*, 14(3) :315–322.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recoguition*, 28 :781–793.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixtures posterior distribution. *J. American Statist. Assoc.*, 95(3) :957–979.

Celeux, G., Lavergne, C., and Martin, O. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statist. Model.*, 5 :243–267.

Chan, P., Jacqmin, P., Lavielle, M., McFadyen, L., and Weatherley, B. (2011). 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.

Chanda, S. (1954). A note on the consistency and maxima of the roots of the likelihood equations. *Biometrika*, 41 :56–61.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Royal. Stat. Soc.*, 34(B) :187–220.

Cramer, S. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New York.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization : Population mixtures and stationary arma processes. *Annals of Statistics*, 27(4) :1178–1209.

Dafny, U. G. and Tsiatsis, A. A. (1998). Evaluating Surrogate markers of clinical outcome measured with error. *Biometrics*, 54 :1445–1462.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurements Data*. Chapman & Hall., London.

De la Cruz-Mesia, R., Fernando, A. Q., and Guillermo, M. (2008). Model-based clustering for longitudinal data. *Comput. Statist. and Data Analysis*, 52 :1441–1457.

DeGruttola, V. and Tu, X. M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, 50 :1003–1014.

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

Delattre, M., Savic, R., Miller, R., Karlsson, M., and Lavielle, M. (2012). Analysis of exposure-response of CI-945 in patients with epilepsy : application of novel mixed hidden Markov modeling methodology. *J. Pharmacokinet. Pharmacodyn.*, 39 :263–271.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence Of A Stochastic Approximation Version Of The EM Algorithm. *The Annals Of Statistics*, 27 :94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *JRSS*, 39(B) :1–38.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *J. Classification*, 5 :249–282.

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

Dubois, A., Lavielle, M., Gsteiger, S., Pigeolet, E., and Mentré, F. (2011). Model-Based Analyses of Bioequivalence Crossover Trials Using the SAEM Algorithm. *Statistics in Medicine*, 30 :582–600.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model. Statistics for Biology and Health*. Springer., New York.

Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator : observed versus expected Fisher information. *Biometrica*, 65 :457–487.

Evans, W. E. and Relling, M. V. (1999). Pharmacogenomics : Translating functional genomics into rational therapeutics. *Science*, 286 :487–491.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman & Hall., London.

Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates : A Gibbs sampling approach. *Stat. Med.*, 15 :1663–1685.

Feller, W. (1943). On a General Class of Contagious Distributions. *Annals of Mathematical Statistics*, 14 :389–400.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer, New York.

Gaffney, S. J. and Smith, P. (2003). Curve clustering with random effects regression mixtures. *In : Bishop, C.M., Frey, B.J. (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key-West, FL.*

Ge, Z., Bickel, P. J., and Rice, J. A. (2004). An approximate likelihood approach to nonlinear mixed effects models via spline approximation. *Comput. Statist. Data. Anal.*, 46 :747–776.

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

Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72 :593–599.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with markov chain monte-carlo method for incomplete data estimation problems. *Proc. Natl. Acad. Sci. U. S. A.*, 95 :7270–7274.

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by markov chain monte carlo stochastic approximation. *J. R. Stat. Soc. B*, 63 :339–355.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Biometrika*, 61 :383–385.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1 :465–480.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *(eds). Bayesian Nonparametrics.* Cambridge University Press., Cambridge.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for independent not identically distributed case. *Ann. Math. Stat*, 42(6) :1977–1991.

Hosmer, D. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Comm. Statist.*, 3(10) :995–1010.

Hsieh, F., Tseng, Y., and Wang, J. (2006). Joint modeling of survival and longitudinal data : Likelihood approach revisited. *Biometrics*, 62(1037–1043) :995–1010.

Huang, X. and Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics*, 63 :389–397.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *J. Comput. and Graphical Statist.*, 12(1) :55–79.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer Series in Statistics, New York.

Jones, P. N. and McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Austral. J. Statist*, 34(2) :233–240.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd edition John Wiley, New York.

Karlsson, K. E., Plan, E. L., and Karlsson, M. O. K. (2011). Performance of three estimation methods in repeated time-to-event modeling. *The AAPS Journal*, 13(1) :83–91.

Kelly, P. J. and Jim, L. L. (2000). Survival analysis for recurrent event data : an application to childhood infectious disease. *Statistics in Medicine*, 19(1) :13–33.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis - Techniques for Censored and Truncated Data.* Springer-Verlag, New York.

Kongerud, J. and Samuelsen, S. O. (1991). A longitudinal study of respiratory symptoms in aluminum potroom workers. *American Review of Respiratory Diseases*, 144 :10–16.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with a mcmc procedure. *ESAIM P& S*, 8 :115–131.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49 :1020–1038.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4) :963–974.

Lavielle, M., Samson, A., Fermin, A. K., and Mentré, F. (2011). Maximum likelihood estimation of long term HIV dynamic models and antiviral response. *Biometrics*, 67 :250–259.

Lehmann, E. (1983). *Theory of point estimation.* John Wiley, New York.

Li, B. (2006). A new approach to cluster analysis : the clustering-function-based method. *J. Roy. Statist. Soc. Ser. B*, 68 :457–476.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed-effects models for repeated measures. *Biometrics*, 46 :673–687.

Liu, L., Wolfe, R., and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60 :747–756.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *J. Royal. Statist. Soc.*, 44 :226–233.

Mbogning, C. and Lavielle, M. Inference in mixtures of non linear mixed effects models. *Submitted.*

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.*, 92 :162–170.

McLachland, G. J. and Basford, K. E. (1988). *Mixture models : Inference and Applications to clustering.* Marcel Dekker Inc., New York.

McLachland, G. J. and Peel, D. (2000). *Finite Mixture models.* Wiley-Interscience, New York.

Miller, R. and al. (2003). Exposure response analysis of pregabalin add-on treatment of patients with refractory partial seizures. *Clin Pharmacol Ther*, 73 :491–505.

Nelson, W. B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications.* ASA-SIAM Series on Statistics and Applied Probability, Philadelphia.

Nie, L. (2006). Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63 :123–143.

Nie, L. (2007). 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.

Nie, L. and Yang, M. (2005). Strong Consistency of MLE in Nonlinear Mixed-effects Models with large cluster size. *Sankya : The Indian Journal of Statistics*, 67 :736–763.

Papastamoulis, P. and Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *J. Comput. Graph. Stat.*, 19 :313–331.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *J. Amer. Statist. Assoc.*, 83 :719–726.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 185 :71–110.

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

Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.*, 35 :362–378.

Pfeifer, C. (2004). Classification of longitudinal profiles based on semiparametric regression with mixed effects. *Statist. Model.*, 4 :314–323.

Pinheiro, J. and Bates, D. (1995). Approximation to the log-likelihood function in the non-linear mixed-effect models. *J. Comput. Graph. Statist.*, 4 :12–35.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69 :339–342.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *J. Amer. Statist. Assoc.*, 57 :306–310.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.*, 73 :730–738.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Racine-Poon, A. (1985). A Bayesian approach to nonlinear random effects models . *Biometrics*, 41 :1015–1023.

Redner, R. A. (1981). Note on the consistency of the maximum-likelihood estimate for non-indentifiable distributions. *Ann. Statist.*, 9 :225–228.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26 :195–239.

Rizopoulos, D. (2010). JM : An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9) :1–33.

Rizopoulos, D. (2012a). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics and Data Analysis*, 56(3) :491–501.

Rizopoulos, D. (2012b). *Joint Models for Longitudinal and Time-to-Event Data : With Applications in R.* Chapman & Hall/CRC Biostatistics, Boca Raton.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *J. Roy. Stat. Soc. B*, 71 :637–654.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22 :400–407.

Robert, C. (1996). *Methode de Monte Carlo par chaînes de Markov.* Economica, Paris.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16 :351–367.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistician Association*, 92 :894–902.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation : application on cancer events. *Biostatistics*, 8 :708–721.

Rossi, P., Allenby, G., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons, New York.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model : Application to HIV dynamics model. *Comput. Statist. Data Anal.*, 51 :1562–1574.

Samson, A., Lavielle, M., and Mentré, F. (2007). The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model. *Statistics in Medicine*, 26 :4860–4875.

Samuelsen, S. O. and Kongerud, J. (1993). Evaluation of applying interval censoring on longitudinal data on asthmatic symptoms. Statistical Research Report 2, Institute of Mathematics, University of Oslo.

Savic, R. and Lavielle, M. (2009). A new SAEM algorithm : Performance in Population Models for Count Data. *Journal of Pharmacokinetics and Pharmacodynamics*, 36 :367–379.

Savic, R., Mentré, F., and Lavielle, M. (2011). Implementation and evaluation of an SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacometrics. *The AAPS Journal*, 13(1) :44–53.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statist. Medicine*, 11 :1861–1870.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology : Methodological Issues*. Birkhaüser, Boston.

Shaked, M. (1980). On mixtures from exponential families. *J. R. Statist. Soc. B*, 42 :192–198.

Sheiner, L. B. and Beal, S. L. (1985). Pharmacokinetic parameter estimates from several least squares procedures : superiority of extended least squares. *J. Pharmacokinet. Biop.*, 13 :185–201.

Sheiner, L. B., Rosenberg, B., and Melmon, K. L. (1972). Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res.*, 5 :441–459.

Snoeck, E., Chan, P., Lavielle, M., Jacqmin, P., Jonsson, N., Jorga, K., Goggin, T., Jumbe, S., and Frey, N. (2010). Hepatitis C Viral Dynamics Explaining Breakthrough, Relapse or Response after Chronic Treatment. *Clinical Pharmacology and Therapeutics*, 87(6) :706–713.

Tanner, M. (1996). *Tools for statistical inference, methods for the exploration of*

posterior distributions and likelihood functions., volume 34. Springer-Verlag, New York.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34 :1265–1269.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, 22 :1701–1762.

Titterington, D. M. and Smith, A. F. M. (1985). *Finite Mixture Distributions.* John Wiley & Sons., Chichester.

Tsiatis, A. A. and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data : An overview. *Statistica Sinica*, 14 :809–834.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error : applications to survival and CD4 counts in patients with AIDS. *J. Amer. Statist. Assoc.*, 90 :27–37.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Ann. Statist.*, 27 :439–460.

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

Vonesh, E. G. and Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurments.* Marcel Dekker, New York.

Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models. *J. Am. Stat. Assoc.*, 91 :62–75.

Wakefield, J., Smith, A., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models by using the gibbs sampler. *J. Roy. Stat. Soc., Ser. C, Appl. Stat.*, 43 :201–221.

Wakefield, J. C., Aaron, L., and Racine-Poon, A. (1998). *The Bayesian approach to population pharmacokinetic/pharmacodynamic modelling. In case studies in Bayesian Statistics.* Springer, New York.

Wald, A. (1949). Note on the consistency of the maximum-likelihood estimate. *Ann. Math. Statist.*, 20 :595–600.

Walker, S. (1996). An EM algorithm for non-linear random effects models. *Biometrics*, 52 :934–944.

Wang, X., Schumitzky, A., and D'Argenio, D. Z. (2007). Non linear random effects mixture models : Maximum likelihood estimation via the EM algorithm. *Comput. Stat. Data Anal.*, 51 :6614–6623.

Wang, X., Schumitzky, A., and D'Argenio, D. Z. (2009). Population pharmacokinetic/pharmacodynamic mixture models via maximum a posteriori estimation. *Comput. Stat. Data Anal.*, 53 :3907–3915.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistician Association*, 85 :699–704.

Wolfinger, R. (1993). Laplace approximation for non-linear mixed effects models. *Biometrica*, 80 :791–795.

Woodbury, M. (1971). Discussion of paper by Hartley and Hocking. *Biometrics*, 27 :808–817.

Wu, C. (1983). On the convergence property of the EM algorithm. *Ann. Statist.*, 11 :95–103.

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *J. Am. Stat. Assoc.*, 99 :700–709.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53 :330–339.

Yakowitz, S. and Spragins, J. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39 :209–214.

Yao, W. (2012). Model based labeling for mixture models. *Stat Comput*, 22 :337–347.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Am. Stat. Assoc*, 104 :758–767.