UNIVERSITÉ PARIS DESCARTES

Document de synthèse

présenté par

Adeline LECLERCQ SAMSON

en vue de l'obtention de

l'Habilitation à Diriger des Recherches

Spécialité : Mathématiques Appliquées

Contribution à l'étude des modèles mixtes et des modèles stochastiques en vue d'applications en biologie

---

Soutenue le 21 novembre 2012

devant le jury :

Fabienne Comte	Professeur (Université Paris Descartes)
Jean-Francois Dupuy	Professeur (INSA de Rennes), *rapporteur*
Reinhard Hoepfner	Professeur (Johannes-Gutenberg Universität), *rapporteur*
Marc Hoffmann	Professeur (Université Paris Dauphine)
Marc Lavielle	Directeur de recherche (INRIA Saclay)
Jean-Michel Marin	Professeur (Université Montpellier 2), *rapporteur*
France Mentré	PU-PH (Université Paris Diderot)
Jean-Christophe Thalabard	PU-PH (Université Paris Descartes)

Table des matières

Remerciements

Je remercie sincèrement Jean-Francois Dupuy, Reinhard Hoepfner et Jean-Michel Marin pour leurs rapports et l'intérêt qu'ils ont porté à mon travail. C'est un honneur pour moi. Merci également pour le temps que vous consacrez à mon jury.

Je remercie Marc Hoffmann d'avoir accepté de participer à mon jury et pour nos discussions informelles toujours appréciées.

Je remercie particulièrement Marc Lavielle et France Mentré, qui m'ont initiée à la recherche lors de ma thèse. Vous avez éveillé chez moi une curiosité pour les problématiques statistiques du domaine biomédical et m'avez montré la richesse des collaborations entre statisticiens et médecins. Cette curiosité est restée le fil conducteur de mes travaux de recherche. Ce manuscrit vous doit beaucoup. Merci.

Je remercie Jean-Christophe Thalabard de participer à mon jury et de m'avoir fait confiance dès mon arrivée au MAP5 en me proposant de co-encadrer le stage de M2 de Julien Stirnemann, puis sa thèse. Merci également de m'avoir invitée à participer au programme STAFAV, au travers des cours au Sénégal, et de la collaboration avec Solange Whegang. Travailler avec toi est très enrichissant tant du point de vue scientifique qu'humain.

Je remercie très sincèrement Fabienne Comte pour sa participation à mon jury, en particulier, et sa très grande disponibilité, en général. C'est un grand plaisir de partager un bureau avec toi. Tu es pour beaucoup dans mes trajets incessants entre l'IUT et le MAP5 et ma motivation à revenir dès que je peux au laboratoire. Notre collaboration continuera je l'espère encore longtemps.. Merci aussi (et surtout) pour ton soutien sans faille dans les moments moins faciles de la vie.

Je remercie Valentine Genon-Catalot pour sa disponibilité constante, et pour m'avoir, elle aussi, fait confiance dès mon arrivée au MAP5, en me proposant une collaboration. J'ai beaucoup appris grâce à toi et pris plaisir à travailler ensemble. Merci également de tes conseils sur la rédaction de ce manuscrit et de ta relecture minutieuse.

Parmi mes collaborateurs, Sophie Donnet a une place toute particulière. Merci pour toutes ces années de travail ensemble, pour tes idées sans cesse renouvelées, ton enthousiasme même quand les bugs nous résistent pendant des jours, des semaines ou.... J'espère que nous continuerons encore longtemps à travailler ensemble, quelque soit le pays où nous serons l'une et l'autre !

Merci à Antoine Chambaz, Jérôme Dedecker, Xaviere Panhard et Marie-Luce Taupin, pour nos collaborations et pour leur amitié, leur soutien indéfectible et leurs conseils, en toutes circonstances.

Merci à Julien Stirnemann, pour son dynamisme qui ne fléchit jamais, sa curiosité statistique insatiable, ses questions exigeantes et renouvelées tant que je ne l'ai pas convaincu !

Merci à Susanne Ditlevsen, pour notre collaboration fructueuse, son accueil lors de mes séjours à Copenhague, sa bonne humeur et sa gentillesse.

Je remercie aussi tous mes autres collaborateurs, pour les nombreux échanges passionnants et enrichissants : Thierry Bastogne, Gabriel Baron, Emmanuelle Comets, Charles-André Cuenod, Maud Delattre, Christophe Denis, Benjamin Favetto, Jean-Louis Foulley, Sylvie Retout, Frédéric Richard, Yves Rozenholc, Michèle Thieullen, Solange Whegang.

Je remercie tous les membres du laboratoire MAP5, qui m'ont offert des conditions de travail très agréables. Je remercie particulièrement Sylvain Durand, pour avoir coordonné ma candidature à l'habilitation ; Christine Graffigne et Annie Raoult, directrices successives du laboratoire ; Marie-Hélène Gbaguidi, notre gestionnaire si disponible ; Azedine Mani, Arnaud Meunier et Thierry Raedersdorff, du

service informatique ; mes collègues statisticiens Avner, Chantal, Christophe, Flora, Grégory, Hector, Jean-Claude, Olivier, Servane et Rachid, ainsi qu'Hermine Biermé, ma collègue de bureau.

Je tiens à remercier particulièrement mes collègues du département STID de l'IUT. Je tiens profondément à notre souci incessant de l'intérêt des étudiants et de la relation avec les professionnels. Merci à Jean-Michel Poggi, de m'avoir fait, dès ma première année à l'IUT, l'honneur de sa confiance, en m'ayant donné la responsabilité de la licence pro Santé. Je te remercie aussi de ta très grande disponibilité et de tes précieux conseils. Merci à Florence Muri, pour m'avoir accueillie si généreusement depuis mon année d'ATER, pour tous ses conseils pédagogiques et autres, prodigués au fil des années et pour son amitié. Merci à Florence et Philippe Chabault pour leur collaboration dynamique au travers de la licence pro ainsi qu'à Noël Lucas et Jérôme Paget, pour leur confiance. Merci à Elisabeth, François, FX, Guillaume, Jérôme, Marc, Michel, Mohamed, Mourad, Olivier, Serge, Servane, Thomas et Yves. Un merci spécial à Clarisse et Anna-Bella, pour leur gentillesse et les services rendus.

A Gwen, mon moteur pour avancer dans mes projets, et Merlin, pour sa joie de vivre.

Curriculum Vitæ

Adeline LECLERCQ SAMSON

Née le 10/12/1979 à La Roche-Sur-Yon (85)

Situation Professionnelle

Fonction : Maître de conférences à l'IUT de l'Université Paris Descartes
Adresses Professionnelles :
Université Paris Descartes, Laboratoire MAP5, UMR CNRS 8145, 45 rue des Saints-Pères, 75006 PARIS
Téléphone : (+33) 1 83 94 58 77, Télécopie : (+33) 1 42 82 41 44
et
IUT Paris Descartes, département STID, 143 avenue de Versailles, 75016 PARIS
Téléphone : (+33) 1 42 86 48 28
E-mail : adeline.samson@parisdescartes.fr
Web : `http://www.adeline.e-samson.org`

Postes occupés

Depuis 09/2007	Maître de conférences à l'IUT Paris Descartes (laboratoire MAP5 CNRS 8145)
2006-2007	A.T.E.R. (96h) à l'IUT Paris Descartes (laboratoire MAP5 CNRS 8145)
2003-2006	Allocataire de recherche de l'Université Paris 6
2003-2006	Monitrice à l'École Nationale Supérieure d'Arts et Métiers, Paris

Formation

2003-2006	Doctorat de Biostatistiques : thèse effectuée sous la direction de Marc Lavielle et France Mentré soutenue à l'Université Paris 6 en mai 2006 **Titre** : Estimation dans les modèles non-linéaires à effets mixtes : extensions de l'algorithme SAEM pour l'analyse de la dynamique virale sous traitement anti-VIH **Rapporteurs** : Jean-Marc Azais et Daniel Commenges
2003-2004	DEA Probabilité et applications, option statistiques, Université Paris 6, *mention bien*
1999-2003	Magistère de Mathématiques, Université Louis Pasteur, Strasbourg, *mention très bien*
2002-2003	DEA Santé Publique, option biostatistiques, Université Paris 11, *mention très bien*
2000-2001	Agrégation de mathématiques, option probabilités et statistiques, *rang 107*
2000-2001	Maîtrise de mathématiques, Université Louis Pasteur, Strasbourg, *mention bien*
1999-2000	Licence de mathématiques, Université Louis Pasteur, Strasbourg, *mention bien*

Domaines de recherche

Mes domaines de recherche concernent à la fois le développement théorique et algorithmique de méthodes statistiques, le développement de modèles stochastiques pour des applications biologiques ou médicales et enfin l'utilisation de ces méthodes pour la résolution de problèmes biomédicaux.

Les thèmes théoriques au coeur de ma recherche sont les modèles mixtes, avec une approche par estimation paramétrique ou non paramétrique. Dans le cadre paramétrique, je propose des estimateurs du maximum de vraisemblance, des estimateurs par minimum de contraste, des estimateurs fondés sur des pseudo-vraisemblances, ou des estimateurs bayesiens. Dans le cadre non paramétrique, je propose des estimateurs de déconvolution. J'étudie également l'estimation paramétrique d'équations différentielles stochastiques, en considérant des diffusions observées à temps discrets avec ou sans bruit, des diffusions hypoelliptiques et des diffusions multidimensionnelles partiellement observées.

Les outils algorithmiques que j'utilise sont des dérivés stochastiques de l'algorithme Expectation-Maximization, des algorithmes Monte Carlo par Chaine de Markov (MCMC) et des algorithmes de filtrage particulaire.

Enfin, les problématiques biomédicales auxquelles je m'intéresse sont variées : pharmacocinétique et pharmocodynamie, dynamique du VIH, croissance tumorale, prédiction de croissance foetale, activité neuronale... Elles ont en commun d'être basées sur la modélisation de processus biologiques observés au cours du temps. Mes travaux consistent à proposer des modèles stochastiques modélisant ces processus, modèles dont les paramètres ont une interprétation biologique ou physiologique, et que l'on cherche à estimer.

Liste de publications

Dans cette liste, je présente mes articles publiés et sous presse dans des revues internationales avec comité de lecture, un chapitre d'une lecture notes, des actes de conférences qui ont fait l'objet d'une relecture par un comité de lecture, une discussion d'article et enfin mes articles soumis. L'ensemble de ces articles sont disponibles sur ma page internet, en particulier sous forme de prépublications pour les articles soumis.

J'ai choisi de présenter la liste de mes articles par ordre chronologique, sans faire de distinction entre les différents types de publication (méthodologie statistique théorique et algorithmique, applications de méthodes statistiques à des problèmes biologiques ou analyse de données réelles biologiques).

Articles publiés ou à paraître dans des revues internationales avec comité de lecture

1. Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model : application to HIV dynamics model. Samson A, Lavielle M, Mentré F. Computational Statistics and Data Analysis, 51(3) :1562-74, 2006.
2. Estimation of parameters in incomplete data models defined by dynamical systems. Donnet S, Samson A. Journal of Statistical Planning and Inference, 137(9) :2815-31, 2007.
3. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. Samson A, Lavielle M, Mentré F. Statistics in Medicine, 26(27) :4860-4875, 2007.
4. Design in nonlinear mixed effects models : optimization using the Fedorov-Wynn algorithm and power of the Wald test for binary covariates. Retout S, Comets E, Samson A, Mentré F. Statistics in Medicine, 26(28) :5162-5179, 2007.
5. Parametric inference for mixed models defined by stochastic differential equations. Donnet S, Samson A. ESAIM P&S, 12 :196-218, 2008.
6. Missing data in randomized controlled trials of rheumatoid arthritis with radiographic outcomes : a simulation study. Baron G, Ravaud P, Samson A, Giraudeau B. Arthritis Care & Research, 59(1) :25-31, 2008.

7. Extension of the SAEM algorithm for the nonlinear mixed models with two levels of random effects. Panhard X, Samson A. Biostatistics, 10 :121-35, 2009.
8. A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. Richard F, Samson A, Cuenod CA. Statistics and Computing, 19 :465-478, 2009.
9. Phenomenological modeling of tumor diameter growth based on a mixed effects model. Bastogne T, Samson A, Vallois P, Wantz-Mézières S, Pinel S, Bechet D, Barberi-Heyob M. Journal of Theoretical Biology, 262 :544-552, 2010.
10. Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. Donnet S, Foulley JL, Samson A. Biometrics, 66(3) :733–741, 2010.
11. Parameter estimation for a bidimensional partially observed Ornstein-Uhlenbeck process with biological application. Favetto B, Samson A. Scandinavian Journal of Statistics, 37(2) :200-220, 2010.
12. Maximum likelihood estimation of long term HIV dynamic models and antiviral response. Lavielle M, Samson A, Fermin AK, Mentre F. Biometrics, 67(1) :250–259, 2011.
13. Parameter estimation and change-point detection from Dynamic Contrast Enhanced MRI data using stochastic differential equations. Cuenod CA, Favetto B, Genon-Catalot V, Rozenholc Y, Samson A, Mathematical Biosciences, 233(1) :68–76, 2011.
14. Individual predictions based on population nonlinear mixed modeling : application to prenatal twin growth. Stirnemann J, Samson A, Thalabard JC, Statistics in Medicine, 2012, to appear.
15. Density estimation of a biomedical variable subject to measurement error using an auxiliary set of replicate observations. Stirnemann J, Comte F, Samson A, Statistics in Medicine, 2012, to appear.
16. Multiple Treatment Comparisons (MTC) in a series of antimalarial trials with an ordinal primary outcome and repeated measurements. Whegang S, Samson A, Basco LK, Thalabard JC. Malaria Journal, 2012, 11(1) :147.
17. Contrast estimator for completely or partially observed hypoelliptic diffusion. Samson A, Thieullen M. Stochastic Processes and Their Applications, 2012, to appear.

Lecture notes

1. Introduction to stochastic models in biology. Ditlevsen S, Samson A. In Bachar, Batzel and Ditlevsen (Eds.), Stochastic Methods and Neuron Modeling. Springer. 2012

Actes de conférences (Proceedings)

1. System identification of tumor growth described by a mixed effects model. Bastogne T, Samson A, Mézières-Wantz S, Vallois P, Pinel S, Barberi-Heyob M. Proceedings of IFAC Symposium on system identification, 2009.
2. Metropolis-Hasting techniques for finite element-based registration. Richard F, Samson A. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

Discussion d'articles

- Discussion on "Parameter estimation for differential equations : a generalized smoothing approach" (by Ramsay JO, Hooker G, Campbell D and Cao J). Donnet S, Samson A. Journal of the Royal Statistical Society : Series B, 69(5) :741-796, 2007.

Articles soumis

1. Nonparametric estimation of random effects densities in linear mixed-effects model. Comte F, Samson A, soumis. Prepublication MAP5 2012-01.
2. Maximum likelihood estimation for stochastic differential equations with random effects. Delattre M, Genon-Catalot V, Samson A, soumis. Prepublication MAP5 2011-31.

3. Estimation in autoregressive model with measurement error. Dedecker J, Samson A, Taupin ML, soumis. Prepublication MAP5 2011-18.
4. Deconvolution estimation of onset of pregnancy with replicate observations. Comte F, Samson A, Stirnemann J, soumis. Prepublication MAP5 2011-15.
5. EM algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. Donnet S, Samson A. Prepublication MAP5 2010-24.
6. Ditlevsen, S. Samson, A. Parameter estimation in the stochastic neuronal Morris-Lecar model with particle filter methods. soumis.

Communications

Je présente ici la liste des communications que j'ai données dans des congrès internationaux, des congrès nationaux ou des groupes de travail. Certaines communications font suite à des invitations et sont indiquées comme telles.

Communications orales dans des congrès internationaux

- Parametric estimation of a partially observed two-dimensional stochastic differential equation. Application to neuronal data analysis, *BIO-SI workshop on biostatistics, Rennes, 2011*
- Parameter estimation of a two-dimensional stochastic differential equation partially observed with application to neuronal data analysis, *Statistics and Modeling for Complex Data, Université Marne La Vallée, 2011* (Invitée)
- Minimum contrast estimate for the parameters of the stochastic Morris-Lecar model, *Stochastic models in neurosciences, CIRM, France, 2010*
- Bayesian analysis of growth curves using mixed models defined by stochastic differential equations, *DYNSTOCH meeting, Berlin, Germany, 2009.*
- Parameter estimation of a bidimensional partially observed Ornstein-Uhlenbeck process, *Summer School and Workshop, Middelfart, Denmark, 2008.* (Invitée)
- Estimation of microcirculation parameters, *International Biomedical Modeling School and Workshop, Bangalore, India, 2008.*
- Maximum Likelihood estimation in non-linear mixed models via the SAEM-MCMC algorithm, *New directions in Monte Carlo methods workshop, Fleurance, France, 2007.*
- Estimation in mixed models with left-censored data and differential systems, *Rencontres Espagne - France - Venezuela de Probabilité et Statistiques Mathématiques, Choroni, Venezuela, 2006.*
- The SAEM algorithm for non-linear mixed models with left-censored data and differential systems : application to the joint modeling of HIV viral load and CD4 dynamics under treatment, *Sheiner Student Session, 15th Meeting of the Population Approach Group in Europe, Brugge, Belgium, 2006.*
- Stochastic Approximation EM algorithm in nonlinear mixed effects models : an evaluation by simulation, *13th Meeting of the Population Approach Group in Europe, Upssala, Sweden, 2004.*

Communications orales dans des congrès nationaux

- Modèles mixtes définis par équations différentielles stochastiques. Estimation par l'algorithme SAEM et filtre particulaire, *43èmes journées de la SfdS, Tunis, France, 2011.*
- Parameter estimation for a bidimensional partially observed Ornstein-Uhlenbeck process and application to anti-cancer therapy, *Seminaire Européen de Biostatistiques, Paris, France, 2008.* (Invitée)
- Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model : application to HIV dynamics model, *39èmes journées de la SfdS, Angers, France, 2007.* (Invitée)
- Estimation par l'algorithme SAEM des paramètres de modèles non linéaires mixtes : application à la modélisation de la dynamique virale VIH, *Journée Jeunes Chercheurs de la Société Française de Biométrie, Villejuif, 2006.*

- Estimation des paramètres d'un modèle non-linéaire mixte de la dynamique virale VIH par l'algorithme SAEM, *JOBIM, Bordeaux, 2006.* (Invitée)
- Estimation paramétrique d'un processus de diffusion à partir d'observations bruitées et à temps discrets, *Colloque "Jeunes probabilistes et statisticiens", Aussois, 2006.*
- Estimation paramétrique dans des modèles définis par un système d'équations différentielles ordinaires, *2ème congrès de la Société Mathématiques Appliquées et Industrielles, Evian, France, 2005.*
- Estimation paramétrique dans des modèles définis par un système d'équations différentielles ordinaires, *37èmes journées de la SfdS, Pau, France, 2005.*

Séminaires et groupes de travail

- Séminaire statistique, Grenoble, janvier 2012.
- Groupe de travail de Probabilité, Université Paris 5, juin 2011.
- Séminaire statistique, Montpellier, mai 2011.
- Groupe de travail de statistique, Strasbourg, mai 2011.
- Groupe de travail de biostatistique, Villejuif-Université Paris 6, avril 2011.
- Statistics Seminar, Copenhagen, Denmark, février 2011.
- Séminaire statistique/biostatistique, Nancy, février 2011.
- Groupe de travail de Statistique, Université Paris 5, octobre 2010.
- Groupe de travail bigMC, octobre 2010.
- Séminaire parisien de statistiques, novembre 2009.
- Séminaire de l'AgroParisTech, octobre 2009.
- Séminaire de Statistiques et Santé Publique de Bordeaux, avril 2009.
- Séminaire de Probabilité et statistiques, Université de Nice, janvier 2008.
- Séminaire de Statistiques, Université Paul Sabatier, Toulouse, janvier 2008.
- Colloquium du MAP5, Université Paris Descartes, octobre 2007.
- Séminaire de Statistiques, Université de Strasbourg, mars 2007.
- Groupe de travail Statistiques et Biologie, Université de Nancy, mars 2007.
- Séminaire de l'équipe Select, INRIA, Université Paris 11, février 2007.
- Séminaire Mathématique et Génome, Université d'Evry, INRA, INA-PG, janvier 2007.
- Séminaire Extraction de Connaissances : Approches Informatique et Statistique, janvier 2007.
- Groupe de travail de Statistique, Université Paris 5, décembre 2006.
- Groupe de travail des thésards et jeunes docteurs, Université Paris 5, décembre 2006.
- Groupe de travail Monolix, Université Paris 11, INSERM, novembre 2006.
- Groupe de travail Monolix, Université Paris 11, INSERM, novembre 2005.
- Groupe de travail de Statistique, Universités Paris 6 et 7, octobre 2005.
- Groupe de travail Monolix, Université Paris 11, INSERM, mars 2005.
- Groupe de travail Monolix, Université Paris 11, INSERM, janvier 2004.

Séjours à l'étranger et collaborations internationales

Depuis 2008	**Susanne Ditlevsen**, *University of Copenhagen, Danemark* Estimation pour les équations différentielles stochastiques et leurs applications en biologie Invitations au département de Mathématiques de Copenhague (un mois en 2011)
Depuis 2008	**Andrea De Gaetano**, *BioMathLab, Rome, Italie* Modélisation de la relation glucose/insuline chez des patients diabétiques Invitations au laboratoire BioMathLab (Italie) (une semaine en 2009)

ANR et autres

Depuis 2011	Responsable du Projet Collaboratif Inter-site, financé par l'université Paris Descartes "Modélisation et analyse statistique de données neuronales". Ce projet réunit des neurophysiologistes (Lee Moore, laboratoire CESEM ; Christophe Pouzat, laboratoire de physiologie cérébrale ; Université Paris Descartes) et des statisticiens (Antoine Chambaz, Fabienne Comte, Jérôme Dedecker, laboratoire MAP5, Université Paris Descartes ; Marie-Luce Taupin, laboratoire statistique et génome, Université d'Evry). Les thématiques abordées sont les problèmes d'estimation paramétrique ou non paramétrique de modèles neurophysiologiques. Les modèles mathématiques sous-jacents sont des modèles de Markov cachés, des modèles auto-régressifs observés avec bruit, des modèles d'équations différentielles stochastiques et des modèles d'intensité conditionnelle.
Depuis 2009	Membre du groupe DYNSTOCH (ex réseau européen) sur les méthodes statistiques pour les modèles dynamiques stochastiques
2008-2009	Membre du BQR Paris Descartes "Modélisation de l'infection par le virus de l'hépatite C", dirigé par Jean-Stéphane Dhersin (Laboratoire MAP5, Université Paris Descartes)
2007-2008	Membre du BQR Paris Descartes "Modélisation de l'angiogénèse", dirigé par Yves Rozenholc (Laboratoire MAP5, Université Paris Descartes)
2007-2008	Membre du Projet CNRS "Modélisation de l'activité neuronale", dirigé par Lee Moore (Laboratoire CESEM, Université Paris Descartes)
2004-2007	Membre de l'ANR Monolix, dirigée par Marc Lavielle (INRIA Orsay), orientée sur les modèles non-linéaires à effets mixtes

Encadrement

Encadrement de thèses

Depuis 2009	Co-encadrement de la thèse de Christophe Denis (ED 386, Laboratoire MAP5, Université Paris Descartes) avec Antoine Chambaz (Laboratoire MAP5, Université Paris Descartes) **Titre** : Méthodes statistiques pour la classification de données de maintien postural
Depuis 2009	Co-encadrement de la thèse de Julien Stirnemann (ED 420, Laboratoire MAP5, département d'obstétrique, hôpital Necker-enfants malades, Université Paris Descartes) avec Jean-Christophe Thalabard (Laboratoire MAP5, Centre de diagnostic, Hotel-Dieu, Université Paris Descartes) **Titre** : Prédiction de courbes de croissance observées avec bruit

Mémoire et stages

2009-2010	Co-encadrement du stage de Master 1 Professionnel Ingénierie Mathématiques, Université Paris 11 de Wei Huang, avec Sophie Donnet (Laboratoire Ceremade, Université Paris Dauphine) **Titre** : Estimation des paramètres de l'activité neuronale
2008-2009	Co-encadrement du stage de Master 2 Recherche Probabilités et Statistiques, Université Paris 11 de Christophe Denis, avec Jean-Stéphane Dhersin (Laboratoire MAP5, Université Paris Descartes) **Titre** : Estimation de la prévalence du VHC chez les usagers de drogue

2007-2008	Co-encadrement du stage de Master 2 Recherche Santé Publique, option Biostatistiques, Université Paris 11 de Julien Stirnemann, avec Jean-Christophe Thalabard (Laboratoire MAP5, Université Paris Descartes) **Titre** : Construction de courbes de croissance : comparaison des méthodes paramétriques et non-paramétriques existantes. Cas particulier des individus corrélés et application au poids néonatal des grossesses gémellaires nées dans les Yvelines (2002-2005)
2006-2007	Encadrement du TIPE de 3 étudiants de classe préparatoire MP* du lycée Henri IV, Paris **Titre** : Modélisation de la dynamique du VIH par systèmes différentiels ordinaires

Membre de comités de thèse

2012	Membre du comité de thèse de Bogdan Mirauta, Université Paris 6 ; encadré par Hughes Richard (Université Paris 6) et Pierre Nicolas (INRA) **Titre** : A Sequential Monte Carlo Method for Estimating Transcriptional Landscape at Basepair Level from RNA-Seq Data
2011	Membre du comité de thèse de Marius Kwemou, Université d'Evry, encadré par Marie-Luce Taupin (Université d'Evry) et Abdou Diongue (Université Gaston Berger, Sénégal) **Titre** : Réduction de dimension en régression logistique et application aux données ACTU-PALU

Membre de jurys de thèses

2011	Membre du jury de thèse de Meïli Bagaratti, Université de la Méditerranée, Marseille
2010	Membre du jury de thèse de Benjamin Favetto, Université Paris Descartes
2009	Membre du jury de thèse de Julie Antic, Ecole Nationale Vétérinaire, Toulouse
2008	Membre du jury de thèse de Mylène Duval, AgroParisTech

Animation de la recherche

Rapports de lecture (referee)
Scandinavian Journal of Statistics, Computational Statistics and Data Analysis, Statistics in Medicine, BMC Bioinformatics, Journal of Pharmacokinetic and Pharmacodynamic, Applied Mathematical Modelling, Journal de la SFdS, Statistics in Biosciences

Colloques, Séminaires, Groupe de travail

Depuis 2010	Co-organisation d'une séance par an du séminaire parisien de statistiques
2010	Co-organisation du colloque Jeunes Probabilistes et Statisticiens, Mont Dore
2008	Co-organisation d'une journée "Statistiques et déformations pour l'imagerie médicale", Paris
Depuis 2006	Membre du bureau du groupe de travail "Applications Bayésiennes Utilisant le Gibbs Sampling qui organise deux journées annuelles, Paris

Responsabilités administratives

Je présente ici mes responsabilités dans le domaine de la recherche et dans le cadre de mes enseignements.

Recherche

Depuis 2011	Représentante de la SMAI au conseil de la SFdS
Depuis 2010	Membre de la commission Communication de la SFdS
Depuis 2010	Membre élue du groupe MAS de la SMAI
Depuis 2010	Membre élue du Conseil du laboratoire MAP5 UMR CNRS 8145
2010-2011	Membre élue du Conseil de département STID à l'IUT Paris Descartes
2011	Membre du comités de sélection MCF statistiques, Toulouse
2010	Membre des comités de sélection MCF statistiques, Université Paris 5, Université Dauphine, AgroParisTech
2009	Membre des comités de sélection MCF statistiques, Université Paris 6, Université Grenoble 2

Enseignement

Depuis 2007	Responsable de la Licence Professionnelle Santé : Statistiques et Informatique Décisionnelle pour la santé, IUT Université Paris Descartes : Je suis en charge de recruter les étudiants à partir de dossiers de candidature et d'entretiens de motivation individuels (environ 100 candidats). Je m'occupe de la constitution de l'équipe pédagogique, qui comprend 35 intervenants dont 25 sont des professionnels du secteur. Je m'occupe de l'évolution des programmes, des contacts avec les partenaires professionnels et de l'emploi du temps. Pendant l'année universitaire, je suis également en charge du suivi personnalisé des étudiants, du suivi des stages et de la promotion de la formation auprès de différents publics. Je conseille les professionnels qui souhaitent faire valider ce diplôme par une procédure de validation des acquis de l'expérience (VAE) en les aidant à rédiger leur dossier de VAE et en organisant le jury. Enfin, je suis responsable de la rédaction du dossier de renouvellement dans le cadre des contrats quinquennaux d'évaluation par l'AERES.
2008-2010	Membre du jury du concours de l'agrégation externe de Mathématiques, oral modélisation
2008	Membre du jury du concours de Technicien en Informatique Médicale, Hôpitaux de Chartres

Enseignement

Depuis 2007	Maitre de conférences à l'IUT Paris Descartes Cours : estimation et test, modèle linéaire, modèle de survie, modèle logistique, série chronologique TD : statistique descriptive, estimation et tests, modèle linéaire, modèle de survie TP (R et SAS) : statistique descriptive, estimation et tests, modèle linéaire, modèles de survie
2006-2007	ATER à l'IUT Paris Descartes Cours-TD : modèle linéaire, statistiques descriptives Encadrement d'un projet de première année
2003-2006	Monitrice à l'ENSAM Paris : TD de mathématiques pour l'ingénieur

Activité de vulgarisation

- Participation à l'organisation du 1er Forum Emploi Mathématiques, 26 janvier 2012, co-organisé par l'AMIES, la SFdS et la SMAI (1000 participants, 35 entreprises présentes)
- Participation à l'élaboration de la brochure ONISEP "les métiers de la statistique", en partenariat avec la SFdS, 2010-2011 (20 000 brochures distribuées)
- Présentation de la statistique (Statistique : des mathématiques aux applications) aux étudiants du magistère de Strasbourg, février 2007
- Présentation des métiers de la statistique et des formations en IUT dans des lycées d'Ile de France (06-07)
- Animation d'un stand de présentation des métiers de la statistique au Forum "Les Métiers Scientifiques" à l'intention des lycéens et étudiants d'Ile de France, Université d'Evry et École Polytechnique (05-06)
- Animation d'un stand de vulgarisation des mathématiques à l'intention des collégiens du Bas-Rhin, Conseil général du Bas-Rhin, Strasbourg (01)

Prime, Prix et Distinctions

- Délégation CNRS au 2eme semestre de l'année 2010-2011.
- Titulaire de la PEDR sur la période 2008-2012.
- Prix du jeune chercheur, New directions in Monte Carlo methods workshop, 2007, Fleurance, France.
- Prix 2007 du docteur Norbert Marx, Société Française de Statistiques.
- Prix 2006 de la "Sheiner Student Session", 15th Meeting of the Population Approach Group in Europe, Brugge, Belgium.

Introduction

Ce document présente une synthèse de mes travaux de recherche. Mon travail de thèse (2003-2006), au sein de l'unité INSERM U736 de l'hôpital Bichat-Claude Bernard, a pour origine l'analyse de données longitudinales de décroissance de charge virale chez des patients atteints par le VIH. Je me suis intéressée à l'estimation des paramètres de modèles mixtes, outil statistique classique d'analyse de ces données. La fonction de régression de ces modèles décrit la dynamique de la charge virale. Différents modèles mathématiques ont été proposés, en particulier des modèles basés sur des systèmes d'équations différentielles ordinaires (EDO), qui n'ont en général pas de solution analytique. Cela complexifie l'estimation des paramètres et peu de méthodes sont adaptées à ce cadre. J'ai poursuivi ce type d'études après ma thèse en élargissant au fur et à mesure les modèles considérés et les domaines d'applications.

Un grand nombre de phénomènes biologiques sont naturellement modélisés par un processus en temps continu. Les modèles mathématiques déterministes (systèmes d'EDO) proposés se sont complexifiés au cours des dernières années pour tenter de modéliser plus en détails les multiples réactions biologiques ou physiologiques en jeu. Cependant, malgré cette complexification, ces modèles théoriques restent et resteront toujours faux. Par ailleurs, il a été montré dans de nombreuses expériences biologiques que certains phénomènes sont imprévisibles. Une expérience réitérée dans des conditions expérimentales strictement identiques aboutit à des observations différentes. La variabilité observée est inhérente au processus biologique sous-jacent et doit être prise en compte. Plusieurs auteurs ont montré que des modèles stochastiques (équation différentielle stochastique, chaîne de Markov, modèle autorégressif) sont plus adaptés à l'analyse de données longitudinales biologiques (Ditlevsen et De Gaetano, 2005; Höpfner, 2007). Cette approche est au coeur de mes travaux de recherche. Je travaille sur l'estimation paramétrique de modèles stochastiques, lorsqu'on dispose d'une seule trajectoire individuelle ou de plusieurs trajectoires individuelles (approche par modèles mixtes). Ces travaux sont décrits dans ce document sous la forme de trois chapitres, chacun portant sur des modèles aléatoires spécifiques.

Le chapitre 1 présente les modèles mixtes tels que je les ai étudiés dans ma thèse et par la suite, et pour lesquels beaucoup de questions restent d'actualité. Je rappelle deux résultats obtenus dans ma thèse. Le premier a été développé en collaboration avec Marc Lavielle (INRIA, Orsay) et France Mentré (INSERM, Hôpital Bichat-Claude Bernard). Nous considérons l'estimation des paramètres de modèles mixtes lorsque les données observées sont censurées par une limite de détection de l'appareil de mesure. Nous proposons un algorithme Stochastic Approximation Expectation Maximization (SAEM) combiné à un algorithme de Monte Carlo par Chaine de Markov (MCMC) avec imputation des données censurées. Cet algorithme fournit de meilleurs estimateurs que les méthodes habituellement utilisées. De plus nous prouvons qu'il converge vers le maximum de vraisemblance. Le deuxième résultat, réalisé avec Sophie Donnet (CEREMADE, Université Paris Dauphine), traite du problème d'une fonction de régression définie comme solution d'une EDO mais sans solution analytique. Nous proposons un algorithme SAEM-MCMC utilisant un schéma d'approximation numérique de l'EDO. Nous bornons l'erreur induite par l'utilisation de ce schéma numérique sur l'estimateur obtenu. Cette borne n'avait jamais été étudiée auparavant. Ensuite, je présente un résultat obtenu avec Fabienne Comte (MAP5, Université Paris Descartes) qui permet de s'affranchir de l'hypothèse de normalité des paramètres aléatoires individuels. Cette hypothèse, classique, peut s'avérer totalement irréaliste. Nous proposons une estimation non paramétrique de la densité des paramètres aléatoires basée sur la déconvolution,

sous différentes hypothèses sur la densité des erreurs résiduelles du modèle mixte. Nous proposons une méthode de sélection de l'estimateur, qui réalise le compromis biais/variance. Cette approche de déconvolution est innovante dans le cadre des modèles mixtes. Le cas particulier de l'utilisation d'un deuxième échantillon constitué de données répétées à un temps donné est considéré avec Fabienne Comte et Julien Stirnemann (Département d'obstétrique, Hôpital Necker). Nous étudions en particulier le risque de l'estimateur en fonction de la taille des deux échantillons. Les comparaisons numériques aux estimateurs précédemment proposés dans la littérature sont favorables à notre estimateur.

J'expose ensuite deux exemples d'application biomédicales parmi différents projets transdisciplinaires auxquels j'ai participé. Le premier, réalisé en collaboration avec Marc Lavielle, France Mentré et Ana Karina Fermin (MODAL'X, Université de Nanterre), est une analyse de données longitudinales recueillies lors d'un essai mené par l'ANRS (Agence Nationale de Recherche sur le Sida) qui compare l'efficacité de trois traitements anti-rétroviraux. L'analyse par modèle mixte est complexe car les observations sont bidimensionnelles (charge virale et taux de cellules lymphocytes CD4), la fonction de régression est solution d'une EDO de dimension 4 ou 5 comportant une dizaine de paramètres et les observations sont censurées par une limite de détection de l'appareil de mesure. La méthode d'estimation proposée, basée sur l'algorithme SAEM, montre ici tout son potentiel, en particulier par rapport aux méthodes existantes qui peuvent être limitées en terme de nombre d'effets aléatoires estimables. Dans un deuxième exemple, je présente un projet mené en collaboration avec Julien Stirnemann (MAP5, Département d'obstétrique, Hopital Necker, Université Paris Descartes) et Jean-Christophe Thalabard (MAP5, Université Paris Descartes) portant sur la prédiction de la croissance foetale chez des jumeaux. Dans ce contexte, il existe un niveau supplémentaire de corrélation entre les données des deux jumeaux d'une même grossesse. L'estimation des paramètres d'un tel modèle mixte peut être réalisée par l'algorithme SAEM, ainsi que je l'avais proposé avec Xavière Panhard (INSERM, Hôpital Bichat-Claude Bernard). Nous proposons alors un algorithme de prédiction de croissance individuelle, qui permet de détecter des croissances anormales chez un des deux jumeaux. Alors que la plupart des travaux dans ce domaine sont basées sur des données transversales, l'originalité de ce travail vient de l'analyse de données longitudinales tenant compte des différents niveaux de corrélation. C'est aussi à ma connaissance une des premières réalisations obtenues pour des grossesses gémellaires. Cette approche suppose de connaître avec précision la date de début de grossesse. Nous appliquons la méthode d'estimation non paramétrique développée en collaboration avec Fabienne Comte pour estimer la densité de la variable "début de grossesse" et étudions l'influence de différentes covariables maternelles. A ma connaissance, c'est la première fois qu'un estimateur de déconvolution est utilisé dans ce cadre.

Enfin, j'ai étudié d'autres questions méthodologiques (modèles mixtes ayant un niveau supplémentaire de variabilité, optimalité de protocoles, données manquantes, recalage d'une séquence d'images médicales) et d'autres applications biomédicales (comparaison d'efficacité de traitements anti-cancéreux par modélisation de croissance tumorale, comparaison d'efficacité de traitements anti-paludiques à partir de données catégorielles répétées) mais qui ne sont pas détaillées dans ce manuscrit.

Le chapitre 2 regroupe les travaux que j'ai effectués autour de l'estimation de modèles stochastiques utilisés en biologie lorsqu'on dispose d'une seule trajectoire individuelle. Je me suis intéressée à l'estimation paramétrique d'équations différentielles stochastiques (EDS) observées à temps discrets avec un bruit de mesure, dans le cadre d'un projet collaboratif avec Charles-André Cuenod (Service de Radiologie, Hôpital Européen Georges Pompidou), Benjamin Favetto, Valentine Genon-Catalot et Yves Rozenholc (MAP5, Université Paris Descartes). Les données recueillies sont issues de séquences d'images médicales et mesurent l'évolution au cours du temps de la concentration d'un agent de contraste injecté à un patient. Nous proposons un système différentiel stochastique issu d'un modèle déterministe de pharmacocinétique. Les observations bruitées à temps discrets d'un tel système s'écrivent alors comme un processus ARMA. Nous proposons un estimateur du maximum de vraisemblance, dont on étudie les propriétés sous l'hypothèse de stationnarité du processus. L'application aux données réelles nécessite l'estimation d'un temps de rupture apparaissant dans la fonction de dérive de l'EDS. Nous proposons un estimateur du type moindres carrés de ce temps de rupture, qui a de bonnes propriétés sous des hypothèses raisonnables sur le modèle. L'analyse des données réelles illustre concrètement l'apport du

modèle stochastique sur le modèle déterministe, ce dernier se révélant instable en présence de données "atypiques".

J'expose ensuite différents travaux dont l'objectif commun est la modélisation de données temporelles de l'activité d'un neurone (données de potentiel membranaire) par des modèles stochastiques. Avec Jérôme Dedecker (MAP5, Université Paris Descartes) et Marie-Luce Taupin (Statistique et Génome, Université d'Evry), nous considérons un modèle auto-régressif observé avec bruit. Nous proposons une méthode d'estimation paramétrique de la fonction de régression du processus auto-régressif caché, à partir d'une fonction de contraste de type moindres carrés qui est calculable explicitement via la théorie de Fourier. Nous obtenons la consistance forte et la normalité asymptotique de cet estimateur, sous différentes hypothèses de dépendance du processus auto-régressif et pour une plus large classe de fonctions de régression que les estimateurs précédemment proposés dans la littérature.

Avec Susanne Ditlevsen (Copenhagen University, Denmark) d'une part, et Michèle Thieullen (LPMA, Université Pierre et Marie Curie) d'autre part, nous considérons des systèmes bidimensionnels d'équations différentielles stochastiques dont seulement une composante est observée. L'estimation des paramètres de ces systèmes est complexe car le processus observé n'est plus markovien. Avec Susanne Ditlevsen, nous abordons le problème d'estimation en proposant un algorithme basé sur un filtre particulaire. Nous montrons que l'algorithme converge vers l'estimateur du maximum d'une pseudo-vraisemblance, obtenue en approchant l'EDS par un schéma d'Euler. L'analyse des données neuronales avec cette méthode d'estimation permet de reconstruire le processus non observé d'ouverture et de fermeture des canaux ioniques présents à la surface de la membrane neuronale. C'est à ma connaissance la première fois qu'on est capable d'estimer des paramètres de l'équation cachée de ce système neuronal. Avec Michèle Thieullen, nous considérons le problème plus difficile où l'EDS est hypoelliptique et proposons une fonction de contraste pour estimer ses paramètres. Nous obtenons la consistance forte et la normalité asymptotique de cet estimateur, avec une perte dans la variance asymptotique due à l'observation partielle du système. A ma connaissance, le seul estimateur existant pour ce problème est basé sur un développement de Taylor-Itô du système hypoelliptique mais ses propriétés théoriques ne sont pas établies. De plus, les résultats numériques sont en faveur de notre estimateur, les biais et variances étant plus faibles que ceux obtenus par l'autre approche.

Le chapitre 3 fait le lien entre les deux chapitres précédents. Il est consacré à l'étude de modèles mixtes définis par des équations différentielles stochastiques. J'ai commencé à travailler sur ces modèles à la fin de ma thèse avec Sophie Donnet. A ma connaissance, seules des méthodes basées sur le filtre étendu de Kalman existent mais les propriétés des estimateurs ainsi obtenus ne sont pas démontrées. Nous proposons différentes méthodes d'estimation selon que l'on connaisse ou non la densité de transition de l'EDS. Dans le cadre du paradigme d'estimation bayésienne et lorsque la densité de transition est explicite, nous proposons une méthode basée sur l'algorithme MCMC. Nous montrons l'apport d'une modélisation par EDS mixte par rapport à un modèle mixte régi par une fonction de régression déterministe sur des données réelles de croissance. Lorsque la densité de transition n'est pas explicite, nous proposons une méthode basée sur l'algorithme SAEM-MCMC couplé à un schéma d'Euler d'approximation de l'EDS, ainsi qu'une méthode basée sur un filtre particulaire. Ces deux dernières méthodes d'estimation permettent de construire une suite d'estimateurs qui convergent vers l'estimateur du maximum d'une pseudo-vraisemblance d'un modèle approché par le schéma d'Euler. Nous obtenons une borne sur la distance entre l'estimateur du maximum de la pseudo-vraisemblance et l'estimateur du vrai maximum de vraisemblance (EMV) .

Plus généralement, l'étude des propriétés de l'EMV dans les modèles mixtes définis par EDS est extrêmement complexe. Avec Maud Delattre (Département de Mathématiques, Orsay) et Valentine Genon-Catalot, nous en proposons une première étude, dans le cadre d'EDS linéaires en les paramètres aléatoires. Nous obtenons la consistance forte et la normalité asymptotique de l'EMV. C'est à ma connaissance le premier résultat de consistance obtenu pour l'EMV de modèles mixtes définis par EDS.

Le dernier chapitre rassemble les différentes perspectives de recherche issues de ces travaux.

Chapitre 1

Modèles mixtes

Mes travaux de thèse ont pour origine la modélisation de la décroissance de la charge virale chez des patients VIH sous traitement anti-rétroviral. Les traitements anti-rétroviraux ont pour but de stopper la multiplication du nombre de virus (charge virale) dans le corps, en agissant sur les mécanismes de la reproduction virale. L'évaluation de l'efficacité de ces traitements peut se faire à partir de mesures de la charge virale chez des patients recevant les différents traitements étudiés. Le traitement faisant décroître la charge virale, celle-ci devient rapidement indétectable par les appareils de mesure (inférieure à un seuil de détection). En parallèle de la charge virale, il est habituel de mesurer le taux de cellules lymphocytaires CD4 dans le sang. On dispose alors pour plusieurs patients de mesures à différentes visites de la charge virale et du taux de CD4. Les méthodes les plus classiques pour comparer l'efficacité des traitements sont basées sur des tests de comparaison de pourcentages de patients dont la charge virale est indétectable à la fin de l'essai clinique dans les différents bras de traitement. Ces tests n'utilisent qu'une donnée terminale par patient et pas l'ensemble des données longitudinales disponibles. De plus, les traitements étant de plus en plus efficaces, la proportion d'individus dont la charge virale est indétectable a considérablement augmenté. Il devient alors difficile de montrer la supériorité d'un nouveau traitement par rapport aux traitements de référence par cette approche statistique. Afin d'utiliser toute l'information disponible, une alternative, plus complexe au niveau statistique, est de construire un test de comparaison de traitements à partir de modèles mixtes, approche étudiée dans ce chapitre.

Les modèles mixtes ont été largement employés lorsque la fonction de régression est linéaire. Sous l'hypothèse supplémentaire de normalité des paramètres individuels aléatoires, la fonction de vraisemblance est explicite et l'estimation des paramètres du modèle est relativement simple. Dès que l'on sort du cadre linéaire ou gaussien, l'estimation est complexe. Le premier problème que j'ai abordé avec Marc Lavielle et France Mentré **[A1]** traite le cas d'observations censurées. Un deuxième travail, réalisé en collaboration avec Sophie Donnet, aborde la problématique d'une fonction de régression solution d'un système d'équations différentielles ordinaires (EDO) **[A5]**. Avec Fabienne Comte **[S1]** et Julien Stirnemann **[S4]**, nous considérons le cas où les effets aléatoires ne sont pas gaussiens et proposons une estimation non paramétrique de leur densité. Enfin, je présente dans ce chapitre deux exemples concrets d'utilisation des modèles mixtes, qui concernent la modélisation de la décroissance de la charge virale dans le contexte du VIH **[A12]** et la prédiction de la croissance foetale **[A14, A16]**. J'ai étudié d'autres questions méthodologiques (modèles mixtes ayant un niveau supplémentaire de variabilité **[A7]**, optimalité de protocoles **[A4]**, données manquantes **[A6]**, recalage d'une séquence d'images médicales **[A8]**) et d'autres applications biomédicales (comparaison d'efficacité de traitements anti-cancéreux par modélisation de croissance tumorale **[A9]**, comparaison d'efficacité de traitements anti-paludiques à partir de données catégorielles répétées **[A16]**) mais qui ne sont pas détaillées dans ce manuscrit.

Dans la section 1.1, j'introduis les modèles mixtes et les notations qui serviront tout au long de ce chapitre. Les développements méthodologiques précédemment cités sont développés dans les sections 1.2 et 1.3. Puis les exemples d'utilisation des modèles mixtes pour l'analyse de données biologiques sont présentés dans la section 1.4.

1.1 Modèles et notations

On note $\mathbf{Y}_k$ le vecteur des données Y_{kj} de l'individu k mesurées aux temps t_{kj}, pour $k = 1, \ldots, N$, $j = 1, \ldots, J_k$. On note $\mathbf{Y} = (\mathbf{Y}'_1, \ldots, \mathbf{Y}'_N)'$ le vecteur de l'ensemble des mesures où X' représente la transposée du vecteur X. On suppose qu'il existe deux fonctions de régression paramétriques f et g telles que

$$\begin{array}{rcl} Y_{kj} & = & f(\phi_k, t_{kj}) + g(\phi_k, t_{kj})\, \varepsilon_{kj}, \quad k = 1, \ldots, N,\ j = 1, \ldots, J_k, \\ \varepsilon_{kj} & \sim_{i.i.d.} & \mathcal{N}(0, \sigma^2), \quad k = 1, \ldots, N,\ j = 1, \ldots, J_k, \\ \phi_k & \sim_{i.i.d.} & \mathcal{N}(\mu, \Omega), \quad k = 1, \ldots, N, \end{array} \tag{1.1}$$

où (ε_{kj}) sont des variables aléatoires représentant les erreurs résiduelles et ϕ_k est un vecteur aléatoire non observé, propre à l'individu k, représentant ses paramètres individuels. On suppose que les erreurs résiduelles $(\varepsilon_{kj})_{1 \leq k \leq N, 1 \leq j \leq J_k}$ sont i.i.d., que les vecteurs aléatoires $(\phi_k)_{1 \leq k \leq N}$ sont également i.i.d. et indépendants des erreurs résiduelles. Sauf indication contraire, on suppose que les erreurs résiduelles sont gaussiennes, d'espérance nulle et de variance σ^2. On note $\varepsilon_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kJ_k})'$ le vecteur des erreurs résiduelles de l'individu k. On suppose que le vecteur individuel ϕ_k est de dimension p, de densité gaussienne d'espérance μ et de matrice de variance Ω. On note $\phi = (\phi'_1, \ldots, \phi'_N)'$ le vecteur des paramètres aléatoires.

Le vecteur $\theta \in \Theta$ des paramètres du modèle est :

$$\theta = (\mu, \Omega, \sigma^2).$$

Ces paramètres sont aussi appelés paramètres de population, pour les distinguer des paramètres individuels ϕ_k. On cherche à estimer θ à partir des observations $\mathbf{Y}$. On note θ_0 la vraie valeur du paramètre.

Lorsque les fonctions f et g sont linéaires en ϕ_k, le modèle est dit linéaire mixte ou à effets mixtes (en distinguant les paramètres fixes μ des paramètres aléatoires ϕ_k). On considère le cas plus général où les fonctions f et/ou g sont non linéaires par rapport à ϕ, le modèle est alors appelé modèle non-linéaire mixte. Les modèles mixtes rentrent dans la famille des modèles à données incomplètes. On distingue alors $\mathbf{Y}$, le vecteur des observations, de $(\mathbf{Y}, \phi)$, le vecteur de l'ensemble des données avec ϕ non observé. Par indépendance des individus, la fonction de vraisemblance d'un modèle mixte est définie par :

$$L(\mathbf{Y}; \theta) \;=\; \prod_{k=1}^{N} L(\mathbf{Y}_k; \theta) = \prod_{k=1}^{N} \int L_c(\mathbf{Y}_k, \phi_k; \theta) d\phi_k,$$

où $L_c(\mathbf{Y}_k, \phi_k; \theta)$ est la vraisemblance des données complètes de l'individu k. Cette vraisemblance complète est égale à

$$L_c(\mathbf{Y}_k, \phi_k; \theta) \;=\; p(\mathbf{Y}_k | \phi_k; \theta)\; p(\phi_k; \theta),$$

où $p(\mathbf{Y}_k | \phi_k; \theta)$ est la densité de probabilité du vecteur $\mathbf{Y}_k$ conditionnellement au vecteur ϕ_k et $p(\phi_k; \theta)$ est la densité de probabilité du vecteur ϕ_k. Si les fonctions f et/ou g sont non-linéaires en ϕ_k, l'intégrale de la vraisemblance complète, et donc la fonction de vraisemblance, n'ont pas, en règle générale, de forme explicite. Alors l'estimateur du maximum de vraisemblance (EMV) $\widehat{\theta}$, défini par

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\mathbf{Y}; \theta),$$

est difficilement calculable.

1.2 Estimation paramétrique

1.2.1 Algorithme SAEM-MCMC

De nombreuses méthodes statistiques ont été proposées pour l'estimation paramétrique des modèles mixtes (voir Pinheiro et Bates (2000) ou Guedj *et al.* (2007)). J'ai choisi de travailler sur des versions stochastiques de l'algorithme Expectation-Maximization (EM) (Dempster *et al.*, 1977). Pour les modèles non-linéaires mixtes, l'étape E n'est pas analytique. En effet la distribution conditionnelle de ϕ_k sachant les observations n'est pas explicite à cause de la non-linéarité de f et g en ϕ_k. L'algorithme Stochastic Approximation EM (SAEM) proposé par Delyon *et al.* (1999) approxime l'espérance conditionnelle calculée dans l'étape E par un algorithme d'approximation stochastique du type Robbins-Monroe. La convergence de l'algorithme SAEM est montrée, entre autres, sous l'hypothèse suivante :

(M) La vraisemblance complète $L_c(\mathbf{Y}, \phi, \theta)$ appartient à la famille des modèles exponentiels

$$\log L_c(\mathbf{Y}, \phi; \theta) = -\Psi(\theta) + \langle S(\mathbf{Y}, \phi), \nu(\theta) \rangle \,,$$

où Ψ et ν sont des fonctions de θ, $S(\mathbf{Y}, \phi)$ sont les statistiques exhaustives du modèle, prenant leur valeur dans un sous-ensemble $\mathcal{S}$ de $\mathbb{R}^d$ et $\langle \cdot, \cdot \rangle$ est le produit scalaire sur $\mathbb{R}^d$.

Sous l'hypothèse **(M)**, l'étape E de l'algorithme SAEM se simplifie en le calcul de l'espérance conditionnelle $\mathbb{E}(S(\mathbf{Y}, \phi)|\mathbf{Y}, \theta')$ où l'espérance est calculée sous la loi de ϕ conditionnellement aux observations $\mathbf{Y}$. Ce calcul est approché via un algorithme stochastique à l'aide d'une étape de simulation et une étape d'approximation stochastique. Cet algorithme nécessite la simulation d'une seule réalisation des variables non observées à l'étape E, permettant de diminuer considérablement le temps de calcul par rapport à des algorithmes Monte Carlo EM. La convergence de l'algorithme SAEM vers un maximum local de la vraisemblance est assurée sous des conditions relativement générales de régularité de la vraisemblance complète L_c et des statistiques exhaustives S. Cet algorithme ne peut pas être directement utilisé pour les modèles non-linéaires mixtes, la distribution conditionnelle de la variable ϕ_k sachant les observations $\mathbf{Y}_k$ n'étant pas explicite. Kuhn et Lavielle (2004, 2005) ont proposé de coupler l'algorithme SAEM avec une méthode de Monte-Carlo par Chaîne de Markov (MCMC) pour réaliser cette étape. L'algorithme SAEM-MCMC s'écrit ainsi :

***Algorithme* 1** (Algorithme SAEM-MCMC)**.**

- *Itération 0 : initialisation de l'algorithme avec une valeur $\widehat{\theta}_0$ du paramètre.*
- *Itération $m \geq 1$, pour la valeur courante $\widehat{\theta}_m$ du paramètre, réalisation des étapes suivantes :*

  Etape S : *Simulation de $\phi^{(m)}$ via la simulation d'une chaîne de Markov ayant pour loi stationnaire la distribution conditionnelle $p(\phi|\mathbf{Y}; \widehat{\theta}_m)$ (algorithme MCMC),*

  Etape SA : *Approximation stochastique de $\mathbb{E}(S(\mathbf{Y}, \phi)|\mathbf{Y}, \widehat{\theta}_m)$ par*

$$s_{m+1} = s_m + \gamma_m (S(\mathbf{Y}, \phi^{(m)}) - s_m),$$

  où (γ_m) est une suite de pas décroissants vers 0 tels que $\sum_{m\geq 1} \gamma_m = \infty$ et $\sum_{m\geq 1} \gamma_m^2 < \infty$,

  Etape M : *Actualisation du paramètre*

$$\widehat{\theta}_{m+1} = \arg\max_{\theta \in \Theta} \{-\Psi(\theta) + \langle s_{m+1}, \nu(\theta) \rangle\}.$$

L'algorithme MCMC de l'étape S dépend du modèle mixte considéré. Kuhn et Lavielle (2004) montrent la convergence presque sûre de l'algorithme SAEM-MCMC vers l'EMV sous les conditions générales de convergence de l'algorithme SAEM (modèle dans la famille exponentielle, régularité des statistiques exhaustives) ainsi que sous l'hypothèse que la chaîne de Markov générée par MCMC est uniformément ergodique.

L'étude théorique de l'EMV est complexe dans les modèles mixtes. La contribution principale dans ce domaine est due à Nie et Yang (2005). Il étudie les propriétés asymptotiques de l'EMV sous différentes asymptotiques, quand le nombre d'individus et/ou le nombre d'observations par individu tendent

vers l'infini. En supposant une série d'hypothèses techniques sur la régularité de la vraisemblance et de la fonction de régression, il montre la consistance faible et la normalité asymptotique de l'EMV.

Depuis ma thèse, j'ai travaillé au développement de l'algorithme SAEM-MCMC pour les modèles mixtes : mises en place de tests de comparaison de groupes de patients basés sur les modèles mixtes (test de Wald et test du rapport de vraisemblance **[A3]**), variabilité inter-occasion **[A7]**, prise en compte de données censurées **[A1]**, fonction de régression définie par système différentiel **[A2]**. Ces différentes versions de l'algorithme SAEM sont programmées dans le logiciel MONOLIX© (Modèles NOn Linéaires à effets mixtes). J'ai activement participé pendant ma thèse au développement de ce logiciel, au sein du groupe MONOLIX, dirigé par Marc Lavielle et France Mentré (http ://software.monolix.org). Ce logiciel est dédié à l'estimation paramétrique pour les modèles mixtes et leur utilisation dans le développement d'un médicament (Lavielle et Mentré, 2007). Une grande librairie de modèles de pharmacocinétique et pharmocodynamie est intégrée dans MONOLIX. Ce logociel représente une alternative sérieuse au logiciel NONMEN, qui était utilisé par la très grande majorité des équipes de biométrie des laboratoires pharmaceutiques. En plus des propriétés théoriques très satisfaisantes de l'estimateur obtenu par l'algorithme SAEM, le logiciel MONOLIX a montré sa capacité à estimer les paramètres de modèles complexes, là où NONMEM atteint très vite des limites numériques. MONOLIX est désormais utilisé par de nombreux laboratoires pharmaceutiques dans les différentes phases de développement d'un médicament.

Les développements concernant les données censurées et les modèles d'équations différentielles ordinaires sont détaillés ci-dessous.

1.2.2 Données censurées à gauche

La mesure d'une quantité biologique peut être soumise à une limite de quantification de l'appareil de mesure. L'observation est donc censurée à gauche. On note LOQ la limite (connue) de quantification. On introduit la variable binaire d'indicatrice de censure

$$\delta_{kj} = \mathbb{1}_{Y_{kj} \geq LOQ}.$$

L'observation disponible n'est pas la variable Y_{kj} mais le couple $(Y_{kj}^{obs}, \delta_{kj})$ où la variable Y_{kj}^{obs} est définie par

$$Y_{kj}^{obs} = \max(LOQ, Y_{kj}) = \left\{ \begin{array}{ccc} Y_{kj} & \text{si} & \delta_{kj} = 1, \\ LOQ & \text{si} & \delta_{kj} = 0. \end{array} \right.$$

Différentes méthodes d'estimation ont été proposées pour estimer le paramètre θ dans ce contexte. La méthode la plus naïve omet les données censurées en ne gardant que les Y_{kj}^{obs} pour les (k, j) tels que $\delta_{kj} = 1$. Une autre proposition est de remplacer la première observation censurée (dans le temps) par une valeur fixe (LOQ ou LOQ/2) et d'omettre les observations censurées suivantes. On peut aussi citer Jacqmin-Gadda *et al.* (2000) et Hughes (1999) pour les modèles linéaires mixtes. Le cas des modèles non-linéaires mixtes a été peu traité, car la vraisemblance n'est alors pas explicite.

Dans **[A1]**, nous développons une version de l'algorithme SAEM adaptée aux données censurées, qui permet d'obtenir un estimateur convergent avec un temps de calcul réduit (de l'ordre de quelques secondes). On introduit les variables censurées $Y_{kj}^{cens} = Y_{kj}$ pour tout (k, j) tel que $\delta_{kj} = 0$. On note $\mathbf{Y}_k^{cens}$ le vecteur des données censurées de l'individu k. Par convention, si un individu n'a pas de données censurées, $\mathbf{Y}_k^{cens} = \varnothing$. On introduit le vecteur $\delta_k = (\delta_{kj})_{1 \leq j \leq J_k}$ de l'ensemble des variables indicatrices du sujet k et $\delta = (\delta_1', \ldots, \delta_N')'$. La vraisemblance du modèle est alors définie par

$$L(\mathbf{Y}^{obs}, \delta, \theta) = \prod_{k=1}^{N} \int L_c(\mathbf{Y}_k^{obs}, \delta_k, \mathbf{Y}_k^{cens}, \phi_k; \theta) d\phi_k d\mathbf{Y}_k^{cens} \tag{1.2}$$

où $L_c(\mathbf{Y}_k^{obs}, \delta_k, \mathbf{Y}_k^{cens}, \phi_k; \theta)$ est la vraisemblance complète, $\mathbf{Y}_k^{cens}, \phi_k$ étant les variables non observées. On peut noter que ces deux variables ne sont pas indépendantes mais que les variables $\mathbf{Y}_k^{cens}$ sont indépendantes entre elles conditionnellement aux variables ϕ_k.

Nous faisons l'hypothèse que le mécanisme de censure est indépendant du modèle de régression, ce qui est raisonnable puisqu'il s'agit d'une censure liée à l'appareil de mesure. Alors conditionnellement aux paramètres individuels ϕ_k, les données censurées $\mathbf{Y}_k^{cens}$ et observées $\mathbf{Y}_k^{obs}$ telles que $\delta_{kj} = 1$, sont indépendantes. On en déduit que la vraisemblance complète est égale à

$$L_c(\mathbf{Y}_k^{obs}, \delta_k, \mathbf{Y}_k^{cens}, \phi_k; \theta) = \prod_{(k,j)|\delta_{kj}=1} p(\mathbf{Y}_k^{obs}|\phi_k; \theta) \prod_{(k,j)|\delta_{kj}=0} p(\mathbf{Y}_k^{cens}|\phi_k; \theta) \; p(\phi_k; \theta),$$

où les densités $p(\mathbf{Y}_k^{obs}|\phi_k; \theta)$ et $p(\mathbf{Y}_k^{cens}|\phi_k; \theta)$ sont gaussiennes. Sous l'hypothèse que la vraisemblance complète vérifie l'hypothèse **(M)**, nous proposons l'algorithme SAEM-MCMC suivant :

***Algorithme* 2** (Algorithme SAEM-MCMC pour données censurées)**.** *A l'itération $m \geq 1$, pour la valeur courante $\widehat{\theta}_m$ du paramètre*

Etape S : *Simulation de $(\mathbf{Y}^{cens(m)}, \phi^{(m)})$ via la simulation d'une chaîne de Markov ayant pour loi stationnaire $p(\mathbf{Y}^{cens}, \phi|\mathbf{Y}; \widehat{\theta}_m)$ par algorithme de Gibbs :*

1. *Simulation de $\phi^{(m)}$ par un algorithme de Metropolis-Hastings ayant $p(\phi|\mathbf{Y}^{obs}, \mathbf{Y}^{cens(m-1)}; \widehat{\theta}_m)$ comme distribution stationnaire,*
2. *Simulation, pour tout $k = 1, \ldots, N$, tout $j = 1, \ldots, J_k$ tel que $\delta_{kj} = 0$ de $\mathbf{Y}_{kj}^{cens(m)}$ avec la loi gaussienne tronquée $p(\mathbf{Y}_k^{cens}|\mathbf{Y}_k^{obs}, \phi_k^{(m)}; \widehat{\theta}_m)$ d'espérance $f(\phi_k^{(m)}, t_{kj})$, de variance $\widehat{\sigma^2}^{(m)} \, g^2(\phi_k^{(m)}, t_{kj})$ et tronquée à droite par la valeur LOQ,*

Etape SA : *Approximation stochastique de $\mathbb{E}(S(\mathbf{Y}^{obs}, \delta, \mathbf{Y}^{cens}, \phi)|\mathbf{Y}, \widehat{\theta}_m)$*

$$s_{m+1} = s_m + \gamma_m(S(\mathbf{Y}, \delta, \mathbf{Y}^{cens(m)}, \phi^{(m)}) - s_m)$$

où (γ_m) est une suite de pas décroissants vers 0 tels que $\sum_{m\geq 1} \gamma_m = \infty$ et $\sum_{m\geq 1} \gamma_m^2 < \infty$,

Etape M : *Actualisation du paramètre*

$$\widehat{\theta}_{m+1} = \arg\max_{\theta\in\Theta}\{-\Psi(\theta) + \langle s_{m+1}, \nu(\theta)\rangle\}.$$

La première étape de l'algorithme de Gibbs peut se réaliser avec un algorithme de Metropolis-Hastings classique. La deuxième étape nécessite une méthode de simulation d'une gaussienne tronquée. Les hypothèses de convergence de l'algorithme SAEM-MCMC restent identiques. Les conditions sur la chaîne de Markov générée lors de l'étape de simulation sont en particulier vérifiées par l'algorithme de Gibbs proposé.

Une étude de simulation illustre le fait que l'algorithme proposé permet de réduire considérablement les biais d'estimation inhérents aux méthodes naïves. L'analyse des données de l'essai clinique TRIANON ANRS81 qui compare l'efficacité de deux traitements anti-rétroviraux contre le VIH confirme la capacité de l'algorithme SAEM-MCMC à estimer les paramètres de modèles mixtes complexes.

La méthode que nous proposons a donc de bonnes propriétés théoriques et est efficace en temps de calcul. Elle représente une très bonne alternative aux méthodes existantes.

1.2.3 Modèles définis par équations différentielles ordinaires

Cette section aborde le problème de l'estimation des paramètres d'un modèle mixte défini par une équation différentielle ordinaire (EDO), quand l'EDO n'a pas de solution analytique. L'approche usuelle est d'évaluer la fonction de régression f par une méthode d'intégration numérique à chaque fois que l'algorithme d'estimation le nécessite. Pour l'algorithme SAEM-MCMC, cela est nécessaire à chaque

itération de l'algorithme MCMC utilisé dans l'étape de simulation de chaque itération de l'algorithme SAEM. Un compromis doit donc être trouvé entre la stabilité de la méthode numérique, sa précision et le temps de calcul qu'elle requiert. La méthode d'estimation est alors utilisée sur un modèle mixte approché où la fonction de régression est une approximation de la solution de l'EDO. On maximise donc une pseudo-vraisemblance. A ma connaissance, l'erreur induite par cette approximation n'avait jamais été étudiée auparavant. Ces aspects sont abordés dans **[A2]** et présentés ci-dessous.

Nous considérons le modèle non-linéaire mixte (1.1) avec $g = 1$ (modèle d'erreur homoscédastique), appelé modèle $\mathcal{M}$ dans cette section. Nous supposons que les temps d'observations (t_{kj}) appartiennent à l'intervalle de temps $[t_0, T]$ et que la fonction $f : \mathbb{R}^p \times \mathbb{R} \longrightarrow \mathbb{R}$ est définie comme la solution de l'EDO suivante :

$$\begin{aligned} &\frac{\partial f(\phi,t)}{\partial t} = F(f(\phi,t),t,\phi), \quad t \in [t_0,T], \\ &f(\phi,t_0) = f_0(\phi), \end{aligned} \tag{1.3}$$

où la fonction $F : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p \longrightarrow \mathbb{R}$ est explicite et connue et la condition initiale $f_0(\phi) \in \mathbb{R}$ est connue.

Lors de l'estimation du paramètre θ, un schéma numérique de résolution de l'EDO (1.3) est appliqué sur des sous-intervalles $[\tau_i, \tau_{i+1}[$, $i = 0, \ldots, M-1$ de l'intervalle de temps $[t_0, T]$. On note $h = \max_{0 \leq i \leq M-1}(\tau_{i+1} - \tau_i)$ le pas de temps maximal. La précision du schéma est définie via son ordre. Soit f_h la fonction approchée par le schéma d'intégration numérique de pas h. Le schéma est d'ordre r s'il existe une constante C telle que

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

L'estimation de θ est réalisée sur un modèle mixte approché par le schéma numérique, noté $\mathcal{M}_h$, où

$$Y_{kj} \;=\; f_h(\phi_k, t_{kj}) + \varepsilon_{kj}, \quad k = 1, \ldots, N,\ j = 1, \ldots, J_k,$$

et les hypothèses sur ε_{kj} et ϕ_k sont les mêmes que pour le modèle $\mathcal{M}$. On note avec un indice h toutes les lois qui se rapportent à ce modèle. Dans l'étape S de l'algorithme SAEM-MCMC, l'algorithme MCMC a alors pour loi stationnaire cible la loi $p_h(\phi|\mathbf{Y};\theta)$. La fonction maximisée est la vraisemblance $L_h(\mathbf{Y};\theta)$ du modèle $\mathcal{M}_h$, qui est une pseudo-vraisemblance pour le modèle $\mathcal{M}$. L'erreur induite par le schéma numérique est étudiée dans le théorème suivant.

Théorème 1. *Sous des hypothèses de régularité sur f et F, on a*

1. *Il existe une constante $C_{\mathbf{Y}}$ telle que si h est suffisamment petit,*

$$\|p(\phi|\mathbf{Y};\theta) - p_h(\phi|\mathbf{Y};\theta)\|_{TV} \leq C_{\mathbf{Y}}\, h^r.$$

2. *Soit σ_0^2 la vraie valeur du paramètre de variance. On suppose qu'il existe $\sigma_{min} > 0$ tel que $\sigma_0^2 > \sigma_{min}^2$. Il existe une constante $C'_{\mathbf{Y}}$ telle que*

$$\sup_{\theta=(\beta,\sigma^2)\ |\sigma^2 > \sigma_{min}^2} |L(\mathbf{Y};\theta) - L_h(\mathbf{Y};\theta)| \leq C'_{\mathbf{Y}}\, h^r.$$

La distance entre la vraie loi conditionnelle $p(\phi|\mathbf{Y};\theta)$ et la loi approchée $p_h(\phi|\mathbf{Y};\theta)$ cible de l'algorithme MCMC décroît avec h. De même, la distance entre la vraisemblance du modèle $\mathcal{M}$ et la pseudo-vraisemblance L_h décroît avec h. On introduit ensuite une hypothèse supplémentaire sur les hessiennes des vraisemblances des deux modèles.

(H) 1. Les fonctions $L(\mathbf{Y};\theta)$ et $L_h(\mathbf{Y};\theta)$ sont deux fois différentiables.

2. Soit θ_∞ et $\theta_{h,\infty}$ les maxima des fonctions $L(\mathbf{Y};\theta)$ et $L_h(\mathbf{Y};\theta)$ respectivement. Il existe deux constantes ε_1 et ε_2 telles que pour tout $\theta \in \{\theta_{h,\infty} + t(\theta_{h,\infty} - \theta_\infty), t \in [0,1]\}$, et pour tout x, on a

$$\begin{aligned} -x^t H_L(\theta) x &\geq \varepsilon_1 \|x\|^2 \\ -x^t H_{L_h}(\theta) x &\geq \varepsilon_2 \|x\|^2 , \end{aligned}$$

où H_L et H_{L_h} sont les matrices hessiennes de $L(\mathbf{Y};\cdot)$ et $L_h(\mathbf{Y};\cdot)$ respectivement.

On peut alors en déduire un résultat sur les maxima des vraisemblance des deux modèles.

Corollaire* 1.** *On suppose que l'hypothèse* ***(H) *et les hypothèses du Théorème 1 sont vérifiées. Alors,* $(\widehat{\theta}_m)_{m\geq 1}$*, la suite d'estimateurs fournie par l'algorithme SAEM-MCMC sur le modèle* $\mathcal{M}_h$*, converge presque surement vers* $\theta_{h,\infty}$ *et il existe une constante* C*, indépendante de* θ *telle que*

$$\|\theta_{h,\infty} - \theta_\infty\|^2 \leq C h^r .$$

En conclusion, sous des hypothèses de régularité sur les deux modèles, l'algorithme SAEM-MCMC fournit un estimateur qui converge vers $\theta_{h,\infty}$, dont la distance au vrai maximum de vraisemblance décroît avec h.

Le schéma numérique est utilisé à chaque itération de l'algorithme MCMC et et de l'algorithme SAEM. Le temps de calcul de la résolution de l'EDO est donc un enjeu fort pour l'estimation des paramètres du modèle mixte. Afin d'améliorer le temps de calcul de l'algorithme SAEM-MCMC sur le modèle $\mathcal{M}_h$, nous introduisons une modification de l'algorithme de linéarisation locale (LL) de résolution d'EDO proposé par Biscay *et al.* (1996). Cette modification de LL est adaptée à son inclusion dans un algorithme de Metropolis-Hastings (MH). Dans un algorithme MH, la solution de l'EDO est évaluée en une valeur courante ϕ du paramètre, puis, après proposition d'un paramètre candidat ϕ^c dans un voisinage de ϕ, elle est évaluée en ϕ^c. Nous proposons d'utiliser la solution $(f_h(\phi,\tau_i))_{1\leq i\leq M}$ obtenue avec le schéma LL de pas h en ϕ sur la grille de temps $(\tau_i)_{1\leq i\leq M}$ pour obtenir une évaluation de f en ϕ^c. On note $(f_{h,\phi}(\phi^c,\tau_i))_{1\leq i\leq M}$ cette nouvelle évaluation. Nous proposons de la définir comme la solution de l'EDO linéaire suivante :

$$\begin{aligned} \frac{\partial f_{h,\phi}(\phi^c,t)}{\partial t} &= F(f_h(\phi,\tau_i),\tau_i,\phi) + \frac{dF}{df}(f_h(\phi,\tau_i),\tau_i,\phi)\,(f_{h,\phi}(\phi^c,t) - f_h(\phi,\tau_i)) \\ &\quad + \frac{dF}{dt}(f_h(\phi,\tau_i),\tau_i,\phi)\,(t-\tau_i) + \frac{dF}{d\phi}(f_h(\phi,\tau_i),\tau_i,\phi)\,(\phi^c - \phi). \end{aligned}$$

Cette EDO vient d'une linéarisation en ϕ de l'EDO (1.3). L'ordre de convergence de ce schéma numérique est étudié dans le lemme suivant :

Lemme 1. *Supposons que* ϕ^c *reste dans un compact de* $\mathbb{R}^p$*. Il existe deux constantes* C_1 *et* C_2 *telles que pour tout* $t \in [t_0, T]$ *et tout* ϕ*,*

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

Le nouveau schéma est donc d'ordre 2 en h et ϕ et étend les résultats obtenus par Biscay *et al.* (1996). L'étude théorique de l'algorithme MCMC basé sur ce nouveau schéma est complexe car la probabilité d'acceptation utilisée n'a pas une forme standard. Elle est discutée dans **[A2]**. Une étude de simulation montre que le nouveau schéma numérique s'avère très compétitif en temps de calcul et donne des résultats d'estimation proches de ceux obtenus avec des schémas numériques d'ordre plus élevé.

Finalement, l'algorithme SAEM montre à nouveau ton son potentiel, contrairement à la plupart des autres méthodes d'estimation pour les modèles mixtes qui sont vite limitées pour estimer des systèmes complexe. Une utilisation de cet algorithme pour un système différentiel de dimension 5 est présenté dans la section 1.4.1 sur des données réelles.

1.3 Estimation non paramétrique

En collaboration avec Fabienne Comte, nous cherchons à alléger l'hypothèse de normalité des paramètres individuels ϕ_k faite dans le modèle mixte (1.1), hypothèse qui peut s'avérer forte. Dans la section 1.3.1, nous considérons l'estimation non paramétrique de la densité de ϕ_k dans le cadre d'un modèle linéaire mixte **[S1]**. Dans la section 1.3.2, nous considérons le cas particulier de l'estimation de la densité de l'ordonnée à l'origine quand un sous-échantillon de mesures répétées au temps $t = 0$ est disponible **[S4]** (travail en collaboration avec Julien Stirnemann).

1.3.1 Modèle linéaire mixte

Nous considérons le modèle linéaire mixte simple suivant

$$Y_{kj} = \alpha_k + \beta_k t_j + \varepsilon_{kj}, \quad k = 1, \dots, N \quad j = 1, \dots, J, \tag{1.4}$$

où $\phi_k = (\alpha_k, \beta_k)$ représente le vecteur des variables aléatoires individuelles du sujet k et les temps $t_{kj} = t_j$ sont identiques pour tous les individus $k = 1, \dots, N$. On a en particulier $J_k = J$. On note f_α et f_β les densités inconnues des deux variables α_1 et β_1. On ne fait pas d'hypothèse d'indépendance sur les variables α_1 et β_1, qui peuvent donc être dépendantes ou indépendantes.

Notre but est d'estimer non paramétriquement les densités f_α et f_β lorsque la densité f_ε est connue ou non. L'estimation des densités f_α et f_β dans le cadre des modèles mixtes a été considérée lorsque les erreurs ε_{kj} sont gaussiennes. On peut citer par exemple Mallet *et al.* (1988); Kuhn (2003); Chafai et Loubes (2006); Antic *et al.* (2009). Nous proposons une approche basée sur des outils de déconvolution. Cette approche a été très étudiée dans des contextes variés mais elle est innovante dans le cadre des modèles mixtes. Les travaux les plus proches sont ceux développés dans le cadre de mesures répétées par Delaigle *et al.* (2008); Meister et Neumann (2010) mais qui ne traitent pas du cas des modèles mixtes.

Dans **[S1]**, nous proposons une méthode d'estimation de f_α et f_β dans les cas où la densité de ε est connue ou non. Dans le deuxième cas, à l'aide d'une transformation astucieuse des données et sous l'hypothèse que la loi du bruit est symétrique, nous sommes capables d'estimer la densité de ε^2. Puis, dans le même esprit que l'estimateur proposé par Comte et Lacour (2011) dans le cas où un échantillon de bruit pur est disponible, nous utilisons cet estimateur de la loi du bruit pour estimer f_α et f_β par déconvolution. Nous montrons que ces estimateurs réalisent le compromis biais/variance. La procédure d'estimation est détaillée ci-dessous.

Nous nous plaçons sous l'hypothèse suivante :

(A) les erreurs de mesures ε_{kj} sont i.i.d. avec une densité f_ε telle que $\mathbb{E}(e^{iu\varepsilon}) \neq 0$, pour tout $u \in \mathbb{R}$.

Cette hypothèse permet de définir des estimateurs de densité faisant apparaitre la densité f_ε au dénominateur.

Dans la suite, et pour simplifier les notations, nous supposons que J est pair et on note $\Delta_j = t_{2j} - t_{2j-1}$. On note f_X la densité d'une variable aléatoire X. On introduit quelques notations liées à la théorie de Fourier. Si f est une fonction intégrable, on note $f^*(u) = \int e^{iux} f(x)dx$ sa transformée de Fourier sur $\mathbb{R}$. Pour deux fonctions réelles de carré intégrable f et g, on note $(f \star g)(x) = \int f(x-y)g(y)dy$ le produit de convolution de f et g. Si f et g sont intégrables et de carré intégrables, on a alors $(f \star g)^* = f^* g^*$. Si f est intégrable et de carré intégrable, la formule de Fourier inverse nous donne $f(x) = 1/(2\pi) \int e^{-ixu} f^*(u)du$.

Cas où la densité du bruit f_ε est connue. Nous présentons d'abord un estimateur de f_β. Il faut pouvoir "isoler" la variable β pour en estimer la densité. Pour cela, nous considérons la transformation des données correspondant à la différence normalisée entre deux observations successives, pour $j = 1, \dots, J/2$,

$$Z_{kj} := \frac{Y_{k2j} - Y_{k2j-1}}{\Delta_j} = \beta_k + \frac{\varepsilon_{k2j} - \varepsilon_{k2j-1}}{\Delta_j}.$$

Pour j fixé, les variables $(Z_{kj})_{k=1,\dots,N}$ sont i.i.d. mais les variables Z_{kj} et Z_{kl} pour $j \neq l$ ne sont pas indépendantes. Comme $f^*_{(\varepsilon_{k2j}-\varepsilon_{k2j-1})/\Delta_j}(u) = \mathbb{E}\left(e^{i\frac{u}{\Delta_j}\varepsilon}\right)\mathbb{E}\left(e^{-i\frac{u}{\Delta_j}\varepsilon}\right) = \left|f^*_\varepsilon(\frac{u}{\Delta_j})\right|^2$, on en déduit que

$$f^*_\beta(u) = \frac{2}{J}\sum_{j=1}^{J/2}\frac{f^*_{Z_j}(u)}{|f^*_\varepsilon(u/\Delta_j)|^2}.$$

La formule d'inversion de Fourier et l'estimateur de $f^*_{Z_j}(u)$

$$\widehat{f^*_{Z_j}}(u) = \frac{1}{N}\sum_{k=1}^{N} e^{iuZ_{kj}} = \frac{1}{N}\sum_{k=1}^{N} e^{iu\frac{Y_{k2j}-Y_{k2j-1}}{\Delta_j}}, \tag{1.5}$$

permettent de définir un estimateur de f_β tronqué au seuil πm :

$$\widehat{f_{\beta,m}}(x) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} e^{-ixu}\frac{2}{J}\sum_{j=1}^{J/2}\frac{\widehat{f^*_{Z_j}}(u)}{|f^*_\varepsilon(u/\Delta_j)|^2}du = \frac{2}{NJ}\sum_{k=1}^{N}\sum_{j=1}^{J/2}\frac{1}{2\pi}\int_{-\pi m}^{\pi m} e^{-ixu}\frac{e^{iuZ_{kj}}}{|f^*_\varepsilon(u/\Delta_j)|^2}du. \tag{1.6}$$

Le seuil πm est introduit pour éviter des problèmes de convergence de l'intégrale. La particularité de cet estimateur par rapport à un estimateur de déconvolution classique vient de la dépendance des observations $(Z_{kj})_{k=1,\dots,N}$ d'un même individu. Il faut donc tenir compte de cette dépendance dans l'étude du risque de l'estimateur.

Pour estimer f_α, nous distinguons deux cas. Le premier cas suppose que des observations au temps 0 sont disponibles. On note Y_{k0} l'observation correspondante. On a alors

$$Y_{k0} = \alpha_k + \varepsilon_{k0},\ k = 1,\dots,N,$$

qui est un modèle classique de déconvolution. On considère l'estimateur de f_α étudié par Comte *et al.* (2006) :

$$\widehat{f^0_{\alpha,m}}(x) = \frac{1}{2\pi N}\sum_{k=1}^{N}\int_{-\pi m}^{\pi m} e^{-ixu}\frac{e^{iuY_{k0}}}{f^*_\varepsilon(u)}du. \tag{1.7}$$

Dans le deuxième cas, aucune observation au temps 0 n'est disponible. En suivant un raisonnement similaire à celui de la construction de $\widehat{f_{\beta,m}}$, on introduit pour $j = 1,\dots,J/2$, les variables

$$V_{kj} = \frac{Y_{k2j}}{t_{2j}} - \frac{Y_{k2j-1}}{t_{2j-1}} = \left(\frac{1}{t_{2j}} - \frac{1}{t_{2j-1}}\right)\alpha_k + \left(\frac{\varepsilon_{k2j}}{t_{2j}} - \frac{\varepsilon_{k2j-1}}{t_{2j-1}}\right).$$

On note $p_j = \frac{1}{t_{2j}} - \frac{1}{t_{2j-1}}$. On a pour tout $j = 1,\dots,J/2$,

$$f^*_\alpha(u) = \frac{f^*_{V_j}(u/p_j)}{f^*_\varepsilon(u/(p_j t_{2j}))\, f^*_\varepsilon(-u/(p_j t_{2j-1}))}. \tag{1.8}$$

Le dénominateur peut prendre des valeurs très grandes quand les temps d'observations t_{2j} sont grands, et ainsi engendrer une instabilité numérique. Nous proposons de construire un estimateur de f_α basé uniquement sur la première observation V_{k1} pour les N individus

$$\widehat{f_{\alpha,m}}(x) = \frac{1}{2\pi N}\sum_{k=1}^{N}\int_{-\pi m}^{\pi m} e^{-iux}\frac{e^{iV_{k1}u/p_1}}{f^*_\varepsilon(u/(p_1 t_2))\, f^*_\varepsilon(-u/(p_1 t_1))}du. \tag{1.9}$$

Comme pour l'estimateur $\widehat{f_{\beta,m}}$, la dépendance des observations d'un même individu complique l'étude du risque.

Je présente brièvement cette étude du risque des estimateurs $\widehat{f_{\beta,m}}$ et $\widehat{f_{\alpha,m}}$. On note $f_{\beta,m}$ et $f_{\alpha,m}$ les fonctions telles que $f^*_{\beta,m} = f^*_\beta \mathbf{1}_{[-\pi m;\,\pi m]}$ et $f^*_{\alpha,m} = f^*_\alpha \mathbf{1}_{[-\pi m;\,\pi m]}$, qui sont les fonctions réellement estimées par $\widehat{f_{\beta,m}}$ et $\widehat{f_{\alpha,m}}$. Les risques moyen intégrés (MISE) de $\widehat{f_{\beta,m}}$ et $\widehat{f_{\alpha,m}}$, définis par $\mathbb{E}\left(||f_\beta - \widehat{f_{\beta,m}}||^2\right)$ et $\mathbb{E}\left(||f_\alpha - \widehat{f_{\alpha,m}}||^2\right)$, où $||\cdot||$ désigne la norme de $L^2(\mathbb{R})$, sont étudiés dans la proposition suivante.

***Proposition* 1.** *Si f_β est intégrable et de carré intégrable, alors*

$$\mathbb{E}\left\|\widehat{f_{\beta,m}} - f_\beta\right\|^2 \leq \frac{1}{2\pi}\int_{|u|\geq\pi m} |f^*_\beta(u)|^2\, du + \frac{4}{NJ^2}\sum_{j=1}^{J/2}\left(\frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{du}{|f^*_\varepsilon(u/\Delta_j)|^4}\right) + \frac{m}{N}. \tag{1.10}$$

Si f_α est intégrable et de carré intégrable, alors

$$\mathbb{E}\left\|\widehat{f_{\alpha,m}} - f_\alpha\right\|^2 \leq \frac{1}{2\pi}\int_{|u|\geq\pi m} |f^*_\alpha(u)|^2\, du + \frac{1}{2\pi N}\int_{-\pi m}^{\pi m}\frac{du}{\left|f^*_\varepsilon\left(\frac{u}{p_1 t_2}\right)\, f^*_\varepsilon\left(\frac{u}{p_1 t_1}\right)\right|^2} + \frac{m}{N}. \tag{1.11}$$

Les termes $\frac{1}{2\pi}\int_{|u|\geq\pi m} |f^*_\beta(u)|^2\, du$ et $\frac{1}{2\pi}\int_{|u|\geq\pi m} |f^*_\alpha(u)|^2\, du$ sont les termes de biais classiques qui décroissent quand m croît.

Les autres termes sont des termes de variance qui croissent quand m croît. Le terme m/N vient des covariances entre deux observations V_{kj} et $V_{kj'}$ d'un même individu. Il n'apparait pas dans le risque d'un d'estimateur de déconvolution classique basé sur des observations indépendantes. On peut aussi noter que pour l'estimateur $\widehat{f_{\beta,m}}$, le terme de variance fait apparaitre à la fois N, le nombre d'individus, et J le nombre d'observations par individu. On gagne donc un facteur J en moyennant sur les J observations de chaque individu par rapport à l'estimateur de déconvolution classique.

Il faut trouver un compromis entre les termes de biais et de variance pour choisir convenablement le seuil m. Pour cela, on définit des fonctions de pénalité qui permettent d'estimer le meilleur seuil m en minimisant le risque quadratique pénalisé. Pour $\omega = \alpha$ ou $\omega = \beta$, on définit ce seuil optimal par

$$\hat{m}_\omega = \arg\min_{m\in\mathcal{M}_{\omega,N}}\left\{-\|\widehat{f_{\omega,m}}\|^2 + \mathrm{pen}_\omega(m)\right\},$$

avec $\mathcal{M}_{\omega,N} = \{m \in \{1,\dots N\},\ \text{tel que } \mathrm{pen}_\omega(m) \leq 1\}$ et où pour $\omega = \beta$:

$$\mathrm{pen}_\beta(m) = \kappa_\beta\left(\frac{4}{NJ^2}\sum_{j=1}^{J/2}\left(\frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{du}{|f^*_\varepsilon(u/\Delta_j)|^4}\right) + \frac{m}{N}\right),$$

et pour $\omega = \alpha$,

$$\mathrm{pen}_\alpha(m) = \kappa_\alpha\left(\frac{1}{2\pi N}\int_{-\pi m}^{\pi m}\frac{du}{\left|f^*_\varepsilon\left(\frac{u}{p_1 t_2}\right)\, f^*_\varepsilon\left(\frac{u}{p_1 t_1}\right)\right|^2} + \frac{m}{N}.\right).$$

Les constantes κ_β et κ_α doivent être calibrées par simulation. Les bornes des MISE des estimateurs finaux $\widehat{f_{\omega,\hat{m}_\omega}}$ sont étudiées dans le théorème suivant pour un bruit f_ε "ordinary smooth".

Théorème 2. *Supposons que f_ω, pour $\omega = \alpha, \beta$, soit intégrable et de carré intégrable. Supposons que le bruit est "ordinary smooth", c'est à dire qu'il existe deux constantes $c_\varepsilon, C_\varepsilon$ telles que, $\forall u \in \mathbb{R}$,*

$$c_\varepsilon(1+u^2)^\delta \leq 1/|f^*_\varepsilon(u)|^2 \leq C_\varepsilon(1+u^2)^\delta.$$

Alors, il existe des constantes C et C' telles que

$$\mathbb{E}\left(\|\widehat{f_{\omega,\hat{m}_\omega}} - f_\omega\|^2\right) \leq C\inf_{m\in\mathcal{M}_{\omega,N}}\left(\|f_\omega - f_{\omega,m}\|^2 + \mathrm{pen}_\omega(m)\right) + \frac{C'}{N}.$$

Ce théorème nous assure que l'estimateur obtenu par la méthode de sélection de seuil proposée réalise automatiquement le compromis biais/variance, à une constante multiplicative près.

Nous considérons également le cas très courant dans les modèles mixtes où le bruit est gaussien. Dans ce cas, le choix d'un m optimal peut être obtenu directement, sans passer par une méthode de sélection de modèles sous l'hypothèse que f_β et f_α sont suffisamment régulières, y compris quand σ^2 est inconnu. Nous montrons que l'estimateur ainsi obtenu atteint la meilleure vitesse possible (vitesse logarithmique).

Cas où la densité du bruit f_ε est inconnue. Les observations longitudinales analysées par modèle mixte fournissent un moyen d'estimer la densité f_ε, dès lors que l'on dispose de plus de 6 observations par individus ($J \geq 6$), obtenues à temps réguliers ($\Delta_j = \Delta$). Il faut également faire une hypothèse supplémentaire sur la loi du bruit :

(A') la loi du bruit f_ε vérifie l'hypothèse **(A)** et est symétrique.

L'hypothèse **(A')** implique que f_ε^* est réelle et garde un signe constant. Comme $f_\varepsilon^*(0) = 1$, on en déduit que $f_\varepsilon^*(u) > 0$ pour tout u. Cette propriété permet d'estimer f_ε^* à partir d'un estimateur de $(f_\varepsilon^*)^2$ ou $(f_\varepsilon^*)^4$. Cette hypothèse est plus simple que celle proposée par Delaigle *et al.* (2008).

Plus précisément, l'estimateur de f_ε^* est obtenu par une transformation des observations. On introduit les variables aléatoires pour tout $k = 1, \ldots, N$

$$\begin{aligned} W_k &= Z_{k2} - Z_{k1} \\ &= \beta_k + \frac{\varepsilon_{k4} - \varepsilon_{k3}}{\Delta} - \beta_k - \frac{\varepsilon_{k2} - \varepsilon_{k1}}{\Delta} = \frac{1}{\Delta}(\varepsilon_{k4} - \varepsilon_{k3} - \varepsilon_{k2} + \varepsilon_{k1}). \end{aligned}$$

On peut alors estimer $(f_\varepsilon^*)^4$ par

$$\widehat{(f_\varepsilon^*)^4}(u/\Delta) = \frac{1}{N}\sum_{k=1}^{N} \cos(uW_k).$$

On tronque cet estimateur pour le substituer dans les estimateurs de f_β et f_α comme l'ont proposé Neumann (1997) et Comte et Lacour (2011). On note $\widetilde{(f_\varepsilon^*)^2}$ l'estimateur tronqué de $(f_\varepsilon^*)^2$:

$$\frac{1}{\widetilde{(f_\varepsilon^*)^2}(u)} = \frac{\mathbb{1}_{\widehat{(f_\varepsilon^*)^4}(u) \geq N^{-1/2}}}{\left[\widehat{(f_\varepsilon^*)^4}(u)\right]^{1/2}}.$$

Nous estimons $f_{Z_j}^*$ avec les observations Z_{kj} pour $j \geq 3$. L'estimateur de f_β quand f_ε est inconnu est alors

$$\widetilde{f_{\beta,m}}(x) = \frac{2}{N(J-4)}\sum_{k=1}^{N}\sum_{j=3}^{J/2}\frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{e^{-iu(x-Z_{kj})}}{\widetilde{(f_\varepsilon^*)^2}(u/\Delta)}du. \tag{1.12}$$

En étudiant l'erreur introduite par la troncation, on peut montrer la proposition suivante. On note pour toute fonction f intégrable et de carré intégrable et pour $\ell \in \mathbb{N}$

$$D_\ell(m, f) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{|f^*(u)|^2}{\left(f_\varepsilon^*(u/\Delta)\right)^{2\ell}}du. \tag{1.13}$$

***Proposition* 2.** *Supposons que f_β est intégrable et de carré intégrable, alors l'estimateur (1.12) vérifie*

$$\mathbb{E}(\|\widetilde{f_{\beta,m}} - f_\beta\|^2) \leq \|f_{\beta,m} - f_\beta\|^2 + 16\frac{D_2(m,1)}{N(J-4)} + 4C_0\left(\frac{D_2(m,f_\beta)}{\sqrt{N}}\right) \wedge \left(\frac{D_4(m,f_\beta)}{N}\right) + 6\frac{m}{N}, \tag{1.14}$$

où C_0 est une constante.

Le terme $4C_0\left(\frac{D_2(m,f_\beta)}{\sqrt{N}}\right)\wedge\left(\frac{D_4(m,f_\beta)}{N}\right)+6\frac{m}{N}$ vient de l'estimation de f_ε, les autres termes apparaissent dans la borne de l'estimateur $\widehat{f_{\beta,m}}$ quand f_ε est connue. Nous proposons une méthode de sélection du seuil m similaire à celle proposée lorsque f_ε est connue, mais où la pénalité est inspirée de la borne (1.14). La pénalité théorique fait intervenir le terme $D_2(m, f_\beta)$, qu'il faut estimer en approchant numériquement l'intégrale (1.13). Les détails sont donnés dans **[S1]**. Un estimateur similaire est proposé pour f_α.

Une étude de simulation compare les propriétés de l'ensemble de ces estimateurs, avec un bruit f_ε gaussien ou Laplace, des densités des variables individuelles f_β et f_α gaussiennes, mélange de gaussiennes, gamma ou mélange de gamma. L'estimateur de f_α basé sur les observations mesurées au temps 0 est toujours meilleur que les autres, que la densité du bruit soit connue ou non. L'estimation de f_β s'est révélée très stable et performante, en particulier pour les densités bimodales. L'estimation de la densité du bruit a tendance à améliorer les propriétés des estimateurs de f_β et f_α. Ceci avait déjà été noté dans Comte et Lacour (2011).

Ce travail constitue à ma connaissance la première utilisation des outils de déconvolution pour définir des estimateurs des densités des effets aléatoires dans le cadre des modèles mixtes. Ces estimateurs sont prometteurs et plusieurs perspectives de recherche sont évoquées à la fin de ce manuscrit.

1.3.2 Cas d'un sous-échantillon de mesures répétées à $t = 0$

Le travail précédent montre que le meilleur estimateur de f_α est celui basé sur les observations à $t = 0$, que la densité f_ε soit connue ou inconnue. Dans le cas où f_ε est inconnue, nous étendons ce travail dans **[S4]** en supposant que deux échantillons d'observations à $t = 0$ sont disponibles : un premier échantillon de taille N d'observations en $t = 0$, noté

$$Y_{k0} = \alpha_k + \varepsilon_{k0}, \quad k = 1, \ldots, N \tag{1.15}$$

et un échantillon de taille M de mesures répétées au temps $t = 0$ noté

$$\widetilde{Y}_{k0}^{(1)} = \alpha_k + \varepsilon_{k0}^{(1)}, \quad \widetilde{Y}_{k0}^{(2)} = \alpha_k + \varepsilon_{k0}^{(2)}, \quad k = 1, \ldots, M. \tag{1.16}$$

On suppose que les deux échantillons sont indépendants. On suppose que les erreurs résiduelles ε_{k0}, $\varepsilon_{k0}^{(1)}$, $\varepsilon_{k0}^{(2)}$ des deux échantillons ont la même densité f_ε.

Ce problème rentre dans le cadre d'une variable mesurée avec erreur, dont on cherche à estimer la densité à partir d'un échantillon de mesures répétées. Il a été étudié par Delaigle *et al.* (2008) et Meister et Neumann (2010) mais leurs estimateurs sont limités au cas d'un bruit "ordinary smooth". Nous proposons dans **[S4]** un estimateur adapté au cas d'un bruit "super smooth", qui est inspiré des travaux de Comte et Lacour (2011) et de **[S1]**.

Dans le travail sur les modèles mixtes, nous estimons f_ε^* via les observations répétées dans le temps. Dans **[S4]**, nous proposons d'utiliser les observations du second échantillon de données répétées pour estimer f_ε^*. On se place à nouveau sous l'hypothèse **(A')** qui assure que f_ε^* est strictement positif. On propose alors d'estimer $(f_\varepsilon^*)^2$ par la quantité suivante

$$\widehat{(f_\varepsilon^*)^2}(u) = \frac{1}{M}\sum_{k=1}^{M}\cos\left(u(\widetilde{Y}_{k0}^{(1)} - \widetilde{Y}_{k0}^{(2)})\right).$$

On définit ensuite un estimateur tronqué de $1/f_\varepsilon^*$:

$$\frac{1}{\tilde{f}_\varepsilon^*(u)} = \frac{\mathbb{1}_{\widehat{(f_\varepsilon^*)^2}(u)\geq M^{-1/2}}}{\sqrt{\widehat{(f_\varepsilon^*)^2}(u)}}.$$

L'estimateur de f_α peut être construit à partir des deux échantillons en introduisant un seuil πm

$$\widetilde{f^{(0)}_{\alpha,m}}(x) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} e^{-ixu}\frac{\hat{f}^*_Y(u)}{\tilde{f}^*_\varepsilon(u)}du, \text{ où } \hat{f}^*_Y(u) = \frac{1}{N+M}\left(\sum_{k=1}^{N} e^{iuY_{k0}} + \sum_{k=1}^{M} e^{iu\widetilde{Y}^{(1)}_{k0}}\right). \tag{1.17}$$

Les données $\widetilde{Y}^{(2)}_{k0}$ ne sont pas utilisés pour l'estimation de $f^*_Y(u)$ contrairement à l'estimateur proposé par Delaigle *et al.* (2008). En effet nous montrons que l'utilisation des $N+2M$ variables disponibles détériore la variance de l'estimateur de $f^*_Y(u)$ à cause de la dépendance entre $\widetilde{Y}^{(1)}_{k0}$ et $\widetilde{Y}^{(2)}_{k0}$. La principale différence avec l'estimateur proposé dans **[S1]** réside dans la façon d'estimer la densité f_ε. Ici, on estime $(f^*_\varepsilon)^2$ alors qu'on estime $(f^*_\varepsilon)^4$ dans le cas des données longitudinales. De plus, on utilise un échantillon de taille M ici, qui est indépendant de l'échantillon initial de taille N.

La borne du risque moyen intégré de $\widetilde{f^{(0)}_{\alpha,m}}$ fait intervenir la quantité

$$\tilde{D}_\ell(m,f) = \int_{-\pi m}^{\pi m} \frac{|f^*(u)|^2}{(f^*_\varepsilon(u))^\ell}du \quad \text{pour } \ell \in \mathbb{N}^*.$$

Cette quantité est différente de (1.13) par le fait que Δ n'apparaît pas dans le dénominateur. De plus la puissance de $f^*_\varepsilon(u)$ est ℓ au lieu de 2ℓ puisqu'on estime $(f^*_\varepsilon)^2$ au lieu de $(f^*_\varepsilon)^4$. On peut alors montrer l'équivalent de la proposition 2 pour l'estimateur (1.17) :

***Proposition* 3.** *Il existe une constante C telle que*

$$\mathbb{E}(\|f_\alpha - \widetilde{f^{(0)}_{\alpha,m}}\|^2) \le \|f_\alpha - f_{\alpha,m}\|^2 + C\left(\frac{\tilde{D}_2(m,1)}{N+M} + \frac{\tilde{D}_2(m,f_\alpha)}{\sqrt{M}} \wedge \frac{\tilde{D}_4(m,f_\alpha)}{M}\right). \tag{1.18}$$

Outre le terme de biais $\|f_\alpha - f_{\alpha,m}\|^2$, on reconnait le terme de variance $\tilde{D}_2(m,1)/(N+M)$, qui est usuel dans la borne du MISE quand f^*_ε est connu. Le terme $\tilde{D}_2(m,f_\alpha)/\sqrt{M} \wedge \tilde{D}_4(m,f_\alpha)/M$ est spécifique aux données répétées. Il est analogue au terme qui apparaît dans (1.14) où $N+M$ joue le rôle de $N(J-4)$, le nombre total d'observations utilisées pour estimer le numérateur, et M joue le rôle de N dans le nombre de paires d'observations utilisées pour estimer la densité f_ε.

Nous étudions plus précisément ces bornes dans le cas de densités f_α et f_ε "ordinary smooth". On reprend les notations $\delta, c_\varepsilon, C_\varepsilon$ du théorème 2 pour la régularité de f^*_ε. Nous supposons aussi qu'il existe $a > 1/2$, $\ell > 0$ tels que

$$\int |f^*_\alpha(ux)|^2(u^2+1)^a du \le \ell.$$

Nous montrons que la borne du MISE (1.18) est alors

$$\mathbb{E}(\|f_\alpha - \widetilde{f^{(0)}_{\alpha,m}}\|^2) \le C\left(m^{-2a} + \frac{m^{2\delta+1}}{N+M} + \frac{m^{2(\delta-a)_+}}{\sqrt{M}} \wedge \frac{m^{2(2\delta-a)_+}}{M}\right).$$

Alors si $M \ge N$ et $a \ge \delta - 1/2$, le choix optimal de seuil $m = m_{opt} = M^{1/(2a+2\delta+1)}$ fournit un estimateur tel que

$$\mathbb{E}(\|f_\alpha - \hat{f}_{\alpha,m_{opt}}\|^2) \le CM^{-2a/(2a+2\delta+1)}.$$

Si $M = N^\omega$ avec $\omega < 1$ et $2\delta \le a \le \frac{\omega}{1-\omega}(\delta + \frac{1}{2})$, le choix de seuil $m_{opt} = N^{1/(2a+2\delta+1)}$ fournit un estimateur tel que

$$\mathbb{E}(\|f_\alpha - \hat{f}_{\alpha,m_{opt}}\|^2) \le CN^{-2a/(2a+2\delta+1)}.$$

Ces vitesses sont usuelles pour les estimateurs de déconvolution à bruit connu, avec des échantillons de taille M et N respectivement. Cette vitesse est optimale quand le bruit est connu (Butucea et

Tsybakov, 2008). Lorsque le bruit est inconnu, le cas $M \geq N$ a déjà été mentionné par Delaigle *et al.* (2008). Le cas $M < N$ est nouveau et intéressant puisqu'il correspond à de nombreux cas de données biomédicales (voir la section 1.4). Une étude est aussi réalisée lorsque le bruit est "supersmooth".

Une étude par simulation est réalisée avec différents types de bruit (gaussien ou Laplace), différentes densités f_α (mélange de Gamma, Cauchy, Laplace ou gaussienne), différents rapports signal sur bruit et des échantillons de taille $M = N$ et $M = \sqrt{N}$, pour $N = 200$ ou $N = 2000$. Le risque des estimateurs décroît quand N ou M croît ou quand le rapport signal sur bruit croît. L'estimateur a un risque diminué quand l'erreur est Laplace par rapport à une erreur gaussienne. Enfin, une comparaison à l'estimateur proposé par Delaigle *et al.* (2008) semble montrer que notre estimateur a un risque nettement inférieur.

1.4 Applications en biologie

Je présente dans cette section deux collaborations avec des médecins et biologistes sur différents projets où l'outil des modèles mixtes est approprié pour répondre à la question posée. J'ai participé à d'autres projets mais qui ne sont pas détaillés, j'en donne une très brève description ci-dessous.

J'ai travaillé avec Thierry Bastogne, Muriel Barberi-Heyob (Centre de Recherche en Automatique de Nancy), Pierre Valois, Sophie Wantz-Mézières (Institut de Mathématiques Elie Cartan, Nancy) et Sophie Pinel (Laboratoire Signalisation, Génomique et Recherche Transactionnelle en Oncologie, Nancy) sur la comparaison d'efficacité de traitements anti-cancéreux par modélisation de croissance tumorale **[A9]**. L'approche par modèle mixte y est pragmatique, avec un modèle de croissance et décroissance tumorale le plus simple possible.

En collaboration avec Solange Whegang (Laboratoire MAP5, Université Paris Descartes ; Laboratoire de Recherche sur le Paludisme, OCEAC, Yaoundé Cameroun), Leonardo Basco (IRD Université Marseille ; OCEAC, Yaoundé Cameroun), et Jean-Christophe Thalabard (Laboratoire MAP5, Université Paris Descartes), nous avons proposé une comparaison d'efficacité de traitements anti-paludiques à partir de données catégorielles répétées **[A16]**. L'originalité de ce travail est d'analyser conjointement les données de cinq essais cliniques, dans lesquels les traitements comparés sont différents. Pour ce faire, nous utilisons une approche de méta-analyse et de comparaison de traitements mixtes.

Enfin, dans le cadre de la collaboration avec Charles-André Cuenod (Service de radiologie, Hopital Européen Georges Pompidou, Paris), collaboration qui est détaillée dans la section 2.1, nous avons proposé avec Frédéric Richard (Laboratoire MAP5 et LATP, Université de Provence) une méthode de recalage d'une séquence d'images médicales enregistrées dans le temps. La méthode que nous développons est basée sur les modèles mixtes et l'algorithme SAEM **[A8]**.

La première application présentée dans cette section concerne la modélisation de la dynamique virale du VIH **[A12]**. Elle a été réalisée en collaboration avec Marc Lavielle, Ana Karina Fermin (Université de Nanterre) et France Mentré et est décrite dans la section 1.4.1. Elle a motivé les développements méthodologiques présentés dans les sections 1.2.2 et 1.2.3.

La deuxième application détaillée résulte d'une collaboration avec Julien Stirnemann qui est gynécologue à l'hôpital Necker, spécialiste dans le suivi de grossesses gémellaires et que je co-encadre en thèse avec Jean-Christophe Thalabard. Différentes questions méthodologiques se posent dans ce contexte, dont deux sont présentées dans la section 1.4.2. Un premier travail a consisté à développer un outil de prédiction individuelle de croissance foetale **[A14]**. Ensuite, nous avons étudié l'estimation de la densité de début de grossesse dans le cycle menstruel d'une femme. Ce travail a été réalisé en collaboration avec Fabienne Comte **[S4, A15]**.

1.4.1 Comparaison de l'efficacité de traitements contre le VIH

L'étude mathématique de la dynamique du VIH a permis depuis plusieurs années de mieux comprendre les mécanismes de l'infection. De nombreux modèles ont été proposés, qui décrivent la décroissance de la charge virale et l'augmentation du nombre de cellules lymphocytaires CD4 lorsqu'un

patient prend un traitement anti-rétroviral (Perelson et Nelson, 1997; Nowak et May, 2000). Ils ont permis d'améliorer la compréhension des différents mécanismes impliqués dans cette pathologie. Un des prochains défis est de mieux appréhender la variabilité dans la réponse aux traitements anti-rétroviraux parmi les patients. L'outil des modèles mixtes apparaît alors pertinent pour répondre à cette question mais le défi statistique est grand. En effet, la mesure de la charge virale est sujette à une limite de quantification (LOQ), problème statistique déjà évoqué dans la section 1.2.2. De plus, les premières analyses de données longitudinales de décroissance de charge virale utilisaient des modèles dynamiques simplifiés où la fonction de régression du modèle mixte est explicite (Ding et Wu, 2001). Plus récemment, des modèles plus complexes modélisant la dynamique à long terme ont été considérés (voir Guedj *et al.* (2007) et les références internes). Ces systèmes de dimension supérieure à 4 n'ont pas de solution analytique et sont partiellement observés (observation de dimension 2). On fait alors face au problème d'une fonction de régression solution d'une EDO, problème évoqué dans la section 1.2.3.

Dans **[A12]**, nous proposons d'utiliser l'algorithme SAEM-MCMC adapté aux données censurées et aux systèmes d'EDO et qui est implémenté dans le logiciel MONOLIX©. Cet algorithme, performant numériquement, nous permet d'envisager plusieurs modèles dynamiques complexes avec un nombre élevé d'effets aléatoires et d'en estimer les paramètres sur les données de l'essai COPHAR2-ANRS 111. En général, le coût numérique et les problèmes de convergence des méthodes d'estimation des paramètres d'un modèle mixte conduisent à limiter l'étude des données réelles à un seul modèle et un nombre réduit d'effets aléatoires. C'est à ma connaissance une des premières comparaisons de modèles dynamiques du VIH réalisée sur données réelles. Cette démarche peut permettre à l'avenir de valider différentes hypothèses émises par les biologistes sur le mécanisme de la dynamique virale.

Je détaille maintenant les modèles comparés. Les données de l'essai COPHAR2-ANRS111 sont présentées ensuite.

Nous considérons trois systèmes différentiels ordinaires non-linéaires de la dynamique de VIH. Le plus simple, noté $\mathcal{M}_B$, a 4 composantes : les cellules CD4 non infectées, les cellules CD4 infectées, les virus infectants et les virus non-infectants. Le second modèle, noté $\mathcal{M}_Q$, différencie les CD4 non-infectées entre une catégorie active et une catégorie passive. Enfin, le troisième modèle, noté $\mathcal{M}_L$, différencie les CD4 infectées entre une catégorie de latence et une catégorie active. Le modèle le plus simple est détaillé ci-dessous. On renvoie à **[A12]** pour une description précise des deux autres.

On note T_{NI}, T_I, V_I et V_{NI} les concentrations de CD4 non infectées et infectées, de virus infectants et non infectants. On suppose que les CD4 sont générées à un taux constant λ. Les CD4 sont infectées par les virus infectants à un taux γ par cellule susceptible (non infectée) et par virus. Les CD4 non infectées et infectées meurent à un taux μ_{NI} et μ_I respectivement. Les CD4 infectées produisent des virus à un taux p par cellule infectée. Les virus meurent à un taux μ_V. Deux paramètres η_{RTI} et η_{PI} sont introduits pour décrire l'effet d'un traitement ant-rétroviral. Les molécules RTI ont pour but d'empêcher les CD4 non infectées de s'infecter en inhibant la transcriptase de l'ARN viral en de l'ADN. Le paramètre η_{RTI} représente donc la proportion de CD4 non infectées le restant après contact avec un virus, il prend ses valeurs entre 0 et 1. Les molécules PI ont pour but de produire des virus non infectants au lieu de virus infectants. Le taux de production de ces virus non infectants est $\eta_{\mathrm{PI}}\, p$, η_{PI} étant compris entre 0 et 1. Plus η_{RTI} et η_{PI} sont proches de 1, plus l'efficacité des molécules est grande. Lorsqu'un traitement anti-rétroviral comprend à la fois des PI et des RTI, la dynamique virale peut être modélisée par le modèle $\mathcal{M}_B$ suivant :

$$\frac{dT_{NI}}{dt} = \lambda - (1-\eta_{\mathrm{RTI}})\gamma\, T_{NI}V_I - \mu_{NI}T_{NI}, \qquad \frac{dV_{NI}}{dt} = \eta_{\mathrm{PI}}\, p\, T_I - \mu_V V_{NI},$$
$$\frac{dT_I}{dt} = (1-\eta_{\mathrm{RTI}})\gamma\, T_{NI}V_I - \mu_I T_I, \qquad \frac{dV_I}{dt} = (1-\eta_{\mathrm{PI}})p\, T_I - \mu_V V_I.$$

La condition initiale du système est calculée sous l'hypothèse qu'avant l'initiation du traitement, le système est à l'équilibre. Les mesures dont on dispose correspondent à la charge virale totale, $V = V_I + V_{NI}$ et le nombre total de CD4, $T = T_{NI} + T_I$. Le système étant non-linéaire en T_{NI} et V_I, les observations de la charge virale et du taux de CD4 doivent être analysées conjointement.

Pour l'individu $k = 1, \ldots, N$, on note $V_k = (V_{k1}, \ldots, V_{kJ_k})$ et $Z_k = (Z_{k1}, \ldots, Z_{kJ'_k})$ les vecteurs des $\log_{10}$ charge virale (cp/ml) et de taux de CD4 (cells/mm^3) mesurés aux temps $(t_{kj})_{1\leq j\leq J_k}$ et $(\tau_{kj})_{1\leq j\leq J'_k}$ respectivement. Le modèle mixte proposé est le suivant

$$\begin{aligned} V_{kj} &= \log_{10}(1000\, V(t_{kj};\psi_k)) + \varepsilon_{V,kj}, \quad \varepsilon_{V,k} \sim \mathcal{N}(0, \sigma_V^2 I_{J_k}), \\ Z_{kj} &= T(\tau_{kj};\psi_k) + T(\tau_{kj};\psi_i)\, \varepsilon_{T,kj}, \quad \varepsilon_{T,k} \sim \mathcal{N}(0, \sigma_T^2 I_{J'_k}), \\ \psi_k &= h(\phi_k), \quad \phi_k \sim \mathcal{N}(\mu, \Omega), \end{aligned}$$

où V et T sont solutions du système différentiel considéré, $\varepsilon_{V,k}$ et $\varepsilon_{T,k}$ sont les vecteurs d'erreurs résiduelles, ψ_k sont les paramètres individuels du modèle, obtenus par une transformation $h(\phi_k)$ de vecteurs gaussiens. La charge virale est mesurée avec une limite de détection LOQ, et on dispose uniquement des observations

$$\delta_{kj} = \mathbb{1}_{V_{kj}\geq LOQ} \quad \text{et} \quad V_{kj}^{obs} = \begin{cases} V_{kj} & \text{si} \quad \delta_{kj} = 1, \\ LOQ & \text{si} \quad \delta_{kj} = 0. \end{cases}$$

L'estimation des paramètres $\theta = (\mu, \Omega, \sigma_V^2, \sigma_T^2)$ est réalisée via le logiciel MONOLIX, qui combine les versions de l'algorithme SAEM-MCMC proposées pour le traitement des données censurées (section 1.2.2) et des fonctions de régression définies comme solution d'EDO (section 1.2.3). La comparaison des 3 modèles est réalisée grâce au critère d'information bayésien BIC. Pour un modèle $\mathcal{M}$, il est défini par

$$BIC(\mathcal{M}) = -2\log L_{\mathcal{M}}(V^{obs}, Z; \widehat{\theta}_{\mathcal{M}}) + \log(N) P_{\mathcal{M}}$$

où $L_{\mathcal{M}}(V^{obs}, Z; \widehat{\theta}_{\mathcal{M}})$ est la vraisemblance du modèle et $P_{\mathcal{M}}$ le nombre de paramètres du modèle. Le meilleur modèle est défini comme étant celui ayant le plus petit BIC. La vraisemblance $L_{\mathcal{M}}(V^{obs}, Z; \widehat{\theta}_{\mathcal{M}})$ est estimée par échantillonnage préférentiel. L'effet de différentes covariables est testé par le test de Wald et critère BIC.

Les données analysées sont issues de l'essai COPHAR2-ANRS 111 (Duval *et al.*, 2009). C'est un essai prospectif non-randomisé, au cours duquel 115 patients infectés par le VIH ont reçu un traitement anti-rétroviral comprenant 2 inhibiteurs de la transcriptase inverse (RTI) et un inhibiteur de la protease (PI) parmi l'indinavir, le lopinavir et le nelfinavir : 48 patients ont été traités avec de l'indinavir, 38 avec du lopinavir et 35 avec du nelfinavir. Le but de l'étude est de comparer l'efficacité des 3 traitements. Les patients sont suivis pendant un an, avec des visites aux semaines 0, 2, 8, 16, 24, 36 et 48 au cours desquelles la charge virale et le taux de CD4 sont mesurés. Les données sont représentées sur la figure 1.1. On remarque une grande variabilité dans la réponse au traitement entre les patients.

Après analyse des données de l'essai COPHAR2-ANRS111 par le logiciel MONOLIX, le meilleur modèle est le modèle $\mathcal{M}_L$. Les efficacités η_{PI} du lopinavir et de l'indinavir sont estimées à 0.99, alors que celle du nelfinavir est estimée à 0.75, la différence étant significative (p-valeur = 10^{-12}). Ceci confirme les premières analyses de Duval *et al.* (2009) à partir du taux d'échec viral. Les trajectoires individuelles de charge virale et de taux de CD4 sont bien estimées par ce modèle (figure 1.2). Cette analyse est une des premières par maximum de vraisemblance dans le contexte de la dynamique virale du VIH, les autres références dans le domaine utilisent en général une estimation par approche bayésienne et des algorithmes MCMC extrêmement coûteux en temps de calcul.

La principale limite de cette étude concerne le choix des modèles comparés. On a supposé que l'effet des traitements était constant dans le temps. En réalité, l'efficacité des traitements est fonction de la concentration du médicament dans le plasma, concentration qui évolue dans le temps. Des modèles prenant en compte cette évolution ont été proposés. Cependant ils peuvent faire face à des problèmes d'identifiabilité numérique sur les données disponibles, car seuls la charge virale et le taux de CD4 sont observables. Une alternative consiste à proposer des modèles stochastiques, en considérant que certains paramètres du modèle sont aléatoires. Cette approche est abordée dans le chapitre 2.

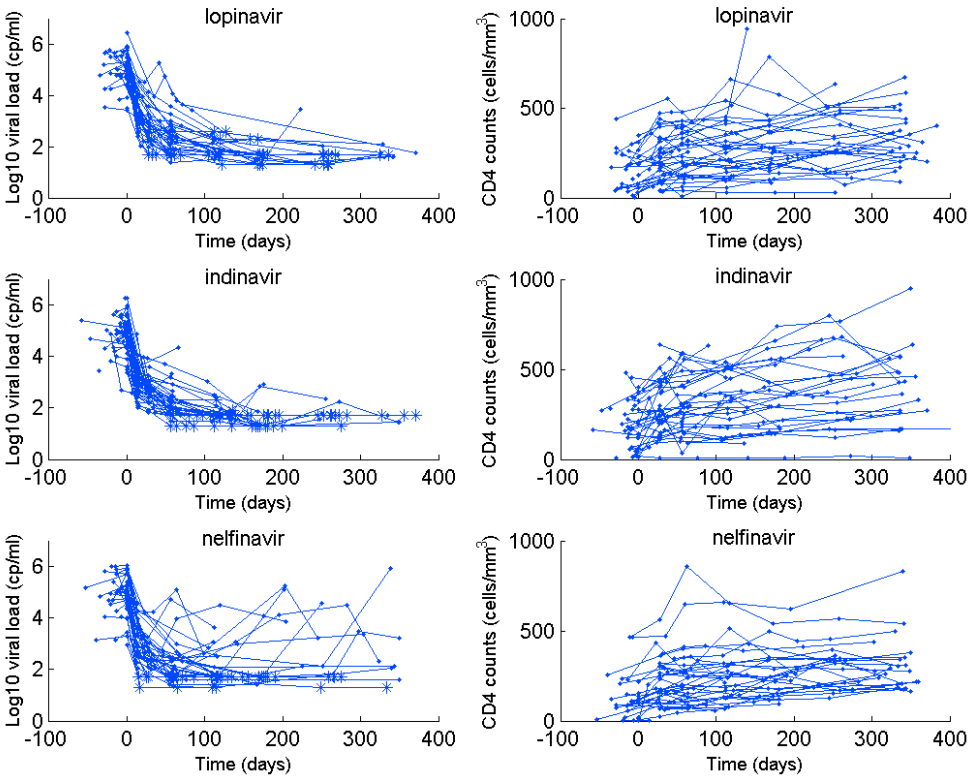

FIG. 1.1 – Décroissance des charges virales (à gauche) et croissance des CD4 (à droite) après l'initiation du traitement antiviral dans les trois groupes : lopinavir (haut), indinavir (milieu) et nelfinavir (bas) des patients de l'essai COPHAR2-ANRS 111.

1.4.2 Prédiction de la croissance foetale

Dans cette section, je présente les travaux issus de la collaboration avec Julien Stirnemann (Laboratoire MAP5, Département d'obstétrique de l'hôpital Necker, Paris), réalisés avec Jean-Christophe Thalabard et Fabienne Comte.

Dans **[A14]**, nous considérons la problématique de la prédiction individuelle de la croissance foetale, en particulier pour les grossesses gémellaires qui sont plus à risque. En général, les courbes de croissance auxquelles se réfèrent les médecins ont été construites à partir d'études transversales sur une population d'individus. Les percentiles de référence estimés à deux âges distincts ne sont donc pas construits avec la même population d'individus. La croissance étant un phénomène dynamique, il est plus précis de construire ces courbes de croissance à partir d'une étude longitudinale. Par ailleurs, dans le contexte particulier de la croissance foetale, un outil de suivi individuel paraît plus approprié. En effet, si un foetus a une petite morphologie, il est plus judicieux de comparer sa courbe de croissance à une courbe "individuelle" plutôt qu'à une courbe "populationnelle". Notre objectif est de construire un intervalle de prédiction individuel correspondant à la taille attendue lors d'une future visite, étant données les observations déjà obtenues précédemment pour ce foetus. Nous proposons de construire cette prédiction individuelle à partir des données individuelles et d'un modèle populationnel construit au préalable, en se plaçant dans une approche bayesienne. Pour cela, nous construisons d'abord un modèle mixte de croissance foetale chez des jumeaux. Les paramètres de population ainsi estimés permettent de définir la loi a priori. Puis, à l'aide d'un algorithme MCMC bayesien, nous estimons l'intervalle de prédiction. Cet outil se révèle tout à fait pertinent pour détecter des "anomalies" de croissance foetale.

Plus précisément, la première étape consiste à construire le modèle mixte de croissance foetale chez des jumeaux à partir de données longitudinales d'une population de référence (croissance considérée

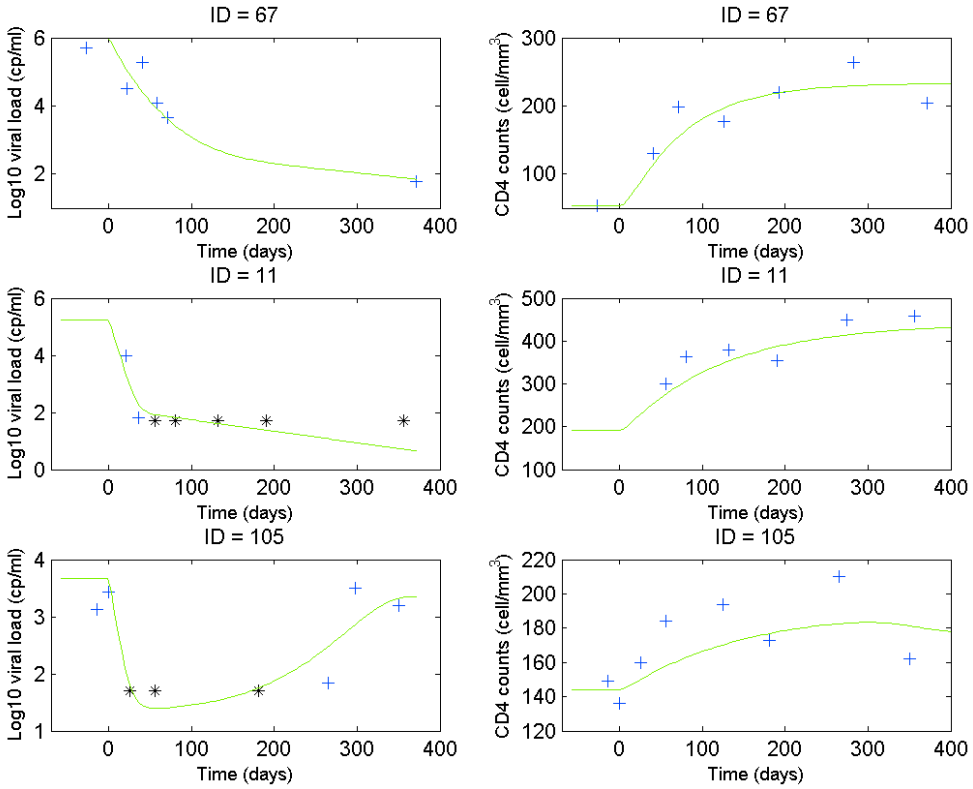

FIG. 1.2 – Données de l'essai COPHAR2-ANRS 111. Exemples de trajectoires individuelles estimées par le modèle mixte $\mathcal{M}_L$ via l'algorithme SAEM-MCMC (charge virale à gauche et taux de CD4 à droite) pour trois patients : patient 67 (groupe lopinavir), patient 11 (groupe indinavir) et patient 105 (groupe nelfinavir). Le symbole + représente les observations non censurées et le symbole $*$ la limite de quantification.

comme "normale"). La croissance des deux jumeaux étant hautement corrélée, nous considérons un modèle mixte à deux niveaux de variabilité. L'analyse statistique est réalisée avec l'algorithme SAEM-MCMC proposé dans **[A7]** pour les modèles mixtes à plusieurs niveaux de variabilité.

Ensuite, à partir de paramètres de population θ et du modèle de croissance construit et validé sur un jeu de données externe, nous proposons la construction des intervalles de prédiction individuels de la taille de deux foetus d'une nouvelle patiente. On suppose que la croissance de ces jumeaux a le même comportement que la population de référence utilisée pour construire le modèle mixte. On considère qu'au temps t_n, n mesures ont déjà été obtenues pour ces deux jumeaux. On note Y_{kj} la mesure au temps t_{kj} pour $j = 1 \dots n$ du k-ième jumeau ($k = 1, 2$) de cette grossesse. On note $Y_{k1:n} = (Y_{k1}, \dots, Y_{kn})$ le vecteur d'observation du jumeau k. On note $Y_{1:n} = (Y'_{11:n}, Y'_{21:n})'$ le vecteur d'observations disponibles au temps t_n pour cette grossesse. On note ϕ_k le vecteur de paramètre du jumeau k et $\phi = (\phi'_1, \phi'_2)'$ le vecteur pour les deux jumeaux.

Le but est de proposer des intervalles de prédiction pour $\phi = (\phi'_1, \phi'_2)'$ et pour un vecteur d'observations futures $Y_t = (Y_{1t}, Y_{2t})$ au temps $t > t_n$, conditionnellement aux données $Y_{1:n}$. Pour la prédiction de ϕ, on s'intéresse à la distribution a posteriori $p(\phi|Y_{1:n}; \theta)$ où $p(\phi, \theta)$ est la distribution a priori construite sur la population de référence par modèle mixte. Grâce au théorème de Bayes, la distribution a posteriori $p(\phi|Y_{1:n}; \theta)$ s'écrit

$$p(\phi|Y_{1:n}; \theta) = \frac{p(Y_{1:n}|\phi; \theta)p(\phi; \theta)}{L(Y_{1:n}; \theta)},$$

où $L(Y_{1:n}; \theta) = \int p(Y_{1:n}, \phi; \theta)d\phi$ est la vraisemblance des données $Y_{1:n}$ et $p(Y_{1:n}|\phi; \theta) = \prod_{j=1}^{n} p(Y_j|\phi; \theta)$

par indépendance des mesures conditionnellement à ϕ. Pour la prédiction d'une future observation Y_t, on considère la distribution prédictive

$$p(Y_t|Y_{1:n};\theta) = \int p(Y_t|\phi;\theta)p(\phi|Y_{1:n};\theta)d\phi.$$

Nous nous concentrons sur l'estimation des espérances conditionnelles $\mathbb{E}(\phi|Y_{1:n};\theta)$ et $\mathbb{E}(Y_t|Y_{1:n};\theta)$ de ces deux distributions et des intervalles de prédiction au niveau $(1-\tau)$ de $\phi|Y_{1:n}$ et $Y_t|Y_{1:n}$. Ces intervalles de prédiction sont basés sur le calcul de quantiles d'une fonctionnelle de ϕ qui sont estimés par algorithme MCMC. Les espérances comme les intervalles de prédiction sont basés sur les données des deux jumeaux et prennent en compte leur corrélation.

Cette méthode est utilisée sur deux types de données recueillies en France, des données issues de 54 croissances gémellaires considérées comme "normales", et des données issues de grossesses avec des croissances pathologiques pour un des jumeaux qui ont nécessité le déclenchement précoce de l'accouchement. Pour les données de croissance "normale", nous disposons de 4 mesures, réalisées tous les mois à partir du 3ème mois de grossesse. Nous cherchons à prédire la 4ème mesure (6ème mois) à partir des mesures précédentes et à la comparer à la vraie mesure. Des intervalles de prédiction à 50% et 90% sont estimés pour les 54 foetus. La proportion de vraies mesures appartenant à ces intervalles de prédiction est de 53% et 94% respectivement, montrant la pertinence de l'outil proposé. Les intervalles individuels sont de plus faible amplitude que les intervalles de population, et peuvent parfois ne pas être inclus dans ces intervalles de population. Pour les données de croissance "anormale", nous considérons les estimations successives de $\mathbb{E}(\phi|Y_{1:j};\theta)$ pour $j = 1,\ldots,n$. A chaque visite j, une nouvelle observation (Y_{1j}, Y_{2j}) est disponible et permet d'estimer une nouvelle valeur de $\mathbb{E}(\phi|Y_{1:j};\theta)$. Nous comparons les prédictions obtenues pour une grossesse à croissance "normale", où l'on dispose de $n = 4$ visites aux temps $t_{1:4} = (21, 26, 32, 34.7)$ semaines, et une grossesse à croissance "anormale" où l'on dispose de $n = 9$ visites aux temps $t_{1:9} = (10.3, 15.6, 17, 19.4, 21.4, 23.3, 25.6, 28.6, 30.3)$. Sur la figure 1.3, colonne de gauche, sont représentés les observations de ces deux grossesses ainsi que les 5ème et 95ème percentiles de population estimés à partir de la population de référence par modèle mixte. Sur la colonne de droite de la figure, nous proposons une représentation graphique dans l'espace des paramètres renormalisés ϕ des estimations de $\mathbb{E}(\phi|Y_{1:j};\theta)$ pour $j = 1,\ldots n = 4$ (grossesse "normale") et $j = 1,\ldots n = 9$ (grossesse "anormale"). Les ellipses de confiance de population calculées à partir du modèle mixte sont également représentées. On voit que pour la grossesse à croissance "normale", les valeurs de $\hat{\phi}$ restent dans les ellipses de référence. Ce n'est pas le cas pour le jumeau dont la croissance est "anormale". Ceci peut fournir un outil de suivi individuel de croissance foetale, même s'il est crucial de rester prudent sur l'implication clinique d'un tel outil.

La prédiction individuelle de la croissance foetale présentée précédemment n'a de sens que si la date de début de grossesse est connue avec précision. En effet, si une erreur de quelques jours est faite sur cette date, on peut interpréter à tort la croissance observée comme une croissance anormale. Actuellement, la date de début de grossesse est estimée à partir des mesures faites à la première échographie. Un modèle de régression polynomiale est ensuite utilisé pour prédire la date de début de grossesse. Cependant, ces modèles de régression ne tiennent pas compte de la date des dernières règles. L'enjeu du projet est donc de proposer une estimation individuelle de la date de début de grossesse tenant compte à la fois de la mesure obtenue à la première échographie et de la date des dernières règles.

Une étape préliminaire est d'estimer la loi de probabilité de l'intervalle entre la date des dernières règles (DDR) et le début de grossesse chez les femmes enceintes. Elle fait l'objet du papier **[A15]**, décrit ci-dessous. Les publications dans le domaine estiment en général la loi de probabilité de l'intervalle entre DDR et le moment de l'ovulation dans un cycle (Wilcox *et al.*, 2000). Cela suppose de suivre de façon intensive une population de femmes en utilisant des techniques particulières pour mesurer l'instant de l'ovulation. A contrario, notre approche est basée sur la mesure obtenue à la première échographie de grossesses dans une population générale, dans le cadre d'une étude transversale de suivi

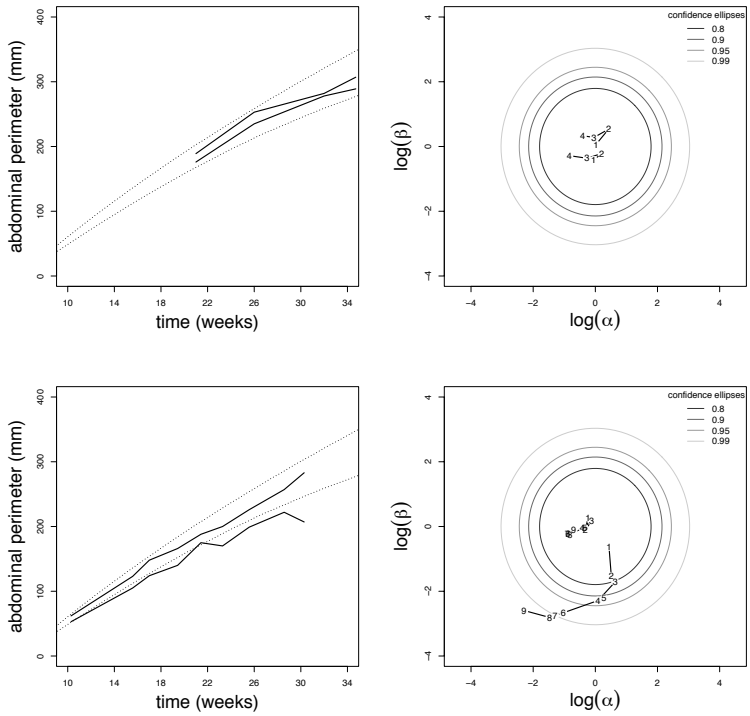

FIG. 1.3 – Données de croissance pour une grossesse gémellaire à croissance "normale" (ligne du haut) et une grossesse à croissance "anormale" (ligne du bas). Sur la colonne de gauche, observations des deux grossesses et 5ème et 95ème percentiles de population estimés à partir de la population de référence par modèle mixte. Sur la colonne de droite, estimations de $\mathbb{E}(\phi|Y_{1:j};\theta)$ pour $j = 1,\dots n = 4$ (grossesse normale) et $j = 1,\dots n = 9$ (grossesse anormale) et ellipses de confiance de population de niveau 80%, 90%, 95% calculées à partir du modèle mixte. L'indice de la visite est donné à chaque point.

gynécologique et obstétrique classique. Cela nous permet d'éviter des biais de sélection et d'impact dus au suivi intensif de l'instant précis de l'ovulation. A ma connaissance, c'est la première fois que cette loi de probabilité est estimée sur la population générale.

Pour formaliser le problème, on note X l'intervalle entre la date des dernières règles et le début de la grossesse, et Y la mesure (bruitée) de cet intervalle obtenue via la première mesure échographique. Le but est d'estimer la densité f_X de X. On dispose de deux échantillons de grossesses, issus de l'hôpital Necker. Le premier concerne $N = 1378$ grossesses uniques pour lesquelles on dispose d'une mesure Y_k bruitée de l'intervalle X_k

$$Y_k = X_k + \varepsilon_k, \quad k = 1,\dots,N.$$

Le second échantillon provient de $M = 86$ grossesses gémellaires, où l'on dispose de deux mesures bruitées (Y_{k1}, Y_{k2}) du même intervalle X_k (le début de grossesse est le même pour les deux jumeaux) :

$$Y_{k1} = X_k + \varepsilon_{k1}, \quad Y_{k2} = X_k + \varepsilon_{k2}, \quad k = 1,\dots,M$$

On suppose que (ε_k), (ε_{k1}) et (ε_{k2}) sont i.i.d. et ont pour densité f_ε. Ce problème rentre dans le cas des deux échantillons (1.15) et (1.16) de la section 1.3.2. Nous proposons donc d'utiliser la méthode développée dans **[S4]** pour estimer la densité f_X.

Nous étudions l'influence de deux covariables, l'âge de la mère et la régularité des cycles. L'âge de la mère est considéré comme une variable qualitative à quatre modalités. Pour chaque classe d'âge, la densité de début de grossesse est estimée à partir des grossesses à cycles réguliers. Ensuite, la densité de début de grossesse est estimée chez les patientes ayant des cycles réguliers, quel que soit leur âge, et chez les patientes ayant des cycles irréguliers, quel que soit leur âge. Les densités estimées sont

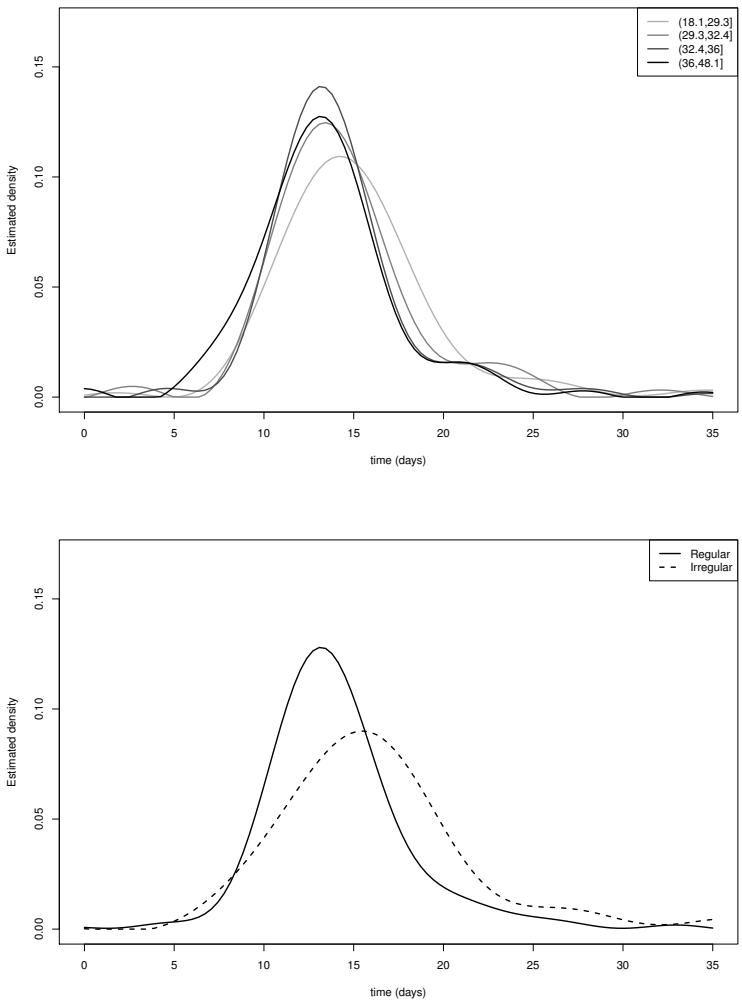

FIG. 1.4 – Figure du haut : densité de début de grossesse estimée en fonction de l'âge maternel. Figure du bas : densité de début de grossesse estimée en fonction de la régularité des cycles de la mère.

présentées dans la figure 1.4. Lorsque l'âge augmente, les intervalles entre les dates des dernières règles et le début de grossesse se réduit. Dans chaque classe d'âge, les modes des densités sont 14.1 jours, 13.1 jours, 13 jours et 13 jours pour des femmes âgées de [18.1-29.3],]29.3, 32.4],]32.4,36] et]36,48.1] ans, respectivement. La densité correspondant aux femmes plus âgées est plus dissymétrique, ce qui est cohérent avec le fait que la première phase d'un cycle menstruel diminue avec l'âge. La densité de début de grossesse des femmes ayant des cycles irréguliers est décalée dans le temps et a une variance plus grande que celle des femmes ayant des cycles réguliers, ce qui est en concordance avec les travaux de Wilcox *et al.* (2000).

Chapitre 2

Modèles stochastiques en biologie

Un grand nombre de phénomènes biologiques sont naturellement modélisés par un processus en temps continu. Par exemple, l'évolution au cours du temps de la concentration d'un médicament dans le sang, l'activité d'un neurone stimulé par un input électrique, la relation entre la glycémie et l'insuline au cours du temps, etc. Des modèles mathématiques sous forme de systèmes d'équations différentielles ordinaires ont été proposés dans ces différents domaines. Cependant, ces systèmes théoriques et rigides ne tiennent pas compte de phénomènes non prévisibles. Par exemple, il est connu que l'anxiété peut modifier la métabolisation de l'insuline chez un diabétique, modifiant la dynamique "théorique" de la relation glycémie/insuline. De même l'élimination d'un médicament n'est pas constante au cours du temps. Les causes de ces changements ne sont pas toujours connues, rendant leur prédiction impossible sous forme déterministe. Le caractère stochastique des neurones a également été largement étudié et confirmé de façon expérimentale. Plusieurs auteurs ont montré qu'une modélisation stochastique de ces phénomènes permettait de tenir compte de ces changements imprévisibles (Ditlevsen et De Gaetano, 2005; Höpfner et Brodda, 2006; Höpfner, 2007). Cette approche est au coeur de mes travaux de recherche. Dans ce chapitre, j'expose deux projets qui rentrent dans ce cadre.

Le premier est réalisé en collaboration avec Benjamin Favetto, Valentine Genon-Catalot, Yves Rozenholc (Laboratoire MAP5, Université Paris Descartes) et Charles-André Cuenod (Service de radiologie, Hôpital Européen Georges Pompidou). Nous étudions la vascularisation de cellules cancéreuses à partir de séquences d'images médicales enregistrées après l'injection d'un agent de contraste au patient. La modélisation de la pharmacocinétique de l'agent de contraste est habituellement réalisée par un modèle différentiel déterministe qui peut s'avérer instable. Nous proposons de transformer ce système en un système d'équations différentielles stochastiques (EDS), en ajoutant une perturbation brownienne à chaque équation déterministe. Dans **[A11]**, nous étudions ce système et proposons une méthode d'estimation des paramètres. Dans **[A13]**, nous utilisons cette approche sur les données médicales et montrons que la solution obtenue par l'EDS est plus stable que celle du modèle déterministe.

Le second problème concerne la modélisation de l'activité neuronale. Il existe plusieurs familles de modèles neuronaux. Les modèles les plus simples sont constitués de modèles stochastiques unidimensionnels. Les observations sont alors des mesures discrètes de ces modèles, incluant ou non un bruit de mesure. En collaboration avec Jérôme Dedecker (Laboratoire MAP5, Université Paris Descartes) et Marie-Luce Taupin (Laboratoire Statistiques et Génome, Université d'Evry), nous proposons une méthode d'estimation paramétrique pour un modèle auto-régressif observé avec bruit **[S3]**. Des modèles neuronaux plus complexes sont constitués d'EDS bidimensionnelles dont on n'observe que la première composante. L'EDS peut être elliptique ou hypoelliptique. Lorsque l'EDS est elliptique, en collaboration avec Susanne Ditlevsen (Copenhagen University, Denmark), nous proposons d'utiliser l'algorithme SAEM pour maximiser une pseudo-vraisemblance obtenue par schéma d'Euler. L'algorithme SAEM est combiné à un filtre particulaire qui permet de "reconstituer" la composante non observée **[S6]**. Lorsque l'EDS est hypoelliptique, nous proposons avec Michèle Thieullen (LPMA, Université Pierre et Marie Curie) une méthode d'estimation basée sur une fonction de contraste **[A17]**.

Le projet sur la modélisation stochastique de la pharmacocinétique d'un agent de contraste est dé-

taillé dans la section 2.1. Mes travaux réalisés sur la modélisation de l'activité neuronale sont présentés dans la section 2.2.

2.1 Imagerie médicale et pharmacocinétique

2.1.1 Description des données et modèle pharmacocinétique

La micro-vascularisation des tissus et le phénomène d'angiogénèse peuvent être étudiés *in vivo* par imagerie dynamique de contraste (DCE-imaging). Ces techniques sont de plus en plus utilisées en imagerie médicale pour suivre l'évolution des cancers. Après l'injection au patient d'un agent de contraste, une séquence d'images est réalisée qui permet de visualiser la cinétique de l'agent de contraste dans les tissus. Dans la littérature, la cinétique est modélisée par des modèles pharmacocinétiques compartimentaux, dont les paramètres ont une interprétation physiologique. Ces systèmes déterministes ne modélisent pas les fluctuations aléatoires dues à des causes environnementales internes ou externes (mouvement du patient, anxiété, variations au cours du temps des paramètres de vascularisation, erreur de mesure de l'entrée artérielle, etc). Dans **[A11, A13]**, nous proposons une version stochastique d'un modèle pharmacocinétique qui tient compte de ces variations aléatoires. Je présente le modèle pharmacocinétique déterministe à partir duquel nous avons travaillé, puis le modèle stochastique que nous proposons. Ensuite, je détaille l'estimation des paramètres de ce modèle et les résultats obtenus sur la séquence d'images médicales fournie par Charles-André Cuenod.

La figure 2.1 schématise le modèle pharmacocinétique à deux compartiments que nous avons considéré. On note $AIF(t)$ (Arterial Input Function) la concentration d'agent de contraste dans l'artère au

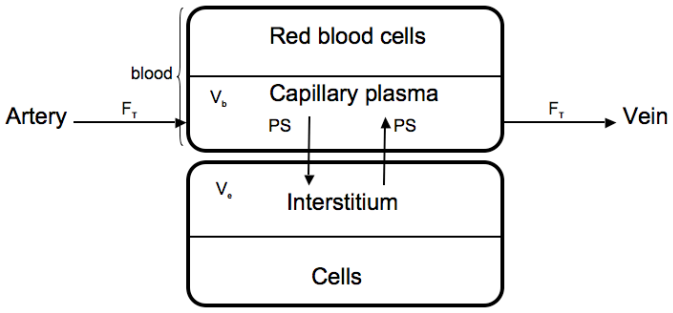

FIG. 2.1 – Modèle pharmacocinétique à deux compartiments.

temps t. L'agent de contraste est injecté dans la veine au temps t_0, il transite par l'artère et arrive dans le plasma avec un flux de perfusion par unité de volume égal à $F_T \geq 0$ (en $\text{ml.min}^{-1}.100\text{ml}^{-1}$), proportionnellement à la concentration d'agent dans l'artère $AIF/(1-h)$, où h est le taux d'hématocrite. Le délai avec lequel l'agent de contraste arrive de l'artère dans le plasma est noté δ. L'agent de contraste est ensuite éliminé du plasma avec un flux F_T, proportionnel à la concentration d'agent dans le plasma $Q_P/(V_b(1-h))$, où $V_b \geq 0$ est la part de volume sanguin (en %). La quantité d'agent de contraste s'échangeant entre le plasma et l'espace interstitiel est égal à PS fois la concentration d'agent de contraste dans le plasma $Q_P/(V_b(1-h))$, où PS est l'aire de la surface perméable par unité de volume (en $\text{ml.min}^{-1}.100\text{ml}^{-1}$). Inversement, la quantité d'agent de contraste s'échangeant entre l'interstitiel et le plasma est égal à PS fois la concentration d'agent dans l'interstitiel Q_I/V_e et $V_e \geq 0$ est la fraction de volume extravasculaire et extracellulaire dans le tissu (en %). On a donc $V_b + V_e \leq 100$. Le taux d'hématocrite h est fixé à $h = 0.4$. Le modèle pharmacocinétique déterministe

correspondant s'écrit

$$\begin{aligned}\frac{dQ_P(t)}{dt} &= \frac{F_T}{1-h}\,AIF(t-\delta)\mathbb{1}_{[\delta,\infty)}(t) - \frac{PS}{V_b(1-h)}Q_P(t) + \frac{PS}{V_e}Q_I(t) - \frac{F_T}{V_b(1-h)}Q_P(t),\\ \frac{dQ_I(t)}{dt} &= \frac{PS}{V_b(1-h)}Q_P(t) - \frac{PS}{V_e}Q_I(t). \end{aligned} \tag{2.1}$$

On suppose qu'il n'y a pas d'agent de contraste dans le corps avant l'expérience. La condition initiale est donc $Q_P(t_0) = Q_I(t_0) = 0$ et $AIF(t_0) = 0$ avec $t_0 = 0$.

2.1.2 Modèle pharmacocinétique stochastique

Le modèle stochastique proposé dans **[A11, A13]** est directement déduit du modèle déterministe en ajoutant un accroissement brownien sur chaque composante :

$$\begin{aligned}dQ_P(t) &= \left(\frac{F_T}{1-h}\,AIF(t-\delta)\mathbb{1}_{[\delta,\infty)}(t) - \frac{PS}{V_b(1-h)}Q_P(t) + \frac{PS}{V_e}Q_I(t) - \frac{F_T}{V_b(1-h)}Q_P(t)\right)dt + \sigma_1 dB_1(t),\\ dQ_I(t) &= \left(\frac{PS}{V_b(1-h)}Q_P(t) - \frac{PS}{V_e}Q_I(t)\right)dt + \sigma_2 dB_2(t), \end{aligned} \tag{2.2}$$

où $B_1(t)$ et $B_2(t)$ sont deux mouvements Browniens indépendants, et σ_1, σ_2 sont des paramètres de volatilité. Les paramètres inconnus de ce modèle d'EDS sont $\theta = (F_T, V_b, PS, V_e, \sigma_1, \sigma_2, \sigma)$ et le paramètre de délai δ qui est traité séparément.

Les données sont issues de séquences d'images médicales, obtenues aux temps $t_0, \ldots, t_n$. On considère la cinétique de l'agent de contraste dans un pixel de l'image, dont on suit la dynamique tout au long de la séquence des $n+1$ images. En chaque pixel, on note $Y_0, \ldots, Y_n$ les niveaux de gris mesurés au cours du temps, qui sont supposés être proportionnels à la quantité totale d'agent de contraste à l'intérieur de ce pixel, à une erreur de mesure près. Le modèle d'observations est le suivant

$$Y_j = Q_P(t_j) + Q_I(t_j) + \sigma\varepsilon_j,$$

où (ε_j) sont les erreurs de mesures, supposées i.i.d. gaussiennes, centrées et réduites et Q_P, Q_I sont solution du modèle pharmacocinétique. On dispose également de mesures de la concentration d'agent de contraste dans l'artère au cours du temps $AIF(t_j)$, pour $j = 0, \ldots, n$.

Le problème statistique est l'estimation du paramètre θ d'une EDS bidimensionnelle dont seulement la somme des deux composantes est observée à temps discrets avec un bruit de mesure. Nous proposons d'utiliser le filtre de Kalman pour calculer de façon exacte la vraisemblance. L'étude du modèle et la méthode d'estimation proposée sont détaillées dans la section 2.1.3. L'application sur les données réelles est détaillée dans la section 2.1.4.

2.1.3 Méthode d'estimation par maximum de vraisemblance

Afin de simplifier les notations, nous considérons la reparamétrisation suivante

$$\alpha = \frac{F_T}{1-h},\ \beta = \frac{F_T}{V_b(1-h)},\ \lambda = \frac{PS}{V_b(1-h)},\ k = \frac{PS}{V_b(1-h)} + \frac{PS}{V_e},$$

et la variable $S(t) = Q_P(t) + Q_I(t)$ de sorte que

$$Y_j = S(t_j) + \sigma\varepsilon_j.$$

On introduit le vecteur $U(t) = (S(t), Q_I(t))'$ qui permet d'écrire le modèle (2.2) sous forme matricielle

$$\begin{aligned} Y_j &= J\,U(t_j) + \sigma\varepsilon_j,\\ dU(t) &= (F(t) + G\,U(t))dt + \Sigma dB(t), \quad U(0) = U_0, \end{aligned}$$

où $J = (1\ 0)$ et

$$F(t) = \begin{pmatrix} \alpha AIF(t-\delta)\mathbb{1}_{[\delta,\infty)}(t) \\ 0 \end{pmatrix}, G = \begin{pmatrix} -\beta & \beta \\ \lambda & -k \end{pmatrix}, dB(t) = \begin{pmatrix} dB_1(t) \\ dB_2(t) \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1 & \sigma_2 \\ 0 & \sigma_2 \end{pmatrix}.$$

Le processus U est une diffusion d'Ornstein-Uhlenbeck bidimensionnelle, qui peut être explicitement résolue. La matrice G est diagonalisable avec deux valeurs propres distinctes $\mu_1 = \frac{-(\beta+k)-\sqrt{d}}{2}$ et $\mu_2 = \frac{-(\beta+k)+\sqrt{d}}{2}$ où $d = (\beta-k)^2 + 4\beta\lambda > 0$. On introduit les matrices des valeurs propres D et des vecteurs propres P

$$D = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}, \ P = \begin{pmatrix} 1 & 1 \\ \frac{\beta-k-\sqrt{d}}{2\beta} & \frac{\beta-k+\sqrt{d}}{2\beta} \end{pmatrix} \text{ avec } D = P^{-1}GP.$$

On peut alors montrer la proposition suivante

***Proposition* 4.** *Soit $X(t) = P^{-1}U(t)$ et $\Gamma = (\Gamma^{k\ell})_{1\le k,\ell\le 2} = P^{-1}\Sigma$. Alors pour $t, h \ge 0$, on a :*

$$\begin{aligned} X(t+h) &= e^{Dh}X(t) + b(t,t+h) + Z(t,t+h), \\ b(t,t+h) &= e^{D(t+h)} \int_t^{t+h} e^{-Ds} P^{-1} F(s) ds \mathbb{1}_{[\delta,\infty)}(t), \\ Z(t,t+h) &= e^{D(t+h)} \int_t^{t+h} e^{-Ds} \Gamma dB(s). \end{aligned}$$

La distribution conditionnelle de $X(t+h)$ sachant $X(s), s \le t$ est

$$\mathcal{N}_2\left(e^{Dh}X(t) + b\left(t,t+h\right), R\left(t,t+h\right)\right) \quad \textit{avec} \quad R(t,t+h) = \left(\frac{e^{(\mu_k+\mu_l)h}-1}{\mu_k+\mu_l}\left(\Gamma\Gamma'\right)^{kl}\right)_{1\le k,l\le 2}.$$

Si on suppose que $AIF(t) \equiv c \ge 0$ avec c une constante, $(X(t))$ a une distribution stationnaire gaussienne dont l'espérance est $M = -D^{-1}P^{-1}F$ et la matrice de variance est $V = \left(\frac{1}{-(\mu_k+\mu_l)}\left(\Gamma\Gamma'\right)^{kl}\right)_{1\le k,l\le 2}$.

Nous proposons une méthode d'estimation du paramètre θ par maximum de vraisemblance. On note $Y_{0:n} = (Y_0, \ldots, Y_n)$. Comme la loi de $(X(t), \varepsilon_j)$ est gaussienne, la vraisemblance $L(Y_{0:n}, \theta)$ peut être explicitement calculée. Cependant sa maximisation requiert l'inversion d'une matrice de taille $2(n+1) \times 2(n+1)$. Cette inversion peut être numériquement instable. Nous proposons de calculer la vraisemblance exacte via l'algorithme de filtrage de Kalman, qui ne nécessite pas d'inversion de matrice. L'EDS vérifiée par $X(t)$ est plus simple que celle vérifiée par $U(t)$, on travaille donc sur le modèle

$$\begin{aligned} Y_j &= JPX(t_j) + \sigma\varepsilon_j, \\ dX(t) &= (DX(t) + P^{-1}F(t))dt + \Gamma dB(t),\ X(0) = P^{-1}U_0 = X_0. \end{aligned}$$

Les vecteurs propres pouvant être choisis à une constante de proportionnalité, on a $H = JP = (1\ 1)$. Le fait que H ne comporte pas de paramètres simplifie grandement les calculs exacts du gradient et de la hessienne de la vraisemblance. On considère également le modèle discrétisé pour $X_j = X(t_j)$

$$\begin{aligned} Y_j &= HX_j + \sigma\varepsilon_j, \\ X_j &= A_jX_{j-1} + b_j + \eta_j,\ \ \eta_j \sim \mathcal{N}(0, R_i), \end{aligned}$$

où $A_j = \exp(D(t_j - t_{j-1}))$, $b_j = b(t_{j-1}, t_j)$, $R_j = R(t_{j-1}, t_j)$. Nous proposons un algorithme récursif de calcul exact de la vraisemblance, son gradient et sa hessienne, basé sur le filtrage de Kalman. Ces

calculs exacts permettent d'obtenir l'estimateur du maximum de vraisemblance (EMV) par maximisation numérique de la vraisemblance $L(Y_{0:n}, \theta)$ via un algorithme du gradient. L'algorithme EM exact (Dempster *et al.*, 1977) peut également être utilisé dans ce contexte.

Les propriétés de l'EMV sont étudiées dans le cas où $X(t)$ est en régime stationnaire, ce qui suppose que $AIF(t) \equiv c$. On suppose également que les temps d'observations sont équidistants, c'est-à-dire $\Delta = t_j - t_{j-1}$, pour tout $j = 1, \ldots, n$. On a alors $A_j = A$, $R_j = R$ et $b_j = b = (I - A)M$. On introduit une nouvelle paramétrisation du modèle $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ où $\theta_\ell = e^{\mu_\ell \Delta}$, $\ell = 1, 2$ et θ_3, θ_4 et θ_5 sont des fonctions explicites de $\mu_1, \mu_2, \sigma_1, \sigma_2$ et Δ telles que

$$A = A(\theta) = \begin{pmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{pmatrix} \quad \text{et} \quad R = R(\theta) = \begin{pmatrix} \theta_3 & \theta_5 \\ \theta_5 & \theta_4 \end{pmatrix}$$

On pose également $\theta_6 = m = HM$. Enfin, on pose $\vartheta = (\theta, \sigma^2)$.

Nous montrons que le processus $(\tilde{Y}_j) = (Y_j - \theta_6)$ est un processus ARMA(2,2). Sa densité spectrale a une forme explicite, qui permet de dégager les paramètres identifiables :

***Proposition* 5.** *Les quantités identifiables sont σ^2, $\theta_1 + \theta_2$ et $\theta_1\theta_2$, et au plus deux parmi les trois paramètres θ_3, θ_4 et θ_5. Si Δ est petit, exactement deux des trois paramètres θ_3, θ_4 et θ_5 sont identifiables.*

On ne peut donc pas identifier les 7 paramètres du modèle dans le cas où $X(t)$ est stationnaire. Ceci est dû au fait que les observations sont de dimension 1 alors que le processus X est de dimension 2. La perte de dimension se traduit dans ce problème d'identifiabilité.

On note ϑ_0 la vraie valeur des paramètres $(\theta_1, \ldots, \theta_6, \sigma^2)$. On suppose que $\vartheta_0 \in \Theta$, qui est un ouvert de $\mathbb{R}^7$. On note $\vartheta^- = (\theta_1, \ldots, \theta_5, \sigma^2) \in \Theta^-$ la projection de ϑ sur $\mathbb{R}^6$. Nous montrons la proposition suivante

***Proposition* 6.** *Soit $\widehat{\theta}$ l'estimateur du maximum de vraisemblance de ϑ_0^- basé sur les observations $\widetilde{Y}_{0:n}$. Si la densité spectrale est une fonction $\mathcal{C}^3$ de $[-\pi, \pi] \times \Theta^-$ et si elle est bijective, alors $\widehat{\theta} \to \vartheta_0^-$ presque sûrement quand $n \to \infty$. De plus, $\sqrt{n}(\widehat{\theta} - \vartheta_0^-)$ converge en loi :*

$$\sqrt{n}(\widehat{\theta} - \vartheta_0^-) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\vartheta_0^-))$$

où $I^{-1}(\vartheta_0^-)$ est la matrice d'information de Fisher.

On obtient donc un estimateur consistant et asymptotiquement gaussien si le processus $X(t)$ est stationnaire.

Dans **[A13]**, nous considérons le cas où $X(t)$ est non stationnaire, c'est-à-dire quand l'entrée artérielle $AIF(t)$ n'est pas constante. Il faut alors estimer le paramètre de délai δ. Pour simplifier l'étude théorique, nous supposons qu'il existe un entier $j^\star \in \mathbb{N}$ tel que $\delta = j^\star\Delta$. L'EMV de $(\theta, j^\star)$ est défini par

$$(\widehat{\theta}, \widehat{j^\star}) = \arg\max_{j,\theta} L(Y_{0:n}, \theta, j),$$

où $L(Y_{0:n}, \theta, j)$ est la vraisemblance du modèle sous l'hypothèse que le délai vaut $\delta = j\Delta$. Nous proposons une méthode d'estimation en deux étapes. La première maximise $L(Y_{0:n}, \theta, j)$ en θ, pour tout j. On obtient alors $\widehat{\theta}(j) = \arg\max_\theta L(Y_{0:n}, \theta, j)$. Ensuite $j^\star$ est estimé par $\widehat{j^\star} = \arg\max_j L(Y_{0:n}, \hat{\theta}(j), j)$ et finalement $\widehat{\theta} = \widehat{\theta}(\widehat{j^\star})$.

L'étude des propriétés de $\widehat{j^\star}$ et $\widehat{\theta}$ dans un cadre général est difficile. Nous nous restreignons au cas où la fonction artérielle est constante par morceaux. On suppose qu'il existe des constantes $(a_j)_{j=1,\ldots,n}$ telles que

$$AIF(t) = \sum_{j=1}^{n} a_j \mathbb{1}_{[(j-1)\Delta, j\Delta[}(t).$$

Nous considérons ensuite deux conditions qui correspondent à deux phases différentes de l'expérience :

Trajectoire	complète		sans les 5 derniers points	
	modèle EDO	modèle EDS	modèle EDO	modèle EDS
F_T	22.40 (1.41)	17.54 (3.53)	20.05 (0.00)	17.50 (3.23)
V_b	36.09 (2.30)	45.23 (7.68)	17.77 (0.00)	45.07 (3.99)
PS	2.03 (0.56)	0.57 (5.77)	120.78 (0.00)	0.77 (0.00)
V_e	63.91 (1.19)	0.03 (4.05)	15.75 (0.00)	0.04 (3.26)
δ	12.00 (−−)	9.60 (−−)	9.60 (−−)	9.60 (−−)
σ	7.69 (0.83)	6.51 (0.00)	8.10 (0.67)	6.37 (0.00)
σ_1	-	1.27 (0.66)	-	1.34 (0.47)
σ_2	-	0.03 (1.02)	-	0.03 (1.03)
BIC	937.82	925.32	897.99	888.04

TAB. 2.1 – Paramètres estimés et critère BIC pour les modèles EDO et EDS. Les valeurs entre parenthèses correspondent aux écart-types estimés par la matrice d'information de Fisher.

(AIF 1) L'AIF a une croissance linéaire ($a_j = j\Delta$) (réaliste dans la première phase de l'expérience).

(AIF 2) L'AIF est constante (réaliste après le pic d'injection de l'agent de contraste).

Sous la condition **(AIF 1)**, on se ramène à l'estimation d'un point de rupture dans la dérive d'un processus ARMA. Alors, en utilisant des arguments de Bai (1994), nous montrons la proposition suivante :

Proposition 7. *Supposons que $j^\star = \lfloor n\tau^\star \rfloor$ pour $\tau^\star \in (0,1)$. On définit l'estimateur $\widehat{\tau} = \widehat{j^\star}/n$. Alors, il existe une constante c qui peut être calculée explicitement en fonction de θ telle que*

$$\widehat{\tau} - \tau^\star = O_P\left(\frac{1}{c^2 n}\right).$$

Ensuite, sous la condition **(AIF 2)**, nous nous ramènons au cas stationnaire traité dans **[A11]**, qui permet d'obtenir des propriétés sur l'estimateur $\widehat{\theta}(j)$ pour j fixé.

2.1.4 Analyse des données réelles

Des séquences de $n = 130$ images médicales du bassin de quatre femmes sont analysées avec les modèles déterministe (EDO) et stochastique (EDS). Pour chaque jeu de données, et chaque modèle, les paramètres et les prédictions de $Q_P(t)$ et $Q_I(t)$ sont estimés. Pour certains jeux de données, les modèles EDO et EDS donnent exactement les mêmes estimations des paramètres. Pour d'autres jeux de données, les modèles diffèrent en terme d'estimation des paramètres. Par exemple, le volume V_b estimé par le modèle EDS ($\widehat{V}_b = 0.03$) est significativement plus petit que celui estimé par le modèle EDO ($\widehat{V}_b = 63.91$) (table 2.1). L'estimation de σ_1 est différente de 0 ($\widehat{\sigma}_1 = 1.27$). L'estimation par le modèle EDS de la quantité $Q_I(t)$ est toujours nulle alors que ce n'est pas le cas avec le modèle EDO. L'estimation du modèle EDO atteint les bornes du domaine d'optimisation ($\widehat{V}_b + \widehat{V}_e = 100$). L'estimation via le modèle d'EDS reste stable quand on enlève les 3 ou les 5 derniers points de la trajectoire (table 2.1), alors que l'estimation des paramètres V_b, PS et V_e via le modèle d'EDO change complètement. Le modèle EDS est donc plus stable et moins sensible à certaines observations, qui n'apparaissent pourtant pas atypiques sur la trajectoire observée.

L'utilisation de modèles stochastiques en pharmacocinétique est donc encourageante comme alternative à la complexification des modèles déterministes évoqués dans le chapitre 1.

2.2 Modélisation de l'activité neuronale

Le second projet concerne la modélisation de l'activité neuronale. On s'intéresse à la modélisation de l'activité du potentiel de la membrane de neurones isolés. Le caractère stochastique intrinsèque des

neurones a été largement montré expérimentalement. De nombreux modèles stochastiques ont donc été proposés pour décrire cette dynamique (Lansky et Ditlevsen, 2008). Je me suis intéressée à deux types de modèles et à l'estimation de leurs paramètres. Dans la section 2.2.1, je présente un travail sur un modèle auto-régressif mesuré avec erreur **[S3]**. Dans les sections 2.2.3 et 2.2.4, je détaille deux travaux portant sur des systèmes différentiels stochastiques partiellement observés **[S6, A17]**.

2.2.1 Modèle auto-régressif observé avec bruit

Plusieurs auteurs ont émis l'hypothèse que la mesure expérimentale du potentiel de la membrane est une mesure bruitée d'un processus $V(t)$ à temps discrets (Jahn *et al.*, 2011). Dans **[S3]**, nous choisissons de travailler avec un modèle auto-régressif observé avec bruit

$$\begin{aligned} Y_j &= V_j + \varepsilon_j, \\ V_j &= f_\theta(V_{j-1}) + \xi_j, \end{aligned} \tag{2.3}$$

où on observe les variables Y_j aux temps t_j et où le processus auto-régressif (V_j) est non observé. On suppose que la fonction de régression f_θ est connue à un paramètre θ près. Les erreurs $(\varepsilon_j)_{j\geq 0}$ sont supposées i.i.d. de variance $\mathrm{Var}(\varepsilon_0) = \sigma^2$, de densité f_ε connue par rapport à la mesure de Lebesgue. On suppose que les variables V_0, $(\xi_j)_{j\geq 1}$ et $(\varepsilon_i)_{i\geq 0}$ sont indépendantes. Les innovations $(\xi_j)_{j\geq 1}$ sont supposées i.i.d., centrées et de variance $\mathrm{Var}(\xi_1) = \sigma_\xi^2$, de distribution inconnue. On considère le cas général où cette distribution n'admet pas nécessairement de densité par rapport à la mesure de Lebesgue. Enfin, on suppose que $(V_j)_{j\geq 0}$ est strictement stationnaire. Finalement, on suppose que la vraie valeur θ_0 du paramètre appartient à l'intérieur d'un espace compact $\Theta \subset \mathbb{R}^p$ et on cherche à estimer ce paramètre à partir des observations Y_j.

Le modèle (2.3) est un modèle de Markov caché, avec espace d'état non compact et continu et une distribution des innovations inconnue. Quand la distribution des innovations est connue à un paramètre près, le modèle est complètement paramétrique et a été largement étudié. Les paramètres peuvent être estimés par maximum de vraisemblance et des résultats de consistance, normalité asymptotiques et efficacité ont été prouvés.

Dans **[S3]**, nous considérons le cas plus général, où la distribution des innovations est inconnue. Peu de résultats existent dans ce contexte, excepté Comte et Taupin (2001) qui proposent une procédure d'estimation basée sur la minimisation d'un critère des moindres carrés modifié. Elles proposent une borne supérieure de la vitesse de convergence de leur estimateur qui dépend de la régularité de la fonction de régression et de la densité f_ε. Leurs résultats sont obtenus en supposant que V_0 admet une densité par rapport à la mesure de Lebesgue et que la chaîne de Markov $(V_j)_{j\geq 0}$ est absolument régulière (β-mélangeante). Cependant, leur critère d'estimation n'est pas explicite et il est difficile de montrer que la vitesse de convergence est la vitesse paramétrique, excepté pour certaines fonctions de régression très particulières. Nous proposons une nouvelle méthode d'estimation qui fournit un estimateur consistant à vitesse paramétrique pour une large classe de fonctions.

Cette approche est basée sur la fonction de contraste

$$S_{\theta_0,P_V}(\theta) = \mathbb{E}[(Y_1 - f_\theta(V_0))^2\, w(V_0)],$$

où w est une fonction de poids à choisir et $\mathbb{E}$ est l'espérance $\mathbb{E}_{\theta_0,P_V}$ où P_V est la loi de V_0.

Pour une fonction $(\theta, u) \mapsto \phi_\theta(u)$ de $\Theta \times \mathbb{R}$ dans $\mathbb{R}$, les dérivées premières et secondes par rapport à θ sont notées $\phi_\theta^{(1)}(\cdot)$ et $\phi_\theta^{(2)}(\cdot)$ respectivement. On note également $\mathbb{P}$, $\mathbb{E}$ et Var les probabilités sous $\mathbb{P}_{\theta_0,P_V}$, l'espérance $\mathbb{E}_{\theta_0,P_V}$ et la variance $\mathrm{Var}_{\theta_0,P_V}$, quand la vraie valeur du paramètre est θ_0. On se place sous le jeu d'hypothèses suivantes.

- **(C1)** Sur $\mathring{\Theta}$, la fonction $\theta \mapsto f_\theta$ admet des dérivées continues jusqu'à l'ordre 3 par rapport à θ.
- **(C2)** Sur $\mathring{\Theta}$, la quantité $w(V_0)(Y_1 - f_\theta(V_0))^2$, et la valeur absolue de ses dérivées par rapport à θ jusqu'à l'ordre 2 ont des espérances finies.

(C3) La quantité $\mathbb{E}[(f_{\theta_0}(V) - f_\theta(V))^2 w(V)]$ admet un unique minimum en $\theta = \theta_0$.

(C4) Pour tout $\theta \in \mathring{\Theta}$, la matrice $S^{(2)}_{\theta_0,P_V}(\theta)$ existe et la matrice $S^{(2)}_{\theta_0,P_V}(\theta_0)$ est définie positive.

(C5) La densité f_ε appartient à l'espace $\mathbb{L}_2(\mathbb{R})$ et pour tout $u \in \mathbb{R}$, $f^*_\varepsilon(u) \neq 0$.

(C6) Les fonctions (wf_θ) et (wf^2_θ) appartiennent à $\mathbb{L}_1(\mathbb{R})$ et les fonctions w^*/f^*_ε , $(f_\theta w)^*/f^*_\varepsilon$, $(f^2_\theta w)^*/f^*_\varepsilon$ appartiennent à $\mathbb{L}_1(\mathbb{R})$.

(C7) Les fonctions $\sup_{\theta\in\Theta}\left|\left(f^{(1)}_{\theta,\ell}w\right)^*/f^*_\varepsilon\right|$ et $\sup_{\theta\in\Theta}\left|\left(f_\theta f^{(1)}_{\theta,\ell}w\right)^*/f^*_\varepsilon\right|$ appartiennent à $\mathbb{L}_1(\mathbb{R})$ pour tout $\ell \in \{1,\dots,d\}$.

La première étape pour construire notre estimateur est d'estimer la fonction de contraste $S_{\theta_0,P_V}(\theta)$ puisqu'on ne dispose pas des observations V_j. On utilise le principe suivant, due à la théorie de Fourier. Pour toute fonction g telle que g et g^*/f^*_ε appartiennent à $\mathbb{L}_1(\mathbb{R})$, par indépendance de ε_0 et V_0 on a

$$\mathbb{E}[g(V_0)] = \mathbb{E}\left(\frac{1}{2\pi}\int g^*(u)e^{-iuV_0}du\right) = \mathbb{E}\left(\frac{1}{2\pi}\int \frac{g^*(u)e^{-iuY_0}}{f^*_\varepsilon(-u)}du\right).$$

On peut donc estimer $\mathbb{E}[g(V_0)]$ à partir des observations $Y_0,\dots,Y_n$ par

$$\frac{1}{2\pi}\mathbb{R}e\int \frac{g^*(u)\frac{1}{n}\sum_{j=1}^n e^{-iuY_j}}{f^*_\varepsilon(-u)}du.$$

Finalement, nous proposons d'estimer $S_{\theta_0,P_V}(\theta)$ par

$$S_n(\theta) = \frac{1}{2\pi n}\sum_{j=1}^n \mathbb{R}e\int \frac{\left(\left(Y_j - f_\theta\right)^2 w\right)^*(u)\, e^{-iuY_{j-1}}}{f^*_\varepsilon(-u)}du,$$

Nous définissons ensuite l'estimateur de θ_0 par

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} S_n(\theta).$$

On peut alors montrer le résultat de consistance suivant :

Théorème 3. *Sous les hypothèses **(C1)-(C7)**, $\widehat{\theta}$ converge en probabilité vers θ_0.*

Pour montrer la normalité asymptotique de $\widehat{\theta}$ et sa vitesse de convergence paramétrique, nous considérons une hypothèse **(C8)** de régularité sur les dérivées secondes et troisièmes de $f_\theta w$ et $f^2_\theta w$ similaire à l'hypothèse **(C7)**. Nous étudions les propriétés asymptotiques de $\widehat{\theta}$ sous des conditions d'α-mélange de la chaîne de Markov V_j et sous des conditions de τ-dépendance, telles que définies par Dedecker et Prieur (2005). On note $\alpha_V(k)$ et $\tau_{V,2}(k)$ les coefficients d'α-mélange et de τ-dépendance de la chaîne. On peut montrer le résultat asymptotique suivant.

Théorème 4. *Soient les hypothèses **(C1)-(C8)**.*

1. *On note $Q_{|V_1|}$ l'inverse cadlag de la fonction $t \to \mathbb{P}(|V_1| > t)$. On suppose que $\sum_{k\geq 1}\int_0^{\alpha_{\mathbf{V}}(k)} Q^2_{|V_1|}(u)du < \infty$, alors $\widehat{\theta}$ est un estimateur $\sqrt{n}$-consistant de θ_0 et il existe une matrice de covariance Σ_1 telle que*
$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\Sigma_1).$$
2. *On note $G(t) = t^{-1}\mathbb{E}(V_1^2\mathbf{1}_{V_1^2>t})$, et G^{-1} l'inverse cadlag de G. On suppose que $\sum_{k>0} G^{-1}(\tau_{\mathbf{V},2}(k))\tau_{\mathbf{V},2}(k) < \infty$, alors $\widehat{\theta}$ est un estimateur $\sqrt{n}$-consistant de θ_0 et*
$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\Sigma_1).$$

L'estimateur proposé atteint donc la vitesse paramétrique pour une grande classe de fonctions, dès lors qu'on est capable d'exhiber une fonction de poids w vérifiant les conditions **(C1)-(C8)**. Le choix de w dépend de la fonction de régression f_θ et de la densité f_ε du bruit. Des exemples de fonctions de poids w sont données pour des fonctions de régression spécifiques.

Pour une fonction de régression linéaire $f_\theta(v) = av+b$, nous considérons deux chaînes de Markov V_j τ-dependantes mais pas α-mélangeantes, dont l'une a une densité stationnaire absolument continue par rapport à la mesure de Lebesgue et l'autre non. Nous proposons deux fonctions de poids w différentes :

$$w(v) = N(v) = \exp\{-v^2/(4\sigma^2)\} \quad \text{et} \quad w(v) = SC(v) = \frac{1}{2\pi}\Big(\frac{2\sin(v)}{v}\Big)^4.$$

Le poids N dépend de de la variance σ^2. Cela permet à l'estimateur de "s'adapter" au niveau de bruit. Inversement, l'estimateur sera plus sensible aux bruits de faible variance. L'estimateur $\widehat{\theta}$ est explicite avec la fonction de poids N. Il doit être calculé par transformée de Fourier pour la fonction de poids SC. Ces deux estimateurs sont comparés à l'estimateur classique d'un modèle ARMA et à l'estimateur naïf basé sur un critère des moindres carrés où on remplace les variables V_j non observées par les variables Y_j. L'étude de simulation montre que l'estimateur naïf a de mauvaises propriétés, ce qui était attendu. L'estimateur d'un processus ARMA a souvent un biais plus grand que les deux estimateurs proposés. L'estimateur basé sur la fonction de poids SC s'avère moins sensible aux rapports signal sur bruit faibles.

Nous considérons aussi une fonction de régression de Cauchy $f_\theta(v) = \theta/(1+v^2)$. A ma connaissance, il n'existe pas d'estimateur consistant pour ce modèle autre que celui que nous proposons. Nous considérons à nouveau deux fonctions de poids

$$N_c(v) = (1+v^2)^2 \exp\{-v^2/(4\sigma^2)\} \text{ et } SC_c(v) = (1+v^2)^2 \frac{1}{2\pi}\Big(\frac{2\sin(v)}{v}\Big)^4.$$

L'estimateur $\widehat{\theta}$ est explicite avec la fonction de poids N_c et doit être calculé par transformée de Fourier pour la fonction de poids SC_c. L'estimateur basé sur la fonction de poids SC_c s'avère à nouveau moins sensible aux rapports signal sur bruit faibles dans les simulations.

Nous proposons donc un estimateur consistant, qui atteint la vitesse paramétrique pour des modèles généraux, en particulier pour une classe plus large que celle considérée dans Comte et Taupin (2001). L'analyse des données réelles neuronales à partir de ces modèles reste à effectuer.

2.2.2 Systèmes d'équations différentielles stochastiques

Des systèmes différentiels ont été proposés pour décrire la dynamique du potentiel de la membrane neuronale. A partir des premiers modèles déterministes de Hodgkin-Huxley, des versions stochastiques ont été introduites, pour tenir compte du caractère stochastique des neurones. L'estimation paramétrique de ces modèles peut s'avérer très complexe.

Nous nous concentrons sur le modèle bidimensionnel de Morris-Lecar stochastique, proposé par Ditlevsen et Greenwood (2012) :

$$\begin{aligned} dV(t) &= f_\theta(V(t), Z(t))dt + \gamma dB_V(t) \\ dZ(t) &= b_\theta(V(t), Z(t))dt + a_\theta(V(t), Z(t))dB_Z(t), \end{aligned} \tag{2.4}$$

où $B_V(t)$ et $B_Z(t)$ sont des mouvements browniens indépendants, la première équation représente le potentiel $V(t)$ de la membrane du neurone, la deuxième équation décrit l'ouverture de canaux ou portes, situés à la surface de la membrane, qui permettent l'échange d'ions entre l'intérieur et l'extérieur de la cellule. Selon la complexité des modèles, on distingue plusieurs types d'ions (potassium, sodium, etc). Dans le modèle de Morris-Lecar, le processus $Z(t)$, qui prend ses valeurs entre 0 et 1, peut être interprété comme la proportion de portes ouvertes au temps t pour les ions potassium K^+, aussi appelée conductance normalisée du courant des ions K^+. La fonction f_θ décrit la dynamique de $V(t)$

et comporte quatre termes, correspondants aux courants Ca^{2+}, K^+, un courant de fuite et un courant d'entrée I. La fonction b_θ modélise les taux d'ouverture et de fermeture des canaux ioniques K^+ et la fonction a_θ leurs fluctuations. Les fonctions f, a et b sont des fonctions paramétriques non-linéaires connues, à un paramètre θ près. Le paramètre γ représente le bruit du courant d'entrée. Dans certains modèles, ce paramètre est considéré égal à 0.

Les observations disponibles sont des mesures à temps discrets de la variable $V(t)$, la variable $Z(t)$ n'est pas mesurable. On note V_j l'observation obtenue au temps t_j, $j = 0, \ldots, n$ et $V_{0:n} = (V_0, \ldots, V_n)$. Les mesures sont généralement disponibles à haute fréquence, toutes les 0.1 ms, pendant une période d'observation de plusieurs secondes.

On cherche à estimer le paramètre θ à partir des observations $V_{0:n}$. Lorsque $\gamma \neq 0$, dans **[S6]**, nous proposons une méthode d'estimation de θ en utilisant des outils de type filtre particulaire (section 2.2.3). Lorsque $\gamma = 0$, le système est hypoelliptique et partiellement observé. Dans **[A17]**, nous proposonss une méthode d'estimation de θ en utilisant une fonction de contraste (section 2.2.4).

2.2.3 Equation différentielle stochastique bidimensionnelle partiellement observée

En collaboration avec Susanne Ditlevsen, nous considérons le modèle (2.4) dans le cas où le paramètre γ n'est pas nul. Comme le processus $(Z(t))$ n'est pas observé, la vraisemblance est définie par

$$L(V_{0:n};\theta) = \int \ldots \int p(V_0, Z_0;\theta) \prod_{j=1}^{n} p(V_j, Z_j | V_{j-1}, Z_{j-1};\theta) dZ_0 \ldots dZ_n,$$

où $p(V_j, Z_j | V_{j-1}, Z_{j-1};\theta)$ est la densité de transition du système (2.4). Cette vraisemblance n'est pas explicite et il est difficile d'estimer les paramètres du modèle de Morris-Lecar par maximum de vraisemblance. A ma connaissance, les seuls travaux existants sont basés sur une linéarisation de l'EDS (Vogelstein *et al.*, 2009).

Dans **[S6]**, nous proposons de maximiser une pseudo-vraisemblance associée au modèle discret obtenu par un schéma d'Euler de pas $\Delta = t_j - t_{j-1}$:

$$\begin{aligned} V_{j+1} &= V_j + \Delta\, f_\theta(V_j, Z_j) + \sqrt{\Delta}\,\gamma\,\tilde{\eta}_i, \\ Z_{j+1} &= Z_j + \Delta\, b_\theta(V_j, Z_j) + \sqrt{\Delta} a_\theta(V_j, Z_j)\eta_i, \end{aligned} \tag{2.5}$$

où $(\tilde{\eta}_i)$ et (η_i) sont des variables gaussiennes standard i.i.d. On note par un indice Δ les lois se rapportant à ce modèle et $L_\Delta(V_{0:n};\theta)$ la pseudo-vraisemblance associée. La maximisation de la pseudo-vraisemblance $L_\Delta(V_{0:n};\theta)$ reste complexe car le processus (Z_j) n'est pas observé.

Nous proposons d'utiliser l'algorithme SAEM qui considère les observations $V_{0:n}$ comme faisant partie d'un vecteur de données complètes $(V_{0:n}, Z_{0:n})$. L'étape de simulation n'est pas explicite car la loi du filtrage $p_\Delta(Z_{0:n}|V_{0:n};\theta)$ n'est pas connue. Nous suggérons de combiner l'algorithme SAEM avec une méthode de filtrage particulaire du type Sequential Monte Carlo (SMC) (Doucet *et al.*, 2001). Le but de l'algorithme SMC est la construction d'un ensemble de L particules $(Z_{0:n}^{(\ell)})_{\ell=1\ldots L}$ et de poids associés $(W_{0:n}^{(\ell)})_{\ell=1\ldots L}$ approchant la distribution $p_\Delta(Z_{0:n}|V_{0:n};\theta)$ par une loi empirique.

Peu de filtres particulaires ont été proposés pour des EDS ou des systèmes auto-régressifs partiellement observés. On peut citer le travail de Del Moral et Jacod (2001) qui traite d'une EDS bidimensionnelle partiellement observée mais dont la seconde équation est autonome. Le filtre qu'ils proposent est aussi basé sur une approximation de l'EDS par un schéma d'Euler mais n'est pas applicable au modèle (2.5). En effet, le fait que la seconde équation de leur système soit autonome rend le processus caché Markovien. Ce n'est pas le cas du modèle (2.5) dont la seconde équation dépend de la première : le processus caché $Z(t)$ n'est pas Markovien, seul le couple $(V(t), Z(t))$ l'est.

Nous proposons un algorithme SMC adapté au cas d'une EDS bidimensionnelle partiellement observée et non autonome. L'algorithme SMC proposé, utilisant la distribution instrumentale $q(Z_j|V_j, V_{j-1}, Z_{j-1};\theta)$ est le suivant :

***Algorithme* 3 (Algorithme SMC).**

– Au temps $j = 0 : \ \forall \ell = 1, \ldots, L$

1. *simulation de* $Z_0^{(\ell)}$,
2. *calcul et normalisation des poids :*

$$w_0\left(Z_0^{(\ell)}\right) = p_\Delta\left(V_0, Z_0^{(\ell)}; \theta\right), \quad W_0\left(Z_0^{(\ell)}\right) = \frac{w_0\left(Z_0^{(\ell)}\right)}{\sum_{\ell=1}^{L} w_0\left(Z_0^{(\ell)}\right)},$$

– Au temps $j = 1, \ldots, n : \ \forall \ell = 1, \ldots, L$

1. *rééchantillonnage des particules via le tirage d'indices* $A_{j-1}^{(\ell)}$ *dans un loi multinomiale telle que*

$$P(A_{j-1}^{(\ell)} = l) = W_j(Z_{0:j-1}^{(\ell)}), \quad \forall\, l = 1, \ldots, L,$$

et définition de $Z_{0:j-1}^{'(\ell)} = Z_{0:j-1}^{(A_{j-1}^{(\ell)})}$,

2. *tirage de* $Z_j^{(\ell)} \sim q\left(\cdot|V_j, V_{j-1}, Z_{i-1}^{'(k)}; \theta\right)$ *et définition de* $Z_{0:j}^{(\ell)} = (Z_{0:j-1}^{'(k)}, Z_j^{(\ell)})$,
3. *calcul et normalisation des poids :*

$$\begin{aligned} w_j\left(Z_{0:j}^{(\ell)}\right) &= \frac{p_\Delta\left(V_{0:j}, Z_{0:j}^{(\ell)}; \theta\right)}{p_\Delta\left(V_{0:j-1}, Z_{0:j-1}^{'(\ell)}; \theta\right) q\left(Z_j^{(\ell)}|V_j, V_{j-1}, Z_{0:j-1}^{'(\ell)}; \theta\right)}, \\ W_j(Z_{0:j}^{(\ell)}) &= \frac{w_j\left(Z_{0:j}^{(\ell)}\right)}{\sum_{\ell=1}^{L} w_j\left(Z_{0:j}^{(\ell)}\right)}. \end{aligned}$$

Finalement, l'algorithme SMC fournit une mesure empirique

$$\Psi_{n\theta}^{L} = \sum_{\ell=1}^{L} W_n(Z_{0:n}^{(\ell)}) \mathbf{1}_{Z_{0:n}^{(\ell)}}$$

qui est une approximation de la distribution de filtrage $p_\Delta(Z_{0:n}|V_{0:n}; \theta)$. Le choix optimal de la distribution instrumentale $q\left(\cdot|V_j, V_{j-1}, Z_{i-1}^{'(\ell)}; \theta\right)$ est la loi conditionnelle $p_\Delta\left(Z_j|V_j, V_{j-1}, Z_{i-1}^{'(k)}; \theta\right)$ qui est explicite pour le modèle (2.5). Contrairement à l'algorithme SMC proposé par Del Moral et Jacod (2001) dans un contexte proche, nous ne resimulons pas les variables observées $V_{0:n}$. Ceci permet d'éviter la dégénérescence des poids des particules.

Pour montrer la convergence de ce nouvel algorithme, nous nous plaçons sous l'hypothèse :

(SMC1) les fonctions $p_\Delta(V_j|V_{j-1}, Z_j, Z_{j-1}; \theta)$ sont bornées uniformément en θ.

On montre alors le lemme suivant pour l'algorithme 3 :

Lemme 2. *Sous l'hypothèse* ***(SMC1)****, pour toute fonction* f *bornée sur* $\mathbb{R}$*, il existe des constantes* C_1 *et* C_2*, indépendantes de* θ*, telles que*

$$\mathbb{P}\left(\left|\Psi_{n,\theta}^{L} f - \pi_{n,\theta} f\right| \geq \varepsilon\right) \leq C_1 \exp\left(-L \frac{\varepsilon^2}{C_2 \|f\|^2}\right)$$

où $\|f\|$ *est la norme supérieure de* f *et* $\pi_{n,\theta} f = \mathbb{E}_\Delta(f(Z_{0:n})|V_{0:n}; \theta)$.

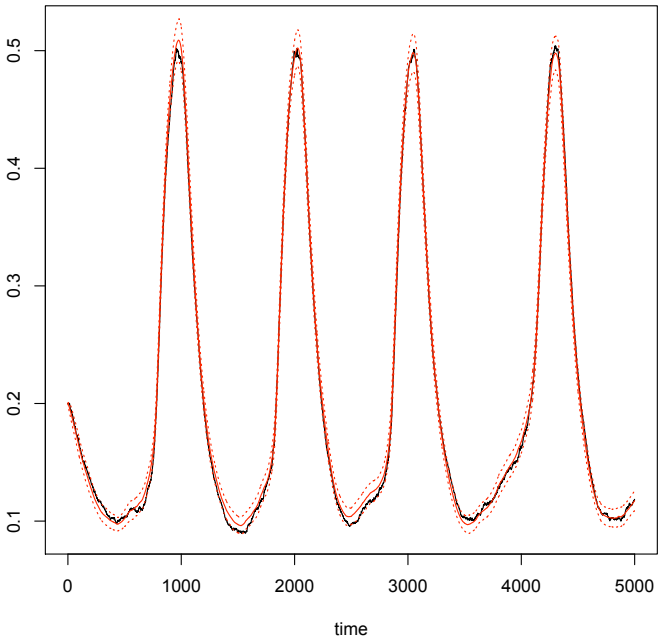

FIG. 2.2 – Filtrage du processus $Z(t)$ obtenu par l'algorithme SMC à θ connu à partir de données $V_{0:n}$ simulées : trajectoire simulée $Z(t)$ cachée (ligne noire), trajectoire moyenne filtrée par SMC (ligne pleine rouge) et intervalle de confiance à 95% (pointillés rouges).

La preuve s'inspire de celle proposée par Del Moral et Jacod (2001) et l'étend au cas d'un processus caché (Z_j) non Markovien. Elle est basée sur des inégalités exponentielles de contrôle de sommes de variables indépendantes et centrées.

Nous étudions ensuite la convergence de l'algorithme SAEM-SMC dont l'étape de simulation est réalisée par l'algorithme SMC. Outre les hypothèses classiques de convergence de l'algorithme SAEM, et en particulier l'hypothèse **(M)** d'appartenance du modèle (2.5) à la famille exponentielle, nous nous plaçons sous les hypothèses :

(SMC2) Les statistiques exhaustives S définies dans l'hypothèse **(M)** sont bornées uniformément en z.

(SMC3) Le nombre de particules L utilisées à chaque itération de l'algorithme SAEM varie avec les itérations de sorte qu'il existe une fonction $g(m) \to \infty$ quand $m \to \infty$ telle que $L(m) \geq g(m)\log(m)$.

On obtient alors la convergence de l'algorithme SAEM-SMC.

Théorème 5. *Sous les hypothèses de convergence de l'algorithme SAEM et les hypothèses **(SMC1)-(SMC3)**, avec probabilité 1, la suite d'estimateurs $(\widehat{\theta}_m)$ construite par l'algorithme SAEM-SMC converge vers un maximum de la pseudo-vraisemblance $L_\Delta(V_{0:n};\theta)$.*

La preuve repose sur une décomposition de l'étape d'approximation stochastique qui comprend un terme supplémentaire par rapport à la décomposition dans l'algorithme SAEM classique. Ce terme supplémentaire provient de l'erreur de simulation réalisée en utilisant l'algorithme SMC au lieu d'une simulation exacte. Il est contrôlé par le lemme 2. Les hypothèses **(SMC1)** et **(SMC3)** ne sont pas restrictives, contrairement à l'hypothèse **(SMC2)**.

Une étude de simulation évalue les algorithmes SMC et SAEM-SMC. A partir d'une trajectoire simulée $V_{0:n}$, nous filtrons le processus caché $Z(t)$ par l'algorithme SMC proposé à θ connu. Un exemple de filtrage obtenu est représenté sur la figure 2.2 et montre la précision de l'algorithme SMC. L'estimation

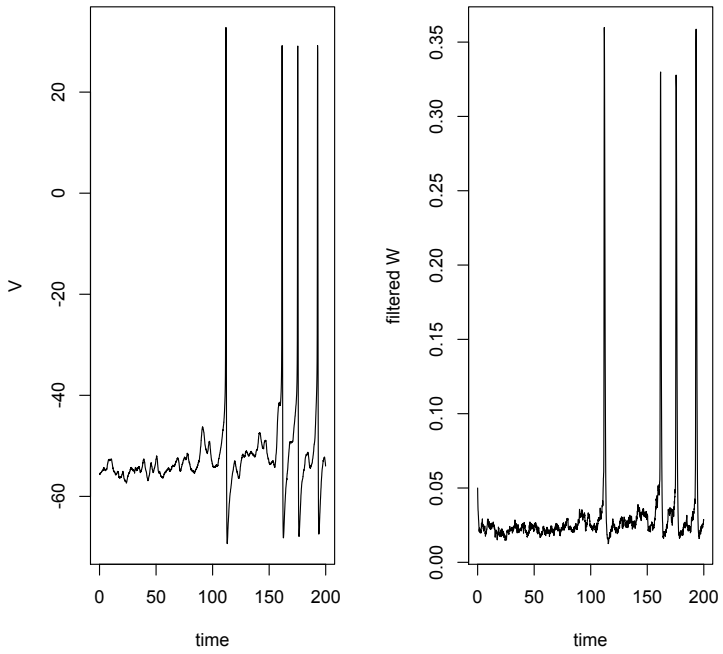

FIG. 2.3 – Données réelles d'un neurone moteur : données observées $V_{0:n}$ à gauche et filtrage du processus $Z(t)$ obtenu par l'algorithme SAEM-SMC à droite.

des paramètres par l'algorithme SAEM-SMC montre des biais faibles, y compris quand le nombre de particules est faible ($L = 50$ en pratique suffit).

Enfin, nous appliquons l'algorithme sur des données d'enregistrement intracellulaire d'un neurone moteur chez une tortue. La trajectoire filtrée de $Z(t)$ obtenue sur ces données est représentée sur la figure 2.3. Les pics de la trajectoire filtrée, que l'on peut interpréter comme l'ouverture des canaux ioniques de la membrane du neurone, coïncident avec les décharges (spikes) observées sur le potentiel de la membrane $V_{0:n}$. Ces premiers résultats sont donc tout à fait encourageants.

2.2.4 Equation différentielle stochastique hypoelliptique

Dans **[A17]**, nous considérons le modèle (2.4) dans le cas où le paramètre γ est nul. On a alors

$$\begin{aligned} dV(t) &= f_\theta(V(t), Z(t))dt, \\ dZ(t) &= b_\theta(V(t), Z(t))dt + a_\theta(V(t), Z(t))dB(t). \end{aligned} \tag{2.6}$$

Ce système est hypoelliptique sous l'hypothèse

(HE1) $\forall (v, z) \in \mathbb{R} \times \mathbb{R}, \quad \partial_z f(v, z) \neq 0.$

L'hypothèse **(HE1)** est vérifiée pour le modèle de Morris-Lecar stochastique.

L'estimation des paramètres de ce modèle a été peu étudiée. Sa densité de transition n'est en général pas explicite. L'estimateur du minimum de contrate proposé par Kessler (1997) ne peut pas être directement utilisé car la matrice de volatilité du système n'est pas inversible. Pokern *et al.* (2009) proposent une approximation de la vraisemblance basée sur un développement d'Itô-Taylor, qui permet de se ramener à une matrice de volatilité inversible. Ils proposent un estimateur bayésien de θ basé sur un algorithme de Gibbs, mais l'étude théorique de l'erreur introduite par le développement d'Itô-Taylor n'est pas réalisée. De plus, ils se limitent aux modèles où $f_\theta(V(t), Z(t)) = Z(t)$, où la fonction de dérive

b_θ est linéaire et la fonction de volatilité a_θ est constante. Lorsque la fonction $f_\theta(V(t), Z(t)) = Z(t)$ et que la seconde équation est autonome, Gloter (2006) propose un contraste basé sur le schéma d'Euler de l'EDS autonome $dZ(t) = b_\theta(Z(t))dt + a_\theta(Z(t))dB(t)$. Il montre la consistance de cet estimateur et sa normalité asymptotique.

Dans **[A17]**, nous proposons un estimateur du minimum de contraste dont nous montrons la consistance et la normalité asymptotique pour une grande classe de fonctions f, a et b. Je présente dans la suite la classe de modèles considérés puis l'estimateur et ses propriétés asymptotiques.

Posons $X(t) := f(V(t), Z(t))$. La condition **(HE1)** implique que $Z(t)$ peut être définie de façon unique comme une fonction de $V(t)$ et $X(t)$ grâce au théorème des fonctions implicites. On a alors

$$\begin{aligned} dV(t) &= X(t)dt \\ dX(t) &= \tilde{b}(V(t), X(t))dt + \tilde{a}(V(t), X(t))dB(t), \end{aligned}$$

où les fonctions $\tilde{b}$ et $\tilde{a}$ sont définies via les fonctions b, a, f, la formule d'Itô et l'application du théorème des fonctions implicites. Afin de disposer d'un modèle paramétrique, nous considérons l'hypothèse supplémentaire

(HE2) Les fonctions $\tilde{b}$ et $\tilde{a}$ sont explicites.

Nous donnons une série d'exemples où la condition **(HE2)** est vérifiée, même si $Z(t)$ n'est pas une transformation explicite de $V(t)$ et $X(t)$. Dans la suite, pour alléger les notations, on notera b et a les fonctions de l'EDS (non-autonome) vérifiée par $X(t)$ et μ et σ les paramètres de ces fonctions, respectivement. On note $\theta = (\mu, \sigma)$ et on suppose que θ appartient à $\Theta = \Theta_1 \times \Theta_2$ pour $\Theta_1 \subset \mathbb{R}^{p_1}$ et $\Theta_2 \subset \mathbb{R}^{p_2}$ deux compacts.

On suppose donc que $(V(t), X(t))$ est l'unique solution du système

$$\begin{aligned} dV(t) &= X(t)dt \\ dX(t) &= b_{\mu_0}(V(t), X(t))dt + a_{\sigma_0}(V(t), X(t))dB(t), \end{aligned} \tag{2.7}$$

où $\theta_0 = (\mu_0, \sigma_0)$ est la vraie valeur des paramètres des fonctions de dérive et de volatilité. Dans la suite on note $b(v, x) = b_{\mu_0}(v, x)$ et $a(v, x) = a_{\sigma_0}(v, x)$.

Afin d'assurer l'existence de la solution de ce système et de montrer les propriétés asymptotiques de notre estimateur, nous introduisons les hypothèses suivantes :

(A1) (a) Il existe une constante c telle que $\sup_{\sigma\in\Theta_2} |a_\sigma^{-1}(v, x)| \leq c(1 + |v| + |x|)$.
(b) Pour tout $\theta \in \Theta$, b_μ et a_σ sont $\mathcal{C}^2(\mathbb{R}^2)$ et il existe une constante c telle que la fonction, ses dérivées première et secondes par rapport à v et x sont bornées par $c(1 + |v| + |x|)$, pour tout $x, v \in \mathbb{R}$, uniformément en θ.

(A2) (a) $\forall k \in]0, \infty[\quad \sup_{t\geq 0} \mathbb{E}(|X(t)|^k + |V(t)|^k) < \infty$.
(b) Il existe une constante c telle que $\forall t \geq 0, \quad \forall \delta \geq 0$,

$$\mathbb{E}(\sup_{s\in[t,t+\delta[} |X(s)|^k | \mathcal{G}_t) + \mathbb{E}(\sup_{s\in[t,t+\delta[} |Y(s)|^k | \mathcal{G}_t) \leq c(1 + |X_t|^k + |V_t|^k)$$

où $\mathcal{G}_t = \sigma(B(s), s \leq t)$.

(A3) $(V(t), X(t))$ admet une mesure de probabilité invariante unique ν_0 dont tous les moments sont finis : $\forall k > 0$, $\nu_0(|\cdot|^k) < \infty$

(A4) $(V(t), X(t))$ satisfait une version faible du théorème ergodique :

$$\frac{1}{T}\int_0^T f(V(s), X(s))ds \xrightarrow[T\to\infty]{} \nu_0(f) \quad p.s.$$

pour toute fonction continue f à croissance polynomiale à l'infini.

Nous détaillons une série de conditions basées sur l'existence d'une fonction de Lyapounov qui implique les hypothèses **(A1)-(A4)**. En particulier, nous montrons que ces hypothèses sont vérifiées pour les systèmes de Langevin, exemple illustré par des simulations.

Nous considérons le cas où on dispose des observations à temps discrets de $V(t)$ et $X(t)$ et le cas où on ne dispose que d'observations à temps discrets de $V(t)$. Le premier cas n'est pas détaillé dans ce manuscrit. Dans le deuxième cas, on suppose que les temps observations sont réguliers et on note $\Delta_n = t_j - t_{j-1}$ de sorte que $t_j = j\Delta_n$.

Lorsque $X(t)$ n'est pas observé, on peut approcher $X_j = X(t_j)$ par des incréments de $V(t)$. On introduit donc le processus des incréments

$$\overline{V}_{j,n} = \frac{V_{j+1} - V_j}{\Delta_n}. \tag{2.8}$$

Le modèle (2.7) implique que

$$\overline{V}_{j,n} = \frac{1}{\Delta_n}\int_{j\Delta_n}^{(j+1)\Delta_n} X(s)ds.$$

Quand Δ_n est petit, $\overline{V}_{j,n}$ est proche de X_j. Plus précisément, on a

***Proposition* 8.** *Sous les hypothèses (A1)-(A2),*

$$\overline{V}_{j,n} - X_j = \Delta_n^{1/2} a(V_j, X_j)\xi'_{j,n} + e_{i,n}$$

où il existe une constante c telle que $|\mathbb{E}(e_{i,n}|\mathcal{G}_j^n) \le c\Delta_n(1+|X_j|+|V_j|)$ et $|\mathbb{E}(e_{i,n}^2|\mathcal{G}_j^n) \le c\Delta_n^2(1+|X_j|^4+|V_j|^4)$.

Le lien entre deux termes successifs du processus non-markovien $\overline{V}_{j,n}$ est étudié dans la proposition suivante.

***Proposition* 9.** *Sous les hypothèses **(A1)-(A2)**, on a*

$$\overline{V}_{j+1,n} - \overline{V}_{j,n} - \Delta_n b(V_j, \overline{V}_{j,n}) \;=\; \Delta_n^{1/2} a(V_j, X_j)U_{j,n} + \varepsilon_{j,n}$$

où $U_{j,n} = \xi_{j,n} + \xi'_{j+1,n}$ avec $\xi_{j,n} = \frac{1}{\Delta_n^{3/2}}\int_j^{j+1}(s-j)dB(s)$ pour $i, n \ge 0$ et $\xi'_{j+1,n} = \frac{1}{\Delta_n^{3/2}}\int_{j+1}^{j+2}(j+1-s)dB(s)$ pour $i \ge -1, n \ge 0$. On a $\mathbb{E}(U_{j,n}|\mathcal{G}_j^n) = 0$ et $\mathbb{E}(U_{j,n}^2|\mathcal{G}_j^n) = 2/3\Delta_n$. De plus, il existe une constante c telle que $\mathbb{E}(\varepsilon_{j,n}|\mathcal{G}_j^n) \le c\Delta_n^2(1+|X_{j+1}|^3+|V_{j+1}|^3)$ et $\mathbb{E}((\varepsilon_{j,n})^2|\mathcal{G}_j^n) \le c\Delta_n^2(1+|X_{j+1}|^4+|V_{j+1}|^4)$.

Dans cette décomposition, la variance de $U_{j,n}$ est $2/3\Delta_n$, alors que la variance du terme similaire dans le cas d'observations complètes est Δ_n. Nous considérons donc un contraste corrigé d'un facteur $3/2$:

$$\mathcal{L}_n(\theta) = \sum_{j=1}^{n-2}\left(\frac{3}{2}\frac{\left(\overline{V}_{j+1,n} - \overline{V}_{j,n} - \Delta_n b_\mu(V_{j-1}, \overline{V}_{j-1,n})\right)^2}{\Delta_n a_\sigma^2(V_{j-1}, \overline{V}_{j-1,n})} + \log(a_\sigma^2(V_{j-1}, \overline{V}_{j-1,n}))\right).$$

Comme $(\overline{V}_{j,n})$ n'est pas markovien, nous introduisons un décalage dans les indices des fonctions de dérive et de volatilité afin d'éviter un terme de corrélation d'ordre $\Delta_n^{1/2}$ entre $(\overline{V}_{j+1,n} - \overline{V}_{j,n})$ et les fonctionnelles $f(V_j, \overline{V}_{j,n})$.

On définit ensuite l'estimateur du minimum de contraste $\widehat{\theta}$ par

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

Nous montrons la consistance et la normalité asymptotique de $\widehat{\theta}$. Pour simplifier les notations, on se limite au cas de paramètres μ et σ unidimensionnels. On introduit l'hypothèse d'identifiabilité

$$\begin{aligned} a_\sigma(v,x) = a_{\sigma_0}(v,x) \quad d\nu_0(v,x) \quad \text{presque partout} \ \Rightarrow \sigma = \sigma_0, \\ b_\mu(v,x) = b_{\mu_0}(v,x) \quad d\nu_0(v,x) \quad \text{presque partout} \ \Rightarrow \mu = \mu_0. \end{aligned}$$

On montre alors le théorème suivant :

Théorème 6. *On suppose que $\Delta_n \to 0$ quand $n \to \infty$. Sous les hypothèses **(A1)-(A4)**, l'estimateur $\widehat{\theta}$ est consistant :*

$$\widehat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0.$$

Pour montrer la normalité asymptotique, on considère de plus que $n\Delta_n^2 \to 0$. Les vitesses de convergence ne sont pas les mêmes pour $\widehat{\mu}$ et $\widehat{\sigma^2}$. On a le résultat suivant.

Théorème 7. *On suppose que $n\Delta_n^2 \to 0$ et que les hypothèses **(A1)-(A4)** sont vérifiées. Alors $\left(\sqrt{n\Delta_n}\left(\widehat{\mu} - \mu_0\right), \sqrt{n}\left(\widehat{\sigma^2} - \sigma_0^2\right)\right)$ converge en loi vers*

$$\mathcal{N}\left(0, \left\{\nu_0\left(\frac{(\partial_\mu b)^2(\cdot,\cdot)}{a^2(\cdot,\cdot)}\right)\right\}^{-1}\right) \otimes \mathcal{N}\left(0, \frac{9}{4}\left\{\nu_0\left(\frac{(\partial_{\sigma^2} a^2)^2(\cdot,\cdot)}{a^4(\cdot,\cdot)}\right)\right\}^{-1}\right)$$

Les preuves de ces deux théorèmes reposent sur l'étude des fonctionnelles

$$\begin{aligned} \overline{\nu}_n(f) &= \frac{1}{n}\sum_{i=0}^{n-1} f(V_j, \overline{V}_{j,n}, \theta) \\ \overline{I}_n(f) &= \frac{1}{n\Delta_n}\sum_{i=1}^{n-2} f(V_{j-1}, \overline{V}_{j-1,n}, \theta)(\overline{V}_{j+1,n} - \overline{V}_i - \Delta_n b(V_{j-1}, \overline{V}_{j-1,n})) \\ \overline{Q}_n(f) &= \frac{1}{n\Delta_n}\sum_{i=1}^{n-2} f(V_{j-1}, \overline{V}_{j-1,n}, \theta)\left(\overline{V}_{j+1,n} - \overline{V}_i\right)^2. \end{aligned}$$

Leur convergence presque sûre s'obtient grâce à l'existence d'une version faible du théorème ergodique (hypothèse A4). Leur convergence en loi s'obtient en utilisant un théorème centrale limite pour martingales.

Une étude de simulation illustre les propriétés des estimateurs dans le cadre des systèmes de Langevin

$$\begin{aligned} dV(t) &= X(t)dt, \\ dX(t) &= [-\gamma X(t) - F_D'(V_t)]dt + \sigma dB(t), \end{aligned}$$

où $\sigma > 0$, $F \in C^\infty(\mathbf{R}, [0, +\infty[)$ est une fonction non-linéaire qui dépend d'un paramètre D et F' est sa dérivée par rapport à v.

Trois modèles sont considérés dans les simulations. Le modèle I correspond à une croissance linéaire stochastique avec $\gamma = 0, F_D \equiv 0$. Le modèle II correspond à un oscillateur linéaire sujet à du bruit et à un amortissement avec $\gamma > 0, F_D(v) \equiv \frac{D}{2}v^2$. Le modèle III est un oscillateur non-linéaire avec bruit et amortissement où $\gamma > 0, F_D(v) \equiv -\Sigma_{j=1}^{n} j^{-1} D_j (\cos v)^j$.

Pour les trois modèles, les fonctions de dérive et de volatilité sont sous-linéaires et vérifient les hypothèses **(A1)-(A4)**. Sur les simulations, l'estimateur montre de bonnes propriétés, en particulier quand Δ_n diminue et n augmente. La comparaison avec les résultats de Pokern *et al.* (2009) sont nettement en faveur de notre estimateur.

En conclusion, à ma connaissance, il s'agit du premier estimateur avec de bonnes propriétés asymptotiques pour cette classe d'EDS hypoelliptique. L'analyse sur le modèle de Morris-Lecar stochastique reste à réaliser mais les résultats numériques sur les systèmes de Langevin sont tout à fait prometteurs. Mes projets de recherche sur ce point sont détaillés dans la dernière section de ce manuscrit.

Chapitre 3

Modèles mixtes et équations différentielles stochastiques

Dans ce chapitre, nous considérons des équations différentielles stochastiques à paramètres aléatoires. Ces modèles mixtes définis par EDS permettent de modéliser la variabilité intra-individuelle des données de chaque sujet, en plus de la variabilité inter-sujets (variance des paramètres aléatoires) et d'une éventuelle erreur de mesure. Les applications principales de ces modèles mixtes définis par EDS sont les domaines de la pharmacocinétique et de la pharmacodynamie. Ils ont été introduits comme alternative aux modèles mixtes définis via des EDO par Ditlevsen et De Gaetano (2005) et Overgaard *et al.* (2005). Ces modèles mixtes définis par EDS sont également de plus en plus développés pour l'analyse de données neuronales (Picchini et Ditlevsen, 2011; Faugeras *et al.*, 2009).

Les méthodes d'estimation pour ces modèles sont complexes. Les problèmes d'estimation paramétrique d'EDS évoqués dans le chapitre 2 sont complexifiés par l'introduction des paramètres aléatoires. Il est donc nécessaire de développer de nouvelles méthodes d'estimation. Lorsqu'on dispose d'observations non bruitées des processus $X_k(t)$, avec Maud Delattre (Laboratoire de Mathématiques, Université Paris Sud) et Valentine Genon-Catalot, nous proposons une méthode d'estimation par maximum de vraisemblance exacte dans le cas où la fonction de dérive est linéaire en les paramètres aléatoires et étudions les propriétés de l'EMV **[S2]**. Cette méthode est présentée dans la section 3.2. Lorsqu'on dispose d'observations bruitées des EDS, en collaboration avec Sophie Donnet et Jean-Louis Foulley (INRA), nous proposons différents algorithmes (MCMC, SAEM-MCMC et par approche de filtrage particulaire) **[A5, A10, S5]**, qui sont exposés dans la section 3.3.

3.1 Introduction

On considère N processus stochastiques réels $(X_k(t), t \geq 0)$, $k = 1, \ldots, N$, dont la dynamique est régie par les EDS suivantes :

$$d\,X_k(t) = b(X_k(t), \phi_k)dt + a_\gamma(X_k(t))\,dB_k(t), \quad X_k(0) = x^k,\ k = 1, \ldots, N, \tag{3.1}$$

où $(B_1, \ldots, B_N)$ sont N processus de Wiener indépendants, $\phi_1, \ldots, \phi_N$ sont N variables aléatoires de $\mathbb{R}^p$ i.i.d., $(\phi_1, \ldots, \phi_N)$ et $(B_1, \ldots, B_N)$ sont indépendants et $x^k, k = 1, \ldots, N$ sont des valeurs réelles. Le coefficient de diffusion $a_\gamma : \mathbb{R} \to \mathbb{R}$ est une fonction connue, qui peut éventuellement dépendre d'un paramètre inconnu noté γ. La fonction de dérive $b(x, \phi)$ est connue au paramètre ϕ près et définie sur $\mathbb{R} \times \mathbb{R}^p$. Chaque processus $(X_k(t))$ correspond à un individu, et chaque variable ϕ_k représente le paramètre aléatoire de l'individu k. On suppose que les variables aléatoires $\phi_1, \ldots, \phi_N$ ont une densité commune $g(\phi, \beta)$ sur $\mathbb{R}^p$, où β est un paramètre inconnu. On appelle β_0 la vraie valeur du paramètre.

L'existence de solution d'une EDS à paramètre fixe a été largement étudiée sous différentes hypothèses (croissance linéaire des fonctions de dérive et de volatilité, hypothèse Lipschitz sur ces fonctions, ...). L'existence des processus (3.1) est plus complexe du fait de l'introduction des paramètres

aléatoires ϕ_k. Un premier travail consiste à vérifier leur existence. Dans **[S2]**, nous introduisons une hypothèse de croissance linéaire. Plus précisément, nous considérons la filtration $(\mathcal{F}_t, t \geq 0)$ définie par $\mathcal{F}_t = \sigma(\phi_k, B_k(s), s \leq t, k = 1, \ldots, N)$ et l'hypothèse suivante

(H1) (i) La fonction $(x, \phi) \to b(x, \phi)$ est C^1 sur $\mathbb{R} \times \mathbb{R}^p$, et :

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

(ii) La fonction $a_\gamma(.)$ est C^1 sur $\mathbb{R}$ et

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

Sous **(H1)**, l'équation (3.1) admet un unique processus solution $(X_k(t), t \geq 0)$, adapté à la filtration $(\mathcal{F}_t, t \geq 0)$ et les processus $(X_k(t), t \geq 0), k = 1, \ldots, N$ sont indépendants. De plus pour tout $\ell \geq 1$, $\mathbb{E}|\phi_k|^{2\ell} < \infty$, et pour tout $T > 0$, $\sup_{t \in [0,T]} \mathbb{E}[X_k(t)]^{2\ell} < \infty$.

L'estimation des paramètres des EDS à paramètres aléatoires diffère selon le type d'observations dont on dispose. Le cas d'observations en temps continu des processus X_k n'a jamais été abordé dans la littérature. On note $\mathbf{X}_k$ le vecteur des observations disponibles pour l'individu k et $\mathbf{X}$ le vecteur des observations pour tous les individus, la vraisemblance s'écrit

$$L(\mathbf{X}; \beta) = \prod_{k=1}^{N} L(\mathbf{X}_k; \beta) = \prod_{k=1}^{N} \int p(\mathbf{X}_k | \phi_k) g(\phi_k; \beta) d\phi_k, \tag{3.2}$$

où $p(\mathbf{X}_k|\phi_k)$ est la vraisemblance du k-ième processus conditionnellement à la valeur ϕ_k du paramètre. Cette vraisemblance peut s'écrire à partir de la formule de Girsanov. Dans **[S2]**, nous traitons ce problème pour une classe de processus linéaires dont la vraisemblance est explicite. Nous proposons un EMV et étudions ses propriétés asymptotiques. Ce problème est abordé dans la section 3.2.

Plusieurs auteurs ont considéré le cas d'observations à temps discrets t_{kj} des processus $(X_k(t))$. Les problèmes d'estimation d'EDS observées à temps discrets évoqués dans le chapitre 2 sont complexifiés par l'introduction des paramètres aléatoires. On garde les mêmes notations que pour le cas d'observations continues, c'est-à-dire que $\mathbf{X}_k$ représente alors le vecteur des observations aux temps discrets (t_{kj}) pour l'individu k et $\mathbf{X}$ le vecteur des observations pour tous les individus. Par la propriété de Markov du processus X_k, la vraisemblance s'écrit

$$L(\mathbf{X}; \beta) = \prod_{k=1}^{N} \int \prod_{j=1}^{J_k} p(X_k(t_{kj}) | X_k(t_{kj-1}); \phi_k) g(\phi_k; \beta) d\phi_k. \tag{3.3}$$

Même lorsque la densité de transition $p(X_k(t)|X_k(0) = x; \phi_k)$ de l'EDS est connue, cette vraisemblance est souvent non explicite car l'intégrale n'a pas de forme analytique. Ditlevsen et De Gaetano (2005) étudient le cas particulier d'un processus Brownien avec dérive ($dX_k(t) = \phi_k dt + \gamma dB_k(t)$) et un paramètre aléatoire ϕ_k gaussien. Ils montrent que la fonction de vraisemblance (3.3) est explicite et étudient l'EMV du paramètre β. Pour des EDS plus générales, Picchini et Ditlevsen (2011) proposent une approximation de la vraisemblance basée sur un développement d'Hermite de la densité de transition de l'EDS et une quadrature de Gauss du calcul de la vraisemblance. Ils montrent la convergence de leur algorithme vers l'estimateur du maximum de vraisemblance d'un modèle approché. Dans **[S2]**, nous abordons le problème autrement, en proposant non pas de maximiser la vraisemblance des observations discrètes (3.3) mais en discrétisant les estimateurs du maximum de la vraisemblance continue (3.2).

Dans la section 3.3, nous considérons le cas où les processus $(X_k(t))$ sont observés à temps discrets avec un bruit de mesure. Le modèle s'écrit alors

$$\begin{aligned} Y_{kj} &= X_k(t_{kj}) + \varepsilon_{kj}, \quad k = 1, \ldots, N, j = 1 \ldots, J_k \\ d\,X_k(t) &= b(X_k(t), \phi_k)dt + a_\gamma(X_k(t))\, dB_k(t), \quad X_k(0) = x^k, \\ \varepsilon_{kj} &\sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ \phi_k &\sim_{i.i.d.} g(\phi, \beta). \end{aligned} \tag{3.4}$$

Le vecteur des paramètres à estimer est $\theta = (\beta, \gamma, \sigma)$. La vraisemblance est plus complexe que précédemment :

$$L(\mathbf{Y}; \theta) = \prod_{k=1}^{N} \int p(\mathbf{Y}_k | \mathbf{X}_k; \theta) L(\mathbf{X}_k; \theta) d\mathbf{X}_k, \tag{3.5}$$

où $L(\mathbf{X}_k; \theta)$ est la vraisemblance (3.3). Elle n'est jamais explicite, y compris pour des EDS très simples comme un Brownien avec dérive. Des méthodes basées sur la linéarisation de l'EDS ont été proposées mais sans résultats théoriques. Nous proposons des méthodes d'estimation basées sur un algorithme MCMC **[A10]** et sur l'algorithme SAEM **[A5, S5]**.

3.2 Equation différentielle stochastique à paramètre aléatoire

En collaboration avec Maud Delattre et Valentine Genon-Catalot, nous traitons le modèle (3.1)

$$d\,X_k(t) = b(X_k(t), \phi_k)dt + a_\gamma(X_k(t))\,dB_k(t), \quad X_k(0) = x^k,\ k = 1, \ldots, N,$$

Nous considérons des observations en temps continu des processus $(X_k(t))$ sur l'intervalle de temps $[0, T_k]$ ou des observations à temps discrets (t_{kj}). On suppose γ connu et on note $a(x) = a_\gamma(x)$ dans la suite de cette section. On cherche à estimer par maximum de vraisemblance les paramètres $\beta \in \Theta$ de la densité des paramètres individuels ϕ_k et à étudier les propriétés de l'estimateur du maximum de vraisemblance quand N tend vers l'infini.

La vraisemblance des observations continues peut être calculée à partir de la formule de Girsanov et vaut

$$L(\mathbf{X}, \beta) = \prod_{k=1}^{N} \int_{\mathbb{R}^p} \exp\left(\int_0^{T_k} \frac{b(X_k(s), \phi_k)}{a^2(X_k(s))} dX_k(s) - \frac{1}{2} \int_0^{T_k} \frac{b^2(X_k(s), \phi_k)}{a^2(X_k(s))} ds \right) g(\phi_k, \beta) d\phi_k.$$

L'estimateur du maximum de vraisemblance est alors défini comme

$$\widehat{\beta} = \arg\max_{\beta \in \Theta} L(\mathbf{X}, \beta).$$

Cet estimateur n'est pas explicite dans le cas général.

Dans **[S2]**, nous restreignons notre étude au cas où la fonction de dérive est linéaire en les paramètres individuels $(b(x, \phi) = \phi b(x))$ et où l'intervalle de temps d'observations est le même chez tous les individus $T_k = T, x^k = x$. Les processus $(X_k(t), t \in [0, T]), k = 1, \ldots, N$ sont alors i.i.d..

Pour étudier l'EMV $\widehat{\beta}$, on suppose que $\int_0^T b^2(X_k(s))/a^2(X_k(s))ds < \infty$ et on introduit les variables

$$U_k = \int_0^T \frac{b(X_k(s))}{a^2(X_k(s))} dX_k(s), \quad V_k = \int_0^T \frac{b^2(X_k(s))}{a^2(X_k(s))} ds,$$

qui sont les statistiques exhaustives du modèle. L'intégrale V_k est ordinaire mais l'intégrale U_k est une intégrale stochastique. Alors la vraisemblance de l'individu k est égale à

$$L(\mathbf{X}_k, \beta) = \int_{\mathbb{R}} g(\phi_k, \beta) \exp\left(\phi_k U_k - \frac{\phi_k^2}{2} V_k\right) d\phi_k.$$

Si on suppose de plus que le paramètre individuel ϕ_k est gaussien, la vraisemblance est explicite. Pour simplifier les notations, on considère que ϕ_k est un scalaire, de loi $\mathcal{N}(\mu, \omega^2)$ et $\beta = (\mu, \omega^2)$. Le cas multidimensionnel est détaillé dans **[S2]**. On peut montrer que

$$L(\mathbf{X}_k, \beta) = \frac{1}{(1 + \omega^2 V_k)^{1/2}} \exp\left[-\frac{V_k}{2(1 + \omega^2 V_k)} \left(\mu - \frac{U_k}{V_k} \right)^2 \right] \exp\left(\frac{U_k^2}{2V_k} \right). \tag{3.6}$$

Les dérivées de la log-vraisemblance sont

$$\begin{aligned}\frac{\partial}{\partial\mu}\log L(\mathbf{X},\beta) &= \sum_{k=1}^{N}\left(\frac{U_k}{1+\omega^2V_k}-\mu\frac{V_k}{1+\omega^2V_k}\right),\\ \frac{\partial}{\partial\omega^2}\log L(\mathbf{X},\beta) &= \frac{1}{2}\sum_{k=1}^{N}\left[\left(\frac{U_k}{1+\omega^2V_k}-\mu\frac{V_k}{1+\omega^2V_k}\right)^2-\frac{V_k}{1+\omega^2V_k}\right].\end{aligned}$$

Quand ω_0^2 est connu, on obtient un estimateur explicite pour μ_0 :

$$\widehat{\mu}=\frac{\sum_{k=1}^{N}\frac{U_k}{1+\omega_0^2V_k}}{\sum_{k=1}^{N}\frac{V_k}{1+\omega_0^2V_k}}\,.$$

Cet estimateur est à comparer à l'EMV d'une EDS observée en temps continu sans paramètre aléatoire $\tilde{\mu}=\sum_{k=1}^{N}U_k/\sum_{k=1}^{N}V_k$. Même quand ω_0^2 est connu, l'étude de l'EMV $\widehat{\mu}$ est donc plus complexe.

Quand les deux paramètres sont inconnus, les EMV de $\beta_0=(\mu_0,\omega_0^2)$ sont donnés par le système :

$$\begin{aligned}\widehat{\mu} &= \left(\sum_{k=1}^{N}\frac{V_k}{1+\widehat{\omega}^2V_k}\right)^{-1}\left(\sum_{k=1}^{N}\frac{U_k}{1+\widehat{\omega}^2V_k}\right),\\ \sum_{k=1}^{N}\left(\widehat{\mu}-\frac{U_k}{V_k}\right)^2\frac{V_k^2}{(1+\widehat{\omega}^2V_k)^2} &= \sum_{k=1}^{N}\frac{V_k}{1+\widehat{\omega}^2V_k}.\end{aligned}$$

On a déjà évoqué dans l'introduction du Chapitre 1 les propriétés de consistance et de normalité asymptotique des EMV pour les modèles mixtes. La principale contribution est due à Nie et Yang (2005). Ce sont des résultats de consistance faible qui sont basés sur une série d'hypothèses techniques difficiles à vérifier pour les modèles mixtes définis par EDS. En particulier, une des hypothèses suppose qu'on peut échanger la dérivation et l'intégration de la fonction de vraisemblance ($\partial_\beta\int L(X_k,\beta)dX_k=\int\partial_\beta L(X_k,\beta)dX_k$). Or cette propriété est très difficile à montrer pour une EDS à paramètre aléatoire. Nous ne pouvons donc pas utiliser les approches classiques d'étude de l'EMV.

Notre preuve consiste à montrer que le score, et plus précisément les variables $\frac{U_k}{1+\omega^2V_k}$, ont une transformée de Laplace finie. On peut alors en déduire que l'espérance de la fonction score est nulle et montrer la convergence en loi de la hessienne de la log-vraisemblance. La consistance et la normalité de l'EMV s'en découlent.

Plus précisément, on se place sous les hypothèses suivantes :

(H2) La function $b(.)/a(.)$ n'est pas constante. Sous la distribution Q du processus canonique, les variables aléatoires (U_1,V_1) admettent une densité $f(u,v)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}\times(0,+\infty)$ qui est continue et positive sur un ouvert de $\mathbb{R}\times(0,+\infty)$.

(H3) L'espace des paramètres Θ est un compact de $\mathbb{R}\times(0,+\infty)$.

(H4) La vraie valeur β_0 appartient à $\mathring{\Theta}$.

(H5) La matrice d'information de Fisher $\mathcal{I}(\beta_0)$ est inversible.

Sous des hypothèses de régularité des fonctions b,a, l'hypothèse **(H2)** est vérifiée par application d'outils de calcul de Malliavin. Le cas $b(.)/a(.)$ constant est simple et traité dans **[S2]**. Les hypothèses **(H3)-(H5)** sont classiques. Ces hypothèses permettent d'obtenir un résultat d'identifiabilité, nécessaire pour l'étude asymptotique de l'EMV.

***Proposition* 10.** *On note $Q^1_{\beta_0}$ la distribution du processus $(X_1(t), t \in [0,T])$ sur l'espace des fonctions continues réelles définies sur $[0,T]$. Soit $K(Q^1_{\beta_0}, Q^1_\beta)$ l'information de Kullback de $Q^1_{\beta_0}$ par rapport à Q^1_β.*

*(i) Sous **(H1)-(H2)**, $Q^1_\beta = Q^1_{\beta_0}$ implique que $\beta = \beta_0$. Donc, $\beta \to K(Q^1_{\beta_0}, Q^1_\beta)$ admet un unique minimum en $\beta = \beta_0$.*

*(ii) Sous **(H1)**, la fonction $\beta \to K(Q^1_{\beta_0}, Q^1_\beta)$ est continue sur $\mathbb{R} \times (0, +, \infty)$.*

On peut alors prouver la consistance et la normalité asymptotique de $\widehat{\beta}$.

***Proposition* 11.** *1. Supposons **(H1)-(H3)** vraies. Sous Q_{β_0}, $\widehat{\beta}$ converge en probabilité vers β_0.*

*2. Supposons **(H1)-(H5)** vraies. L'EMV vérifie, quand N tend vers l'infini,*

$$\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_2(0, \mathcal{I}^{-1}(\beta_0)),$$

où $\mathcal{I}(\beta_0)$ est la matrice d'information de Fisher.

Nous considérons ensuite le cas d'observations discrètes des processus $X_k(t)$. On suppose qu'on observe simultanément les N processus $X_k(t)$ aux temps $t_j = j\frac{T}{J}$, pour $j = 0, \ldots, J$. Il est possible de travailler sur la vraisemblance (3.3) du processus discrétisé ou de discrétiser la vraisemblance (3.6). Dans les deux cas, il est difficile d'étudier les propriétés des estimateurs obtenus sauf pour des modèles très simples (Ditlevsen et De Gaetano, 2005).

Comme alternative, nous proposons de construire un estimateur en discrétisant l'EMV $\widehat{\beta}$ de la vraisemblance des observations en temps continu. Ceci consiste à remplacer les variables U_k, V_k, $k = 1, \ldots, N$ par leurs versions discrétisées :

$$\begin{aligned} U_k^J &= \sum_{j=0}^{J-1} \frac{b(X_k(t_j))}{a^2(X_k(t_j))}(X_k(t_{j+1}) - X_k(t_j)), \\ V_k^J &= \sum_{j=0}^{J-1} \frac{b^2(X_k(t_j))}{a^2(X_k(t_j))}(t_{j+1} - t_j), \end{aligned}$$

dans la définition de $\widehat{\beta}$. On note $\widehat{\beta}^{(J)}$ l'estimateur déduit de $\widehat{\beta}$ en substituant les variables U_k^J, V_k^J aux variables U_k et V_k. Il suffit d'étudier les différences $U_k - U_k^J$ et $V_k - V_k^J$ pour déduire les propriétés de l'estimateur $\widehat{\beta}^{(J)}$ de celles de $\widehat{\beta}$:

Lemme 3. *Supposons que b/a est bornée et Lipschitz, $a(.) \geq \epsilon > 0$, b et a Lipschitz, alors pour tout $p \geq 1$ et tout $k = 1, \ldots, N$, il existe une constante C telle que*

$$\mathbb{E}_{\beta_0}(|V_k - V_k^J|^p + |U_k - U_k^J|^p) \leq \frac{C}{J^{p/2}}.$$

On en déduit la proposition suivante.

***Proposition* 12.** *Si $J \to +\infty$, alors $\widehat{\beta} - \widehat{\beta}^{(J)} = o_{P_{\beta_0}}(1)$. Si $J = J(N) \to +\infty$ de sorte que $\frac{J}{N} \to +\infty$, alors $\sqrt{N}(\widehat{\beta} - \widehat{\beta}^{(J)}) = o_{P_{\beta_0}}(1)$.*

A ma connaissance, c'est la première fois que la consistance de l'EMV est démontrée pour les EDS linéaires à paramètres aléatoires, dans le cas d'observations en temps continu ou à temps discrets. Dans le cas des observations à temps discrets, on voit que la consistance est établie quand le nombre d'individus N et le nombre d'observations par individu J tendent vers l'infini.

Le comportement de l'estimateur discrétisé $\widehat{\beta}^{(J)}$ est illustré par une étude de simulation. Dans le cas simple où $b(x) = c\, a(x)$, nous considérons un processus Brownien avec dérive

$$dX_k(t) = \phi_k dt + \gamma\, dB_k(t), \quad X_k(0) = 0,$$

avec $\phi_k \sim \mathcal{N}(\mu, \omega^2)$, un processus Brownien géométrique $b(x) = x$, $a(x) = \gamma\, x$ et le modèle où $b(x) = \sqrt{1+x^2}$, $a(x) = \gamma\sqrt{1+x^2}$, avec γ connu. Les simulations illustrent l'amélioration de la précision des estimateurs $\widehat{\mu}$ et $\widehat{\omega}^2$ quand N augmente. L'augmentation de l'intervalle de temps T ne semble pas avoir d'influence sur les propriétés des estimateurs. Dans le cas plus général où $b(x) \neq ca(x)$, nous considérons un processus d'Ornstein-Uhlenbeck avec un paramètre aléatoire

$$dX_k(t) \quad = \quad \phi_k X_k(t)dt + \gamma\, dB_k(t), \quad X_k(0) = 0,$$

avec $\phi_k \sim \mathcal{N}(\mu, \omega^2)$, et deux paramètres aléatoires

$$dX_k(t) \quad = \quad (-\phi_k^1 X_k(t) + \phi_k^2)dt + \gamma\, dB_k(t), \quad X_k(0) = 0,$$

avec $\phi_k = (\phi_k^1, \phi_k^2)' \sim \mathcal{N}_2(\mu, \Omega)$, $\mu = (\mu_1, \mu_2)'$ et une matrice diagonale Ω d'éléments (ω_1^2, ω_2^2). Un processus plus complexe non linéaire en X

$$dX_k(t) \quad = \quad \phi_k X_k(t)dt + \gamma\sqrt{1 + X_k(t)^2}dB_k(t), \quad X_k(0) = 0,$$

avec $\phi_k \sim \mathcal{N}(\mu, \omega^2)$ est aussi étudié. Ces trois simulations illustrent également les bonnes propriétés de l'EMV.

3.3 Equation différentielle stochastique à paramètre aléatoire observée avec bruit

Dans cette section, nous considérons l'estimation des paramètres $\theta = (\beta, \gamma, \sigma)$ du modèle (3.4) où les processus X_k sont observés à temps discrets et avec un bruit de mesure.

3.3.1 Approche bayésienne

Dans **[A10]**, notre objectif est de montrer l'apport de la modélisation par EDS par rapport à un modèle déterministe dans le cadre d'analyse de données longitudinales. Nous analysons des données de croissance d'animaux, issues d'une étude dont le but est de différentier différents phénotypes en terme de croissance. Les données de croissances sont classiquement modélisées par une fonction de régression monotone telle que la fonction de Weibull, de Gompertz, ou de Richards. Les données étudiées dans ce travail ont été analysées par cette approche dans Jaffrézic *et al.* (2006). Cependant, ces fonctions monotones ne sont pas forcément adaptées à la modélisation du véritable processus de croissance, la prise de poids de certains individus pouvant être parfois ralentie ou même décroître par des causes internes ou externes imprévisibles. Une mauvaise prise en compte de ces phénomènes dans la modélisation peut affecter le reste de l'analyse (biais dans les effets de covariables génétiques, etc). Afin de contourner ce problème, on peut proposer des fonctions déterministes plus complexes. Mais si le phénomène sous-jacent possède un caractère aléatoire, il ne sera pas pris en compte. Une alternative est de considérer une EDS directement déduite de la fonction de régression déterministe utilisée.

Plus précisément, nous analysons les données de croissance de $N = 50$ poulets, dont le poids est mesuré au temps $t_j = 0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36, 40$ semaines après la naissance. Dans Jaffrézic *et al.* (2006), ces données sont analysées par un modèle mixte avec une fonction de régression de Gompertz $f(\phi, t) = \phi_1 \exp(-\phi_2 e^{-\phi_3 t})$ et un modèle d'erreur hétéroscédastique. Nous considérons donc un modèle sur le logarithme des observations $\mathbf{Y}$ avec un modèle d'erreur additif homoscédastique

$$\begin{aligned} \log(Y_{kj}) &= \log\phi_{k1} - \phi_{k2}e^{-\phi_{k3}t_{kj}} + \varepsilon_{kj}, \quad k = 1, \ldots, N, j = 0, \ldots, J_k \\ \varepsilon_{kj} &\sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ \phi_k &= (\log\phi_{k1}, \phi_{k2}, \log\phi_{k3}) \sim_{i.i.d.} \mathcal{N}(\mu, \Omega). \end{aligned}$$

Une log-paramétrisation est utilisée pour les paramètres ϕ_{k1} et ϕ_{k3}, afin d'assurer la positivité des paramètres. On note $\mu = (\log(\phi_1), \phi_2, \log(\phi_3))$. La fonction de Gompertz $f(\phi, t)$ est solution de l'EDO

$$f'(t) = \phi_2\phi_3 e^{-\phi_3 t} f(t), \quad f(0) = \phi_1 e^{-\phi_2}.$$

On peut donc en déduire une EDS

$$dZ(t) = \phi_2\phi_3 e^{-\phi_3 t} Z(t)dt + a_\gamma(Z(t))dB(t),\ Z_0 = \phi_1 e^{-\phi_2}, \tag{3.7}$$

où la fonction de volatilité $a_\gamma(Z(t))$ doit être choisie. Etant donné le caractère hétéroscédastique du processus observé, nous proposons une volatilité de la forme $a_\gamma(Z(t)) = \gamma Z(t)$. Comme le modèle mixte porte sur le log des observations $\mathbf{Y}$, nous introduisons le processus $X(t) = \log Z(t)$ qui est solution de l'EDS

$$dX(t) = (\phi_2\phi_3 e^{-\phi_3 t} - \frac{1}{2}\gamma^2)dt + \gamma dB(t).$$

Cette EDS est non-homogène en temps. Le modèle mixte défini par EDS considéré est donc le suivant

$$\begin{aligned}
\log(Y_{kj}) &= X_k(t_{kj}) + \varepsilon_{kj}, \quad k = 1, \dots, N, j = 0, \dots, J_k \\
dX_k(t) &= (\phi_{k2}\phi_{k3} e^{-\phi_{k3} t} - \frac{1}{2}\gamma^2)dt + \gamma dB_k(t), X_k(0) = \log\phi_{k1} - \phi_{k2}, \\
\varepsilon_{kj} &\sim_{i.i.d.} \mathcal{N}(0, \sigma^2), \\
\phi_k &= (\log\phi_{k1}, \phi_{k2}, \log\phi_{k3}) \sim_{i.i.d.} \mathcal{N}(\mu, \Omega),
\end{aligned}$$

qui rentre dans la classe des modèles définis par (3.4). On suppose Ω diagonale, d'éléments diagonaux $\omega_1^2, \omega_2^2, \omega_3^2$. Les paramètres inconnus sont $\theta = (\mu, \Omega, \gamma, \sigma^2)$.

Dans ce travail et dans la continuité du travail de Jaffrézic *et al.* (2006), nous optons pour un cadre d'estimation bayésienne. On cherche donc à estimer la loi a posteriori $p(\theta|\mathbf{Y})$ définie en fonction de la distribution a priori $p(\theta)$ par

$$p(\theta|\mathbf{Y}) = \frac{L(\mathbf{Y};\theta)p(\theta)}{p(\mathbf{Y})},$$

où $L(\mathbf{Y};\theta)$ est la vraisemblance (3.5) et $p(\mathbf{Y}) = \int L(\mathbf{Y};\theta)d\theta$ est la constante de normalisation. La loi a posteriori $p(\theta|\mathbf{Y})$ n'est pas explicite. On recourt alors un algorithme de Gibbs pour l'estimer. Nous spécifions les lois a priori des paramètres θ comme suit : une loi normale pour les composantes du vecteur μ, une loi inverse Wishart pour la matrice Ω et une inverse Gamma pour le paramètre σ^2. Le paramètre γ^2 est un paramètre de variance. Plusieurs lois a priori sont testées : une loi uniforme, inverse-Gamma et de Jeffreys.

Nous proposons ensuite un algorithme d'estimation de Gibbs incluant la simulation des variables auxiliaires ϕ_k et $X(t_{kj})$:

- ETAPE 1 : initialisation du compteur de la chaîne à $m = 1$ et des valeurs initiales $\sigma^{-2(0)}, \gamma^{2(0)}, \mu^{(0)}, \phi^{(0)}, \mathbf{X}^{(0)}$.
- ETAPE 2 : simulation de $\sigma^{-2(m)}, \gamma^{2(m)}, \mu^{(mk)}, \phi^{(m)}, \mathbf{X}^{(m)}$ à partir des générations successives de
  1. $\mathbf{X}^{(m)} \sim p(\mathbf{X}|\phi^{(m-1)}, \gamma^{-2(m-1)}, \sigma^{-2(m-1)}, \mathbf{Y})$
  2. $\phi^{(m)} \sim p(\phi|\sigma^{-2(m-1)}, \gamma^{-2(m-1)}, \mu^{(m-1)}, \Omega^{(m-1)}, \mathbf{X}^{(m)}, \mathbf{Y})$
  3. $\mu^{(m)} \sim p(\mu|\phi^{(m)})$ et $\Omega^{(m)} \sim p(\Omega|\phi^{(m)})$
  4. $\sigma^{-2(m)} \sim p(\sigma^{-2}|\mathbf{X}^{(m)}, \phi^{(m)}, \mathbf{Y})$ and $\gamma^{-2(m)} \sim p(\gamma^{-2}|\mathbf{X}^{(m)}, \phi^{(m)})$
- ETAPE 3 : changer m en $m + 1$ et reprendre l'ÉTAPE 2 jusqu'à convergence.

Dans le cas de l'EDS basée sur le modèle de Gompertz, la densité de transition du processus $(X(t))$ est explicite. Cela permet d'obtenir des calculs exacts de certaines lois conditionnelles. La simulation réalisée est alors exacte. Pour les lois non explicites, un algorithme de Metropolis-Hastings est utilisé.

Une étude de simulation est réalisée pour comparer les modèles mixtes définis par EDO et EDS. 100 jeux de données avec $N = 50$ individus et $J = 9$ mesures par individus sont simulés avec les deux

Modèle de simulation	vraie	EDO ($\gamma^2 = 0$)		EDS ($\gamma^2 = 1$)	
Modèle d'estimation	valeur	EDO	EDS	EDO	EDS
$\log \phi_1$	8.01	8.00 (0.04)	8.03 (0.06)	7.84 (0.07)	8.02 (0.09)
ϕ_2	5.00	4.99 (0.08)	5.02 (0.08)	4.83 (0.09)	5.01 (0.11)
$\log \phi_3$	2.64	2.64 (0.04)	2.63 (0.04)	2.69 (0.05)	2.63 (0.05)
ω_1^{-2}	100.00	122.24 (39.95)	160.84 (27.84)	8.63 (3.08)	113.58 (29.25)
ω_2^{-2}	100.00	106.70 (22.17)	103.16 (23.74)	87.38 (34.72)	103.50 (24.02)
ω_3^{-2}	100.00	126.27 (45.69)	131.03 (55.02)	125.31 (47.20)	114.53 (47.69)
γ^2		- (-)	0.19 (0.02)	- (-)	0.96 (0.25)
σ^{-2}	5.00	5.05 (0.39)	5.35 (0.43)	3.67 (0.26)	5.12 (0.40)

Tab. 3.1 – Moyennes des estimateurs (et erreurs standards) obtenus à partir de l'analyse par modèle mixte défini par EDO et EDS de 100 jeux de données simulés avec un modèle mixte défini par EDO ou EDS.

modèles. Puis, les deux algorithmes de Gibbs adaptés pour les modèles mixtes définis par EDO et EDS respectivement sont utilisés pour estimer les distributions a posteriori. Les résultats sont présentés dans la table 3.1. Quand les données sont simulées à partir d'un modèle déterministe, les estimations obtenues par les deux algorithmes sont satisfaisantes. La valeur estimée du paramètre γ du modèle EDS est alors petite (0.19). Quand les données sont simulées à partir du modèle EDS, les estimations obtenues par l'algorithme de Gibbs du modèle EDS sont très satisfaisantes. En revanche, les estimations obtenues par l'algorithme de Gibbs du modèle déterministe peuvent être très biaisées (le paramètre ω_1^{-2} est estimé à 8.63 alors que la vraie valeur est 100).

L'analyse des données réelles fournit aussi des résultats satisfaisants. L'estimation du paramètre γ est strictement positive et son intervalle de crédibilité ne contient pas 0. Une comparaison des modèles déterministe et EDS par le critère bayésien DIC indique clairement une meilleure capacité prédictive du modèle EDS. Le calcul des lois a posteriori prédictives est aussi en faveur du modèle EDS. Sur ce jeu de données, le modèle EDS doit donc être préféré au modèle déterministe. Une illustration de ce phénomène peut être observée sur la figure 3.1 où les trajectoires individuelles de 4 individus sont représentées, avec la prédiction obtenue par le modèle mixte déterministe, la prédiction moyenne obtenue par le modèle EDS et un intervalle de confiance à 95% calculé empiriquement à partir de 1000 trajectoires simulées par l'algorithme de Gibbs. On voit que les sujets 4 et 13 sont des exemples d'individus pour lesquels on n'observe pas de ralentissement de la croissance. Les modèles EDO et EDS proposent tous les deux des prédictions satisfaisantes des données. Pour le sujet 1, on observe une perte de poids importante, elle est plus faible chez le sujet 14. Le modèle déterministe ne parvient pas à capter ces phénomènes alors que le modèle EDS fournit une prédiction tout à fait satisfaisante.

Enfin, nous proposons aussi dans **[A10]** un algorithme de Gibbs pour les modèles mixtes définis par EDS dans le cas où la densité de transition $p(X_k(t)|X_k(0);\phi_k)$ de l'EDS n'est pas explicite. Notre approche est basée sur une approximation de la solution de l'EDS par un schéma d'Euler-Maruyama de pas h. L'erreur introduite par le pas h sur la loi a posteriori est étudiée. Cependant sa mise en place est complexe car il faut introduire une grille de temps intermédiaires, de pas h, qui complexifie l'algorithme MCMC.

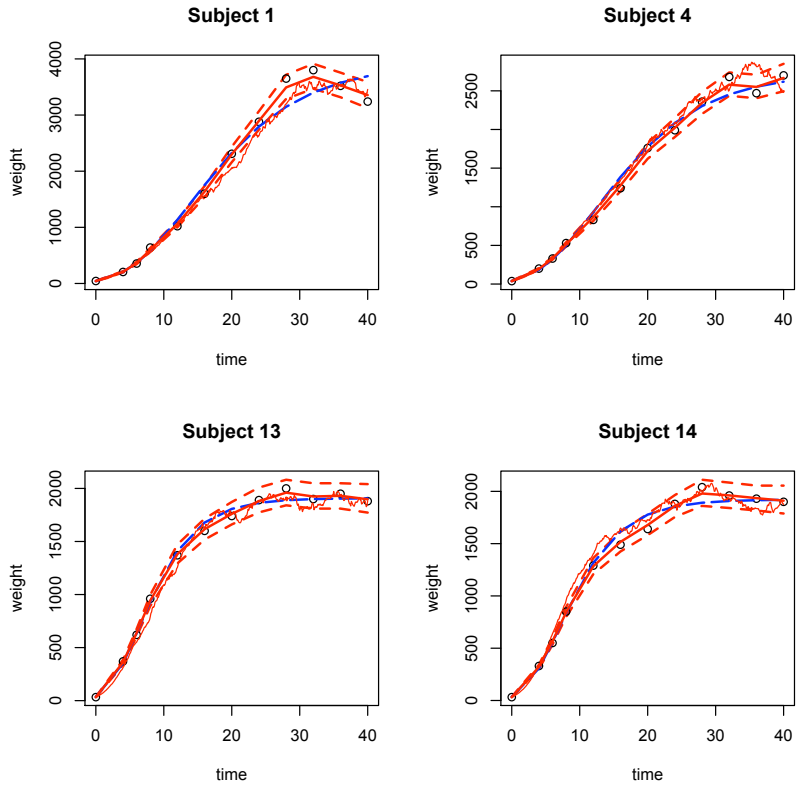

FIG. 3.1 – Données réelles de croissance du poids de 4 individus. Observations ($\circ$), prédictions obtenues par le modèle mixte déterministe (ligne en long pointillés), la prédiction moyenne du modèle mixte défini par EDS (ligne pleine) et l'intervalle de crédibilité à 95% obtenu par le modèle mixte défini par EDS (ligne pointillée). Une réalisation de l'EDS est aussi représentée (ligne pleine).

3.3.2 Approche par maximum de vraisemblance via l'algorithme SAEM-MCMC

Dans cette section et la suivante, je présente deux travaux réalisés en collaboration avec Sophie Donnet qui portent sur le modèle (3.4)

$$\begin{aligned} Y_{kj} &= X_k(t_{kj}) + \varepsilon_{kj}, \quad k = 1, \ldots, N, j = 1 \ldots, J_k, \quad \varepsilon_{kj} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ d\,X_k(t) &= b(X_k(t), \phi_k)dt + a_\gamma(X_k(t))\, dB_k(t), \\ \phi_k &\sim_{i.i.d.} g(\phi, \beta). \end{aligned}$$

L'estimation des paramètres $\theta = (\beta, \gamma, \sigma)$ de ce modèle a été peu étudiée, hormis par des approches de linéarisation de la solution de l'EDS.

Dans **[A5]**, nous considérons l'estimation de θ dans le cas où la densité de transition $p(X_k(t)|X_k(0); \phi_k)$ de l'EDS n'est pas connue. Pour cela, nous introduisons une pseudo-vraisemblance, correspondant à l'approximation des processus $X_k(t)$ par un schéma d'Euler-Maruyama. Nous proposons ensuite de maximiser cette pseudo-vraisemblance par l'algorithme SAEM-MCMC. Nous étudions enfin l'erreur due à l'introduction de la pseudo-vraisemblance sur l'estimateur obtenu. Cette approche est détaillée ci-dessous.

Pour simplifier les notations, on considère qu'on dispose du même nombre J d'observations chez tous les individus, aux temps $t_{kj} = t_j$ compris dans l'intervalle de temps $[0, T = t_J]$. On se restreint aussi au cas où la fonction de volatilité a_γ est constante et égale à γ. On introduit une grille de temps intermédiaires. On note $t_0 = \tau_0 < \tau_1 < \ldots < \tau_\ell < \ldots < \tau_M = T$ une discrétisation de l'intervalle de

temps $[0,T]$. On suppose que pour tout $j=0,\dots,J$, il existe une suite d'entiers ℓ_j tels que $t_j=\tau_{\ell_j}$. On note $(h_\ell)_{1\leq\ell\leq L}$ la suite de pas définis par $h_\ell=\tau_\ell-\tau_{\ell-1}$. On note $h=\max_{1\leq\ell\leq L}h_\ell$ le pas maximum. On peut alors définir le processus approché $\tilde{X}_k$ obtenu par le schéma d'Euler-Maruyama pour un individu k et le paramètre ϕ_k par le schéma itératif suivant. Pour $\ell=0$, $\tilde{X}_{k,0}=x^k$, et pour $\ell=1,\dots,L$,

$$\begin{aligned} h_\ell &= \tau_\ell-\tau_{\ell-1}\,, \\ \tilde{X}_{k,\ell} &= \tilde{X}_{k,\ell-1}+h_\ell\, b(\tilde{X}_{i,\ell-1},\phi_k)+\gamma\sqrt{h_\ell}\,\xi_\ell\,, \\ \xi_\ell &\sim_{i.i.d} \mathcal{N}(0,1). \end{aligned} \tag{3.8}$$

On introduit alors un modèle mixte qui est une approximation du modèle (3.4) défini à partir de l'approximation $\tilde{X}(t)$.

$$\begin{aligned} Y_{kj} &= \tilde{X}_{k,\ell_j}+\varepsilon_{kj}, \qquad 1\leq k\leq N, 0\leq j\leq J\,, \\ \varepsilon_{kj} &\sim_{i.i.d} \mathcal{N}(0,\sigma^2)\,, \\ \phi_k &\sim_{i.i.d} g(\phi;\beta)\,, \\ \tilde{X}_{k,\ell} &= \tilde{X}_{k,\ell-1}+h_\ell\, b(\tilde{X}_{i,\ell-1},\phi_k)+\gamma\sqrt{h_\ell}\,\xi_\ell, \quad \xi_\ell\sim_{i.i.d}\mathcal{N}(0,1), \quad \ell=1,\dots,L. \end{aligned} \tag{3.9}$$

On appelle $\mathcal{M}_h$ ce modèle dans la suite. On note par un indice h les lois se référant à ce modèle.

L'estimation des paramètres θ est réalisée en maximisant la pseudo-vraisemblance L_h se rapportant au modèle $\mathcal{M}_h$. La pseudo-vraisemblance n'étant pas explicite, la maximisation est réalisée par l'algorithme SAEM-MCMC en considérant $(\mathbf{Y},\tilde{X},\phi)$ comme le vecteur des données complètes.

On suppose que la pseudo-vraisemblance des données complètes vérifie l'hypothèse **(M)** (famille exponentielle). L'étape de simulation de l'algorithme SAEM-MCMC consiste donc à générer une chaîne de Markov $(\tilde{X}^{(m)},\phi^{(m)})$ ayant pour loi stationnaire la loi $p_h(\tilde{X},\phi|\mathbf{Y};\widehat{\theta}_m)$. Nous proposons un algorithme de Metropolis-Hastings-Within-Gibbs pour réaliser cette simulation.

***Algorithme* 4** (Algorithme SAEM-MCMC pour EDS mixtes)**.** *A l'itération $m\geq 1$, pour la valeur courante $\widehat{\theta}_m$ du paramètre*

Etape S : *Simulation de $(\tilde{X}^{(m)},\phi^{(m)})$ via la simulation d'une chaîne de Markov ayant pour loi stationnaire la distribution conditionnelle $p_h(\tilde{X},\phi|\mathbf{Y};\widehat{\theta}_m)$ par algorithme de Gibbs : pour chaque individu k, itération des deux étapes suivantes :*

1. *Simulation de $\phi_k^{(m)}$ par un algorithme de Metropolis-Hastings ayant q_1 comme densité instrumentale et $p_h(\phi\,|\mathbf{Y}_k,\tilde{X}_k^{(m-1)};\widehat{\theta}_m)$ comme loi invariante,*
2. *Simulation de $\tilde{X}_k^{(m)}$, avec un algorithme de M-H ayant q_2 comme densité instrumentale et $p_h(\tilde{X}\,|\mathbf{Y}_k,\phi_k^{(m)};\widehat{\theta}_m)$ comme loi invariante,*

Etape SA : *Approximation stochastique de $\mathbb{E}(S(\mathbf{Y},\tilde{X},\phi)|\mathbf{Y},\widehat{\theta}_m)$*

$$s_{m+1}=s_m+\gamma_m(S(\mathbf{Y},\tilde{X}^{(m)},\phi^{(m)})-s_m)$$

où (γ_m) est une suite de pas décroissants vers 0 tels que $\sum_{m\geq1}\gamma_m=\infty$ et $\sum_{m\geq1}\gamma_m^2<\infty$,

Etape M : *Actualisation du paramètre*

$$\widehat{\theta}_{m+1}=\arg\max_{\theta\in\Theta}-\Psi(\theta)+\langle s_{m+1},\nu(\theta)\rangle.$$

Le choix des lois instrumentales q_1 et q_2 est important pour assurer la convergence de l'algorithme et son ergodicité. Pour q_1, nous proposons de générer les candidats ϕ à l'aide d'une marche aléatoire. Pour q_2, nous proposons deux distributions. La première est la loi du schéma d'Euler. Pour la seconde, nous proposons de diviser le vecteur $\tilde{X}_k$ en deux, $(\tilde{X}_k(t_1),\dots\tilde{X}_k(t_J))$ le processus aux temps d'observations t_j et le vecteur $\tilde{X}_{k,aux}$ regroupant les temps auxiliaires (τ_ℓ). La distribution instrumentale pour simuler

les $(\tilde{X}_k(t_1), \dots \tilde{X}_k(t_J))$ candidats est une marche aléatoire. La distribution instrumentale pour simuler les $\tilde{X}_{k,aux}$ candidats est un pont Brownien.

Nous montrons la convergence de l'algorithme SAEM-MCMC vers un maximum de la pseudo-vraisemblance $L_h(\mathbf{Y};\theta)$, qui correspond à la vraisemblance du modèle $\mathcal{M}_h$. Nous étudions ensuite l'erreur introduite par le schéma d'Euler-Maruyama sur la vraisemblance dans le théorème suivant

Théorème 8. *Supposons que la fonction de dérive b est infiniment différentiable et que ses dérivées de tous ordres sont uniformément bornées.*

1. *Soient $p(X,\phi|\mathbf{Y})$ et $p_h(\tilde{X},\phi|\mathbf{Y})$ les distributions conditionnelles des processus $(X(t),\phi)$ et $(\tilde{X}(t),\phi)$ sur les modèles (3.4) et $\mathcal{M}_h$ respectivement. Alors il existe une constante $C(\mathbf{Y})$ dépendant de $\mathbf{Y}$ et une constante H_0 telle que pour tout $0 < h < H_0$,*

$$\left\| p(X,\phi|\mathbf{Y}) - p_h(\tilde{X},\phi|\mathbf{Y}) \right\|_{TV} \leq C(\mathbf{Y})h.$$

2. *Soient L et L_h les vraisemblances des modèles (3.4) et $\mathcal{M}_h$ respectivement. On suppose que la vraie valeur γ_0 du paramètre γ est telle que $\gamma_{min} < \gamma_0 < \gamma_{max}$. Alors il existe une constante $C_2(\mathbf{Y})$ dépendant de $\mathbf{Y}$ et indépendante de θ telle que pour tout $0 < h < H_0$,*

$$\sup_{\{\theta=(\beta,\gamma^2,\sigma^2),\gamma_{min}<\gamma<\gamma_{max}\}} |L(\mathbf{Y};\theta) - L_h(\mathbf{Y};\theta)| \leq C_2(\mathbf{Y})h.$$

La preuve est basée sur la vitesse de convergence de la densité de transition approchée par le schéma d'Euler, vitesse étudiée par Bally et Talay (1996).

Sous l'hypothèse **(H)** sur les hessiennes des vraisemblances des deux modèles, hypothèse déjà utilisée dans la section 1.2.3, on peut en déduire un résultat sur les maxima de vraisemblance des deux modèles.

***Corollaire* 2.** *On suppose que l'hypothèse **(H)** et les hypothèses du Théorème 8 sont vérifiées. Alors, $(\widehat{\theta}_m)_{m\geq 1}$ converge presque surement vers $\theta_{h,\infty}$ et il existe une constante C_3, indépendante de θ telle que*

$$\|\theta_{h,\infty} - \theta_\infty\|^2 \leq C_3 h.$$

Une étude de simulation dans le contexte de la pharmacocinétique illustre la convergence de l'algorithme SAEM-MCMC proposé. Une application sur les donnés réelles de pharmacocinétique de la Théophylline montre l'apport de la modélisation par EDS par rapport à un modèle pharmacocinétique d'EDO.

La principale difficulté de cette approche est le choix du paramètre h, c'est-à-dire le nombre de temps intermédiaires introduits dans le modèle. En effet, plus h est petit, plus l'espace sur lequel est réalisé l'algorithme MCMC est grand. Outre les temps de calcul qui peuvent s'avérer très longs, on peut aussi être confronté à des problèmes de mélange de la chaîne de Markov générée par l'algorithme MCMC, qui explore mal l'espace multidimensionnel sur lequel elle est définie. Une alternative est présentée dans la section suivante.

3.3.3 Approche par maximum de vraisemblance via un filtre particulaire

L'algorithme SAEM-MCMC proposé précédemment ne tient pas compte du caractère temporel des données manquantes $(X(t_{kj}))$. Un outil plus adapté pour ce type de données est l'approche par filtrage particulaire (algorithme SMC), qui a été particulièrement utilisé pour l'estimation dans les modèles à espace d'états sans paramètre aléatoire. Ces techniques sont cependant difficiles à adapter au cas

de paramètres aléatoires (Casarin et Marin, 2009). Récemment, Andrieu *et al.* (2010) ont développé un algorithme puissant, le Particle Markov Chain Monte Carlo (PMCMC), combinant les approches MCMC et SMC, dont la chaîne de Markov générée converge vers la loi exacte d'intérêt. Nous proposons dans **[S5]** de coupler l'algorithme SAEM à l'algorithme PMCMC pour l'estimation des paramètres θ du modèle mixte défini par EDS (3.4).

Dans un premier temps, nous considérons le cas où la densité de transition de l'EDS est explicite. L'algorithme SAEM est utilisé sur le modèle exact (3.4), en considérant $(\mathbf{Y}, X, \phi)$ comme vecteur des données complètes. L'étape de simulation consiste alors à simuler des réalisations de (X_k, ϕ_k) sous la loi conditionnelle $p(X_k, \phi_k|\mathbf{Y}_k; \widehat{\theta}_m)$. L'algorithme PMCMC permet de réaliser cette étape. Il utilise de façon astucieuse l'algorithme SMC, qui a déjà été évoqué dans la section 2.2.

Pour $j = 1, \ldots, J$, on note $X_{k0:j} = (X_{k0}, \ldots, X_{kj})$. L'algorithme SMC produit un ensemble de L particules $(X^{(\ell)}_{k0:J})_{\ell=1\ldots L}$ de poids respectifs $(W^{(\ell)}_{k0:J})_{\ell=1\ldots L}$ approchant la loi conditionnelle $p(X_{k0:J}|Y_{k0:J}, \phi_k; \gamma, \sigma)$ par une mesure empirique

$$\Psi^L_J = \sum_{\ell=1}^{L} W_{k0:J}(X^{(\ell)_{k0:J}} \mathbb{1}_{X^{(\ell)}_{k0:J}}.$$

La simulation d'une trajectoire $X_{k0:J}$ sous la loi approchée de $p(X_{k0:J}|Y_{k0:J}, \phi_k; \gamma, \sigma)$ est réalisée en choisissant aléatoirement une particule parmi les L particules de poids $(W_{k0:J}(X^{(\ell)_{k0:J}})_{\ell=1\ldots L}$. La distribution marginale $p(Y_{k0:J}|\phi_k; \gamma, \sigma)$ (où on a intégré en X) peut être estimée par

$$\widehat{p}^L(Y_{k0:J}|\phi_k; \gamma, \sigma) = \frac{1}{L}\sum_{\ell=1}^{L} w_0\left(X^{(\ell)}_{k0}\right) \prod_{j=1}^{J}\left(\frac{1}{L}\sum_{\ell=1}^{L} w_j\left(X^{(\ell)}_{k0:j}\right)\right). \tag{3.10}$$

L'algorithme PMCMC proposé par Andrieu *et al.* (2010) est le suivant :

***Algorithme* 5 (Algorithme PMCMC).**

- Initialisation *: en démarrant de $\phi^{(0)}_k$, simulation de $X^{(0)}_{k0:J}$ via un algorithme SMC avec L particules de loi cible $p(X_{k0:J}|Y_{k0:J}, \phi^{(0)}_k; \gamma, \sigma)$ et estimation de $p(Y_{k0:J}|\phi^{(0)}_k; \gamma, \sigma)$ par $\widehat{p}^L(Y_{k0:J}|\phi^{(0)}_k; \gamma, \sigma)$,*
- A l'itération $r \geq 1$
  1. *Simulation d'un candidat $\phi^c_k \sim q(\cdot|\phi^{(r-1)}_k)$,*
  2. *Simulation par algorithme SMC avec L particules de $X^c_{k0:J}$ ciblant la loi $p(\cdot|Y_{k0:J}, \phi^c_k; \gamma, \sigma)$ et calcul de $\widehat{p}^L(Y_{k0:J}|\phi^c_k; \gamma, \sigma)$ estimant $p(Y_{k0:J}|\phi^c_k; \gamma, \sigma)$,*
  3. *Mise à jour de $\left(X^{(r)}_{k0:J}, \phi^{(r)}_k\right) = (X^c_{k0:J}, \phi^c_k)$ et $\widehat{p}^L(Y_{k0:J}|\phi^{(r)}_k; \gamma, \sigma) = \widehat{p}^L(Y_{k0:J}|\phi^c_k; \gamma, \sigma)$ avec probabilité*

$$\widehat{\rho}^L(X^c_{k0:J}, \phi^c_k|X^{(r-1)}_{k0:J}, \phi^{(r-1)}_k) = \min\left\{1, \frac{q(\phi^{(r-1)}_k|\phi^c_k)}{q(\phi^c_k|\phi^{(r-1)}_k)} \frac{\widehat{p}^L(Y_{k0:J}|\phi^c_k; \gamma, \sigma)p(\phi^c_k; \beta)}{\widehat{p}^L(Y_{k0:J}|\phi^{(\ell-1)}_k; \gamma, \sigma)p(\phi^{(\ell-1)}_k; \beta)}\right\}.$$

  Si le candidat n'est pas accepté, alors $\left(X^{(r)}_{k0:J}, \phi^{(r)}_k\right) = \left(X^{(r-1)}_{k0:J}, \phi^{(r-1)}_k\right)$ et $\widehat{p}^L(Y_{k0:J}|\phi^{(r)}_k; \gamma, \sigma) = \widehat{p}^L(Y_{k0:J}|\phi^{(r-1)}_k; \gamma, \sigma)$.

La propriété la plus remarquable de l'algorithme PMCMC est que la distribution d'intérêt $p(X_{k0:J}, \phi_k|Y_{k0:J}; \gamma, \sigma)$ est laissée invariante par le noyau de transition de la chaîne, quelque soit le nombre de particules L. Plus précisément, l'algorithme PMCMC génère une suite $(X^{(r)}_{k0:J}, \phi^{(r)}_k)$ dont la loi marginale $\mathcal{L}^L(X^{(r)}_{k0:J}, \phi^{(r)}_k|Y_{0:J}; \theta)$ est telle que pour tout $\theta \in \Theta$ et tout $L > 0$,

$$||\mathcal{L}^L(X^{(r)}_{k0:J}, \phi^{(r)}_k|Y_{0:J}; \theta) - p(X_{k0:J}, \phi_k|Y_{k0:J}; \theta)||_{TV} \xrightarrow[\ell\to\infty]{} 0.$$

Dans **[S5]**, nous proposons d'utiliser l'algorithme PMCMC dans l'étape de simulation de l'algorithme SAEM pour fournir un algorithme SAEM-PMCMC. Grâce à la propriété de convergence de l'algorithme PMCMC, nous montrons la convergence suivante de l'algorithme SAEM-PMCMC.

Paramètres	$\log(\tau)$	κ	ω_τ	ω_κ	γ	σ
Vraie valeur	0.62	1.00	0.10	0.10	0.05	0.05
SAEM-PMCMC avec la distribution a posteriori $L = 25$						
Biais	0.26	0.11	0.32	-4.84	0.64	-0.41
RMSE	0.49	0.30	1.28	1.30	1.04	0.45
SAEM-PMCMC avec la distribution a posteriori, $L = 50$						
Biais	0.38	0.05	-2.55	-3.29	-0.84	0.07
RMSE	0.48	0.30	1.22	1.34	0.93	0.44
SAEM-PMCMC avec la distribution a posteriori, $L = 100$						
Biais	0.04	0.24	-0.97	-4.11	6.86	-2.68
RMSE	0.48	0.29	1.24	1.29	1.15	0.51
SAEM-PMCMC avec la distribution a priori, $L = 1000$						
Biais	-0.63	0.66	-2.78	-4.73	15.42	-5.44
RMSE	0.53	0.34	1.43	1.77	2.17	0.83

TAB. 3.2 – Modèle mixte d'Ornstein-Uhlenbeck, $N = 20$, $J = 40$: biais et RMSE (%) de $\widehat{\theta}$ obtenu par l'algorithme SAEM-PMCMC sur 100 données simulées. L'algorithme PMCMC est implémenté successivement avec la distribution a posteriori $q(X_{t_{kj}}|X_{t_{kj-1}}, Y_{kj}, \phi_k; \theta) = p(X_{t_{kj}}|X_{t_{kj-1}}, Y_{kj}, \phi_k; \theta)$ et $L = 25$, $L = 50$ ou $L = 100$ particules et la distribution a priori $q(X_{t_{kj}}|X_{t_{kj-1}}, Y_{kj}, \phi_k; \theta) = p(X_{t_{kj}}|X_{t_{kj-1}}, \phi_k\theta)$, $L = 1000$ particules.

Théorème 9. *Sous les hypothèses classiques de SAEM, la suite $\widehat{\theta}_m$ fournie par l'algorithme SAEM-PMCMC converge presque sûrement vers un maximum de la vraisemblance $L(\mathbf{Y}; \theta)$.*

Une étude de simulation illustre cette convergence avec un modèle mixte défini par un processus d'Ornstein-Uhlenbeck et un modèle d'erreur homoscédastique :

$$\begin{aligned} Y_{kj} &= X_{kj} + \varepsilon_{kj}, \;\; \varepsilon_{kj} \sim \mathcal{N}(0, \sigma^2), \\ dX_k(t) &= -\left(\frac{X_k(t)}{\tau_k} - \kappa_k\right) dt + \gamma dB_k(t), \quad X(0) = 0 \end{aligned}$$

où $\kappa_k \in \mathbb{R}$, $\tau_k > 0$. On pose $\phi_k = (\log(\tau_k), \kappa_k)$ le vecteur des paramètres individuels. On suppose que $\log(\tau_k) \sim_{i.i.d.} \mathcal{N}(\log(\tau), \omega_\tau^2)$, $\kappa_k \sim_{i.i.d.} \mathcal{N}(\kappa, \omega_\kappa^2)$. Le vecteur des paramètres est $\theta = (\log\tau, \kappa, \omega_\tau, \omega_\kappa, \gamma, \sigma)$.

Le choix du nombre de particules L dans l'algorithme PMCMC ainsi que le choix de la distribution instrumentale q utilisée dans l'algorithme SMC sont étudiés. Les résultats sont présentés dans la table 3.2. Le choix de L a très peu d'influence. En revanche, le choix de la distribution instrumentale q de l'algorithme SMC peut avoir une grande influence. Le choix optimal est la loi conditionnelle aux observations $\mathbf{Y}$.

Un modèle plus complexe est considéré par simulation, qui correspond à une diffusion de Gompertz non-homogène en temps et un modèle d'erreur hétéroscedastique, similaire à celui proposé dans **[A10]** :

$$\begin{aligned} Y_{kj} &= X_{kj}(1 + \varepsilon_{kj}), \; \varepsilon_{kj} \sim \mathcal{N}(0, \sigma^2), \\ dX_k(t) &= \phi_{k2}\,\phi_{k3} e^{-\phi_{k3}t} X_k(t)dt + \gamma X_k(t) dB_k(t), \quad X_k(0) = \phi_{k1} e^{-\phi_{k2}} \end{aligned} \tag{3.11}$$

où $\phi_{k1} > 0, \phi_{k2} > 0, \phi_{k3} > 0$. On pose $\phi_k = (\log\phi_{k1}, \log\phi_{k2}, \log\phi_{k3})$. On suppose que $\log\phi_{k1} \sim_{i.i.d.} \mathcal{N}(\log\phi_1, \omega_1^2)$, $\log\phi_{k2} \sim_{i.i.d.} \mathcal{N}(\log\phi_2, \omega_2)$, $\log\phi_{k3} \sim_{i.i.d.} \mathcal{N}(\log\phi_3, \omega_3^2)$. Les paramètres d'intérêt sont $\theta = (\log\phi_1, \log\phi_2, \log\phi_3, \omega_1, \omega_2, \omega_3, \gamma, \sigma)$. Les propriétés des estimateurs fournis par SAEM-PMCMC restent très satisfaisantes. En particulier, la comparaison avec l'algorithme SAEM-MCMC proposé dans **[A5]** montre une nette amélioration pour l'estimation du paramètre de volatilité. La prise en

compte du caractère temporel du processus $X(t)$ dans l'étape de simulation de l'algorithme SAEM a donc toute son importance.

Enfin, nous considérons brièvement le cas plus complexe d'un modèle mixte dont l'EDS n'a pas de densité de transition connue. Alors, de façon similaire à **[A5]**, nous introduisons un modèle approché $\mathcal{M}_h$ basé sur l'approximation de l'EDS par schéma d'Euler-Maruyama. On peut alors montrer la convergence de l'algorithme SAEM-PMCMC vers un maximum de la pseudo-vraisemblance $L_h(\mathbf{Y};\theta)$ du modèle $\mathcal{M}_h$. Comme pour l'algorithme SAEM-MCMC, le nombre de points intermédiaires à introduire dans le schéma d'Euler peut s'avérer difficile à choisir.

Perspectives

Je présente dans ce dernier chapitre un bref panorama de mes perspectives de recherche. Elles s'articulent naturellement selon la construction de ce manuscrit.

Estimation dans les modèles mixtes

Dans la suite des travaux théoriques présentés dans le chapitre 1, j'aimerais continuer à travailler sur deux projets évoqués ci-dessous. Je continue aussi les applications en biologie des modèles mixtes, je présente ici des perspectives dans le cadre de la prédiction de la croissance foetale.

Estimation non paramétrique de la distribution des effets aléatoires
Dans la section 1.3, en collaboration avec Fabienne Comte, nous proposons une méthode d'estimation non paramétrique de la distribution des effets aléatoires d'un modèle linéaire mixte simple. Nous estimons séparément les densités des variables α et β. Une première extension serait d'envisager l'estimation de la densité bidimensionnelle du couple (α, β), la sélection de modèles étant alors beaucoup plus complexe. Ensuite, une extension naturelle serait de travailler sur un modèle linéaire mixte incluant des covariables binaires (sexe, groupe de traitement) ou continues (âge, poids, etc). Les transformations Z ou V des variables Y proposées dans **[S1]** ne sont pas applicables directement puisque les individus n'ont pas les mêmes valeurs de covariables, contrairement aux temps d'observations. Enfin, une autre perspective de travail serait de considérer des modèles non-linéaires mixtes. Dans ce cadre, excepté pour des fonctions de régression très particulières (fonction exponentielle), on ne peut pas envisager de transformer les variables d'observations pour se replacer dans un contexte de déconvolution. On pourrait alors combiner des méthodes d'estimation non paramétriques avec des algorithmes du type EM.

Données répétées et modèle d'erreur hétéroscedastique
Dans la section 1.3.2, avec Fabienne Comte et Julien Stirnemann, nous considérons un problème de déconvolution dans le cadre de mesures répétées. Nous supposons que le bruit de mesure est un bruit additif homoscédastique. Dans les applications biologiques, il est souvent plus réaliste de supposer un bruit de mesure hétéroscédastique du type

$$Y_j = X_j + (1 + X_j)\varepsilon_j.$$

Lorsque la densité f_ε est connue, des approches de déconvolution classiques peuvent être utilisées. Lorsque la densité du bruit est inconnue, on peut envisager de l'estimer à partir d'échantillon de mesures répétées. Cependant l'hypothèse de symétrie sur la densité f_ε faite dans le cas homoscédastique ne suffit pas dans ce contexte. D'autres hypothèses seront considérées.

Prédiction de la croissance foetale
Dans la section 1.4.2, en collaboration avec Julien Stirnemann, nous estimons la densité de la variable X, qui correspond à l'intervalle entre la date des dernières règles et le début de grossesse. Cette densité est la densité dans la population. On aimerait se servir de cette densité "populationnelle" pour affiner la prédiction de début de grossesse individuelle estimée chez une femme enceinte à partir de la première mesure échographique. Avec la même approche que celle utilisée pour construire des intervalles de prédiction individuels de croissance foetale, on voudrait estimer la distribution individuelle a posteriori

de date de début de grossesse à partir de la distribution estimée dans la population et d'un modèle de prédiction basée sur la mesure échographique. Une approche bayésienne pourrait à nouveau être utilisée.

Equations différentielles stochastiques en biologie

Modèle stochastique de pharmacocinétique
Dans la section 2.1, nous proposons un modèle de pharmacocinétique stochastique en incluant une perturbation brownienne additive sur chaque équation du système. La volatilité de ces perturbations est supposée constante. Il est plus probable que les perturbations aléatoires soient proportionnelles à la quantité d'agent de contraste dans le pixel. Un modèle plus réaliste reposerait sur des fonctions de volatilité dépendant du processus : $\sigma_1 Q_P(t) dB_1(t)$ ou $\sigma_1 \sqrt{Q_P(t)} dB_1(t)$. Le processus bidimensionnel considéré n'a alors pas nécessairement une solution explicite, comme c'était le cas pour le processus d'Ornstein-Uhlenbeck proposé. Cela complique l'estimation des paramètres de ce processus partiellement observé avec bruit. Cependant, en considérant un bruit de mesure lui aussi multiplicatif (hétéroscédastique), il est envisageable d'utiliser des filtres exacts qui ont été proposés par Chaleyat-Maurel et Genon-Catalot (2006). Il devrait être possible de calculer de façon exacte la fonction de vraisemblance et d'obtenir un estimateur du maximum de vraisemblance, dont les propriétés restent à étudier. Dans le cas d'un modèle d'erreur plus général, on pourrait également utiliser des filtres particulaires pour calculer et maximiser la vraisemblance du modèle. Les propriétés de l'EMV devraient pouvoir se déduire des propriétés de convergence des filtres particulaires.

Modèle neuronal observé avec bruit de mesure
Dans la section 2.2.1, en collaboration avec Jérôme Dedecker et Marie-Luce Taupin, nous proposons une méthode d'estimation d'un modèle auto-régressif observé avec un bruit de mesure. Afin de valider ou non cette hypothèse d'existence d'un bruit de mesure dans les données de potentiel neuronal recueillies expérimentalement, nous voulons mettre en place un test statistique d'existence du bruit de mesure. Nous testerons alors cette hypothèse sur les données neuronales fournies par Lee Moore (Laboratoire CESEM, Université Paris Descartes).

Une extension possible du modèle auto-régressif (2.3) est de supposer que le paramètre θ est aléatoire. Ces modèles auto-régressifs à paramètre aléatoire peuvent être vus comme une alternative à la construction d'EDS pour modéliser la variabilité d'un processus biologique. En effet, si on part d'un modèle déterministe d'EDO, qu'on le discrétise en temps afin d'obtenir un modèle auto-régressif déterministe, un moyen de modéliser la variabilité au cours du temps du processus biologique modélisé est de considérer les paramètres du modèle auto-régressif comme aléatoires. L'estimation des paramètres de ces modèles est donc une question tout à fait pertinente et enthousiasmante dans le but d'applications en biologie.

Modèle neuronal bidimensionnel partiellement observé
Dans la section 2.2.3, en collaboration avec Susanne Ditlevsen, nous proposons une méthode d'estimation combinant l'algorithme SAEM et une méthode de filtrage particulaire pour l'estimation paramétrique d'une EDS bidimensionnelle partiellement observée. Nous montrons la convergence de l'algorithme SAEM vers le maximum de la vraisemblance d'un modèle approché par un schéma d'Euler. L'étude des propriétés de l'EMV est complexe dans ce cadre. On peut montrer que la distance entre la vraisemblance du modèle approché (la pseudo-vraisemblance) et celle du modèle exact est bornée par $C^n \Delta$, où C est une constante qui peut être grande, n est le nombre d'observations et Δ est le pas de temps utilisé par le schéma d'Euler. Mais cette borne n'est pas suffisante. Une alternative serait d'utiliser les résultats de convergence de Del Moral et Jacod (2001) sur le filtre particulaire appliqué aux EDS. Leurs résultats lient le nombre de particules utilisées dans l'algorithme de filtrage au pas de temps utilisé pour approcher l'EDS par un schéma d'Euler. Dans l'algorithme d'estimation SAEM-SMC que nous proposons, il faudrait donc augmenter le nombre de particules au cours des itérations

de l'algorithme SAEM, ce qui aurait pour effet de modifier le modèle approché considéré à chaque itération. La convergence de cet algorithme reste donc une question ouverte.

Les premiers résultats obtenus sur données neuronales réelles sont tout à fait prometteurs. Une analyse plus poussée est envisagée, permettant de comparer le comportement neuronal dans différentes conditions expérimentales. Ce travail sera réalisé en collaboration avec Susanne Ditlevsen et Rune Berg, neurophysiologiste à l'université de Copenhague.

Méthode d'estimation pour EDS hypoelliptique
Dans la section 2.2.4, nous considérons l'estimation paramétrique d'un système différentiel stochastique hypoelliptique (2.6). L'application au modèle de Morris-Lecar stochastique (2.4) reste à faire. Ce système vérifie les conditions **(HE1)-(HE2)** et la transformation en une diffusion intégrée est donc possible. Cependant les fonctions de dérive a et b sont hautement non-linéaires en les paramètres. Les estimateurs n'ont pas de forme explicite et il faut utiliser une méthode numérique de minimisation du contraste. Une étude sur simulations sera nécessaire avant l'analyse des données réelles neuronales. Il sera alors intéressant de comparer les résultats avec ceux obtenus dans la section précédente.

Modèle mixte et équations différentielles stochastiques

Propriétés du maximum de vraisemblance pour des EDS à paramètres aléatoires
Dans la section 3.2, en collaboration avec Maud Delattre et Valentine Genon-Catalot, nous proposons des premiers résultats de consistance et de normalité asymptotique de l'estimateur du maximum de vraisemblance pour des EDS à paramètres aléatoires dont la fonction de dérive est linéaire en les paramètres. Ce travail s'est limité au cas où l'EDS est unidimensionnelle. L'extension au cas multidimensionnel ne devrait pas être problématique. En revanche, il serait intéressant d'étudier des EDS dont la dérive est non linéaire en ϕ, ce qui permettrait de couvrir un champ plus vaste de modèles. Si la loi des paramètres aléatoires n'est pas gaussienne, la vraisemblance n'est plus explicite et l'étude théorique des propriétés de l'EMV reste aussi une question ouverte. Enfin, cette approche ne permet pas d'estimer les paramètres de la fonction de volatilité. On pourrait envisager que ce coefficient de volatilité est décrit par un paramètre aléatoire, ayant une loi Gamma par exemple. Pour des modèles simples, la vraisemblance est explicite. L'étude des propriétés de cet estimateur reste à réaliser.

Méthode d'estimation pour modèles mixtes définis par EDS
Dans la section 3.3, en collaboration avec Sophie Donnet, nous proposons un algorithme SAEM-PMCMC quand la densité de transition de l'EDS est connue. Lorsque la densité n'est pas explicite, il est possible d'utiliser cet algorithme sur un modèle approché par un schéma d'Euler. Mais la convergence de l'algorithme vers le maximum de la vraisemblance exacte est complexe à montrer. Si on considère un algorithme "théorique" où l'étape de simulation de l'algorithme SAEM est réalisée par un algorithme SMC, alors, en utilisant des résulats de convergence de l'algorithme SMC, il est possible de montrer la convergence de l'algorithme SAEM-SMC, comme on l'a fait avec Susanne Ditlevsen. Si on suppose de plus que le paramètre de volatilité γ est connu, il est possible de montrer la convergence de l'algorithme SAEM-SMC vers le maximum de vraisemblance du modèle exact, en utilisant de façon astucieuse la forme de la vraisemblance complète du modèle mixte. Cependant, les algorithmes SMC sont connus pour ne pas être adaptés au cas de paramètres aléatoires. L'alternative est d'utiliser un algorithme PMCMC comme nous l'avons proposé. Mais la question de la convergence de l'algorithme vers le maximum de la vraisemblance exacte reste alors ouverte.

Estimation non paramétrique de la densité des effets aléatoires
Dans la continuité des travaux réalisés avec Fabienne Comte dans le cadre d'un modèle mixte (section 1.3), nous envisageons avec Fabienne Comte et Valentine Genon-Catalot de proposer des méthodes d'estimation de la densité des paramètres aléatoires d'une EDS à paramètres aléatoires. En se basant sur les statistiques exhaustives U_k et V_k étudiées dans **[S2]**, il est possible de se ramener à un problème de déconvolution pour les EDS simples où les fonctions de dérive et de volatilité sont égales. Des estimateurs par projection peuvent aussi être envisagés. Il sera alors intéressant de comparer les vitesses

de ces différents estimateurs, l'estimateur par déconvolution ayant l'avantage d'être à densité de bruit connue (par construction de l'EDS) mais gaussienne, ce qui implique des vitesses logarithmiques. L'extension aux EDS plus générales est plus complexe.

Liste de Publications

Articles publiés ou à paraître dans des revues avec comité de lecture

[A1] Samson, A., Lavielle, M., Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model : application to HIV dynamics model. *Comput. Stat. Data An.* **51**, 1562-74.

[A2] Donnet, S., Samson, A. (2007). Estimation of parameters in incomplete data models defined by dynamical systems. *J. Stat. Plan. Infer.* **137**, 2815-31.

[A3] Samson, A., Lavielle, M., Mentré, F. (2007). The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Stat. Med.* **26**, 4860-4875.

[A4] Retout, S., Comets, E., Samson, A., Mentré, F. (2007). Design in nonlinear mixed effects models : optimization using the Fedorov-Wynn algorithm and power of the Wald test for binary covariates. *Stat. Med.* **26**, 5162-5179.

[A5] Donnet, S., Samson, A. (2008). Parametric inference for mixed models defined by stochastic differential equations. *ESAIM P&S* **12**, 196-218.

[A6] Baron, G., Ravaud, P., Samson, A., Giraudeau, B. (2008). Missing data in randomized controlled trials of rheumatoid arthritis with radiographic outcomes : a simulation study. *Arthrit. Care Res.* **59**, 25-31.

[A7] Panhard, X., Samson, A. (2009). Extension of the SAEM algorithm for the nonlinear mixed models with two levels of random effects. *Biostatistics* **10**, 121-35.

[A8] Richard, F., Samson, A., Cuenod, C.A. (2009). A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. *Stat. Comput.* **19**, 465-478.

[A9] Bastogne, T., Samson, A., Vallois, P., Wantz-Mézières, S., Pinel, S., Bechet, D., Barberi-Heyob, M. (2010). Phenomenological modeling of tumor diameter growth based on a mixed effects model. *J. Theor. Biol.* **262**, 544-552.

[A10] Donnet, S., Foulley, J.L., Samson, A. (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics* **66**, 733–741.

[A11] Favetto, B., Samson, A. Parameter estimation for a bidimensional partially observed Ornstein-Uhlenbeck process with biological application. *Scand. J. Stat.* **37**, 200-220.

[A12] Lavielle, M., Samson, A., Fermin, A.K., Mentré, F. (2011). Maximum likelihood estimation of long term HIV dynamic models and antiviral response. *Biometrics* **67**, 250–259.

[A13] Cuenod, C.A,. Favetto, B., Genon-Catalot, V., Rozenholc, Y., Samson, A. (2011). Parameter estimation and change-point detection from Dynamic Contrast Enhanced MRI data using stochastic differential equations. *Math. Biosci.* **233**, 68–76.

[A14] Stirnemann, J., Samson, A., Thalabard, J.C. (2012) Individual predictions based on population nonlinear mixed modeling : application to prenatal twin growth. *Stat. Med.* to appear.

[A15] Stirnemann, J., Comte, F., Samson, A. (2012) Density estimation of a biomedical variable subject to measurement error using an auxiliary set of replicate observations. *Stat. Med.* to appear.

[A16] Whegang, S., Samson, A., Basco, L.K., Thalabard, J.C. (2012) Multiple Treatment Comparisons (MTC) in a series of antimalarial trials with an ordinal primary outcome and repeated measurements. *Malaria Journal*, 11(1) :147.

[A17] Samson A, Thieullen M. Contrast estimator for completely or partially observed hypoelliptic diffusion. (2012) *Stochastic Process. Appl.*, to appear.

Lecture notes

[L1] Ditlevsen, S., Samson, A. (2012). Introduction to stochastic models in biology. In Bachar, Batzel and Ditlevsen (Eds.), Stochastic Methods and Neuron Modeling. Springer.

Actes de conférences (Proceedings)

[C1] Bastogne, T., Samson, A., Mézières-Wantz, S., Vallois, P., Pinel, S., Barberi-Heyob, M. (2009). System identification of tumor growth described by a mixed effects model. *Proceedings of IFAC Symposium on system identification.*

[C2] Richard, F., Samson, A. (2007). Metropolis-Hasting techniques for finite element-based registration. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Articles soumis

[S1] Comte, F., Samson, A. Nonparametric estimation of random effects densities in linear mixed-effects model. Prepublication MAP5 2012-01 (soumis)

[S2] Delattre, M., Genon-Catalot, V., Samson, A. Maximum likelihood estimation for stochastic differential equations with random effects. Prepublication MAP5 2011-31 (soumis)

[S3] Dedecker, J., Samson, A., Taupin, M.L. Estimation in autoregressive model with measurement error. Prepublication MAP5 2011-18 (soumis)

[S4] Comte, F., Samson, A., Stirnemann, J. Deconvolution estimation of onset of pregnancy with replicate observations. Prepublication MAP5 2011-15 (soumis)

[S5] Donnet, S., Samson, A. EM algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. Prepublication MAP5 2010-24 (soumis)

[S6] Ditlevsen, S. Samson, A. Parameter estimation in the stochastic neuronal Morris-Lecar model with particle filter methods. (soumis)

Bibliographie

Andrieu, C., Doucet, A. et Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B* **72**, 269Ð342.

Antic, J., Laffont, C., Chafai, D. et Concordet, D. (2009). Comparison of nonparametric methods in nonlinear mixed effects models. *Comput. Stat. Data An.* **53**, 642–656.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *J. Time Ser. Ana.* **15**, 453–472.

Bally, V. et Talay, D. (1996). The law of the Euler scheme for stochastic differential equations (II) : convergence rate of the density. *Monte Carlo Methods Appl.* **2**, 93–128.

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

Butucea, C. et Tsybakov, A. B. (2008). Sharp optimality in density deconvolution with dominating bias. II. *Theor. Probab. Appl.* **52**, 237–249.

Casarin, R. et Marin, J. (2009). Online data processing : comparison of Bayesian regularized particle filters. *Electron. J. Stat.* **3**, 239–258.

Chafai, D. et Loubes, J.-M. (2006). On nonparametric maximum likelihood for a class of stochastic inverse problems. *Statist. Probab. Lett.* **76**, 1225–1237.

Chaleyat-Maurel, M. et Genon-Catalot, V. (2006). Computable infinite dimensional filters with applications to discretized diffusions. *Stoch. Proc. and Applic.* **116**, 1447–1467.

Comte, F. et Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *J. Roy. Stat. Soc. B* **73**, 601–627.

Comte, F., Rozenholc, Y. et Taupin, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Can. J. Stat.* **34**, 431–452.

Comte, F. et Taupin, M. (2001). Semiparametric estimation in the (auto)-regressive β-mixing model with errors-in-variables. *Math. Methods Statist.* **10**, 121–160.

Dedecker, J. et Prieur, C. (2005). New dependence coefficients. examples and applications to statistics. *Probab. Theory Relat.* **132**, 203–236.

Del Moral, P. et Jacod, J. (2001). Interacting particle filtering with discrete observations. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pp. 43–75. Springer, New York.

Delaigle, A., Hall, P. et Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.* **36**, 665–685.

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

Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Jr. R. Stat. Soc. B* **39**, 1–38.

Ding, A. et 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.

Ditlevsen, S. et De Gaetano, A. (2005). Stochastic vs. deterministic uptake of dodecanedioic acid by isolated rat livers. *Bull. Math. Biol.* **67**, 547–561.

Ditlevsen, S. et Greenwood, P. (2012). The Morris-Lecar neuron model embeds a leaky integrate-and-fire model. *Journal of Mathematical Biology* .

Doucet, A., de Freitas, N. et Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pp. 3–14. Springer, New York.

Duval, X., Mentré, F., Rey, E., Auleley, S., Peytavin, G., Biour, M., Métro, A., Goujard, C., Taburet, A., Lascoux, C., Panhard, X., Tréluyer, J. et Salmon-Céron, D. (2009). Benefit of therapeutic drug monitoring of protease inhibitors in HIV-infected patients depends on PI used in HAART regimen - ANRS 111 trial. *Fundam. Clin. Pharm.* **23**, 491–500.

Faugeras, O., Touboul, J. et Cessac, B. (2009). A constructive mean field analysis of multi population neural networks with random synaptic weights and stochastic inputs. *Front. Comput. Neurosci* **3**, 1–28.

Gloter, A. (2006). Parameter estimation for a discretely observed integrated diffusion process. *Scand. J. Statist.* **33**, 83–104.

Guedj, J., Thiébaut, R. et Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198–2006.

Höpfner, R. (2007). On a set of data for the membrane potential in a neuron. *Math. Biosci.* **207**, 275–301.

Höpfner, R. et Brodda, K. (2006). A stochastic model and a functional limit theorem for information processing in large systems of neurons. *Journal of Mathematical Biology* **52**, 439–457.

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

Jacqmin-Gadda, H., Thiebaut, R., Chene, G. et Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* **1**, 355–368.

Jaffrézic, F., Meza, C., Lavielle, M. et Foulley, J. (2006). Genetic analysis of growth curves using the SAEM algorithm. *Genet. Sel. Evol.* **38**, 583–600.

Jahn, P., Berg, R., J., H. et Ditlevsen, S. (2011). Motoneuron membrane potentials follow a time inhomogeneous jump diffusion process. *J. Comput. Neurosci.* **31**, 563–579.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, 211–229.

Kuhn, E. (2003). Maximum likelihood estimation in non linear inverse problems. Ph.D. thesis, Université Paris Sud - Paris XI.

Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S* **8**, 115–131.

Kuhn, E. et Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.* **49**, 1020–1038.

Lansky, P. et Ditlevsen, S. (2008). A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models. *Biol. Cybern.* **99**, 253–262.

Lavielle, M. et Mentré, F. (2007). Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. *J Pharmacokinet Pharmacodyn* **34**, 229–249.

Mallet, A., Mentré, F., Steimer, J. et Lokiec, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to cyclosporine. *J. Pharmacokinet. Pharmacodyn.* **16**, 311–327.

Meister, A. et Neumann, M. H. (2010). Deconvolution from non-standard error densities under replicated measurements. *Statist. Sinica* **20**, 1609–1636.

Neumann, M. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Statist.* **7**, 307–330.

Nie, L. et Yang, M. (2005). Strong consistency of the mle in nonlinear mixed-effects models with large cluster size. *Sankhya Ser. A* **67**, 736–763.

Nowak, M. et May, R. (2000). *Virus dynamics : mathematical principles of immunology and virology.* Oxford University Press.

Overgaard, R., Jonsson, N., Tornøe, C. et Madsen, H. (2005). Non-linear mixed-effects models with stochastic differential equations : Implementation of an estimation algorithm. *J Pharmacokinet. Pharmacodyn.* **32**, 85–107.

Perelson, A. et Nelson, P. (1997). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review* **41**, 3–44.

Picchini, U. et Ditlevsen, S. (2011). Practicle estimation of high dimensional stochastic differential mixed-effects models. *Comput. Stat. Data An.* **55**, 1426–1444.

Pinheiro, J. et Bates, D. (2000). *Mixed-effect models in S and Splus.* Springer-Verlag.

Pokern, Y., Stuart, A. et Wiberg, P. (2009). Parameter estimation for partially observed hypoelliptic diffusions. *J. Roy. Stat. Soc. B* **71**, 49–73.

Vogelstein, J. T., Watson, B. O., Packer, A. M., Yuste, R., Jedynak, B. et Paninski, L. (2009). Spike inference from calcium imaging using sequential monte carlo methods. *Biophys J* **97**, 636–655.

Wilcox, A. J., Dunson, D. et Baird, D. D. (2000). The timing of the "fertile window" in the menstrual cycle : day specific estimates from a prospective study. *BMJ (Clinical Research Ed.)* **321**, 1259–1262.