Stochastic Expectation Maximization algorithms for estimation in latent variable models: developments, analysis and applications.

Estelle Kuhn

To cite this version:

Estelle Kuhn. Stochastic Expectation Maximization algorithms for estimation in latent variable models: developments, analysis and applications.. Mathematics [math]. Université Paris Sud - Paris 11, 2015. tel-02795652

HAL Id: tel-02795652
https://hal.inrae.fr/tel-02795652

Submitted on 5 Jun 2020

Université Paris-Sud

Faculté des sciences d'Orsay
École doctorale de mathématiques Hadamard (ED 574)
Unité de Recherche INRA, MaIAGE (UR 1404)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Estelle KUHN

Stochastic Expectation Maximization algorithms for estimation in latent variable models : developments, analysis and applications.

Rapporteurs :

Christophe ANDRIEU
Didier CONCORDET
Pascal MASSART

Date de soutenance : 5 novembre 2015

Composition du jury :

Didier CONCORDET	(Rapporteur)
Florence FORBES	(Examinatrice)
Jean-Michel MARIN	(Examinateur)
Pascal MASSART	(Rapporteur)
Christian ROBERT	(Examinateur)
Stéphane ROBIN	(Examinateur)
Sophie SCHBATH	(Invitée)

Merci !

Mes premiers remerciements vont à mes trois rapporteurs : Christophe Andrieu, Didier Concordet et Pascal Massart. Je vous suis sincèrement reconnaissante d'avoir accepté de consacrer de votre temps précieux à la lecture de mon manuscrit. Je remercie également tous les autres membres du jury : Florence Forbes, Jean-Michel Marin, Christian Robert, Stéphane Robin et Sophie Schbath. C'est un grand honneur pour moi que vous soyez tous là aujourd'hui. Vous avez toute mon estime et toute ma considération. Merci à vous !

Je remercie ensuite chaleureusement les personnes qui ont contribué de près ou de loin à ce que je prenne la décision d'écrire mon manuscrit en vue d'obtenir le diplôme d'habilitation à diriger les recherches. Merci aussi à celles qui m'ont encouragée et soutenue lors de la rédaction proprement dite et de toutes les étapes qui ont suivi jusqu'à ces derniers jours. Je leur en suis extrêmement reconnaissante. Leurs manisfestations, qu'il s'agisse d'un petit sourire en passant devant mon bureau, d'un petit mot, d'un message ou d'un échange plus profond, furent infiniment précieuses pour moi.

J'adresse également mes remerciements à tous mes collaborateurs avec lesquels j'ai pris beaucoup de plaisir à travailler au cours de ces années d'activités de recherche. Sans eux, ce manuscrit ne serait pas tel qu'il est. Merci aussi à tous les collègues avec lesquels j'ai eu l'occasion de partager des connaissances plus ou moins profondes, d'appréhender des interrogations plus ou moins sérieuses, d'échanger des opinions plus ou moins marquées, de confronter des points de vue plus ou moins déterminés. Ces moments de discussion, et l'émulation qui en découle, contribuent considérablement à mes yeux à l'activité de recherche.

Enfin, mes derniers remerciements vont aux membres de tous les laboratoires dans lesquels j'ai exercé avec plaisir et enthousiasme mes activités de recherche et d'enseignement depuis quinze ans : mes débuts en thèse au laboratoire de mathématiques d'Orsay, puis le CEREMADE de l'Université Paris Dauphine et le laboratoire Statistique et Génome de l'Université Evry Val d'Essonne comme ATER, le laboratoire d'analyse, de géométrie et de leurs applications de l'Université Paris Nord en tant que maître de conférence et actuellement l'unité MaIAGE du centre INRA de Jouy-En-Josas comme chargée de recherche. Dans tous ces laboratoires, j'ai trouvé un environnement harmonieux et apaisé, propice au travail. Je souhaite vivement pouvoir poursuivre mon activité dans des conditions aussi agréables.

Contents

Introduction en français

Parcours scientifique

Mon stage de DEA effectué sous la direction de Marc Lavielle fut ma première expérience de recherche en statistique. A cette occasion, nous avons *collaboré étroitement* avec un géophysicien pour résoudre un *problème d'estimation* en tomographie dans un contexte à *données manquantes* en *implémentant un algorithme numérique.* Ce projet initial rassemble à lui seul les composantes principales de mes activités de recherche depuis lors.

J'ai ensuite effectué une thèse en statistique à l'Université Paris Sud sous la direction de Marc Lavielle. J'ai proposé, étudié et mis en oeuvre des algorithmes pour l'estimation par maximum de vraisemblance dans des modèles à variables latentes non linéaires. J'ai effectué des applications en pharmacologie et en traitement du signal en collaboration avec des scientifiques de ces deux domaines. Après mon doctorat, j'ai exercé en tant que maître de conférence à l'Université Paris Nord au Laboratoire d'Analyse, de Géométrie et de leurs applications pendant quatre ans. J'y ai poursuivi mes activités de recherche en statistique et ai développé de nouvelles collaborations motivées par des applications en analyse d'image et en nutrition. Depuis 2009, je suis chargée de recherche à l'INRA au département Mathématiques et Informatique Appliquées (MIA). J'exerce ma fonction au sein de l'équipe DYNENVIE de l'unité MIA de Jouy-en-Josas. Je m'intéresse à de nouvelles applications telles que l'étude de dynamique de population en épidémiologie ou en agronomie, ou encore l'étude de la croissance des plantes en collaboration avec des scientifques de l'INRA et d'autres organismes. J'oriente et développe ma recherche théorique en statistique pour apporter des réponses pratiques aux questions soulevées par ces applications.

Contexte scientifique

Les problématiques statistiques auxquelles je m'intéresse découlent majoritairement de l'analyse de phénomènes complexes dans lesquels interviennent différentes quantités : certaines sont observables, plus ou moins directement, d'autres pas. Un exemple classique est celui de la déconvolution de signal où l'on observe le signal d'intérêt bruité. L'objectif est d'obtenir à partir des seules quantités observées des informations sur les quantités d'intérêt non observées ou partiellement observées. Les outils probabilistes habituellement utilisés dans ce cas sont les modèles à variables latentes. Les quantités d'intérêt non accessibles sont modélisées par des variables aléatoires non observées aussi appelées latentes. Les quantités accessibles sont modélisées par des variables aléatoires pour lesquelles on observe une réalisation. La dépendance entre les observations et les variables latentes est modélisée par la distribution jointe. Un des enjeux principaux pour le statisticien est de caractériser cette distribution à partir des seules observations. Dans la suite, on supposera que cette distribution jointe est paramétrique. L'objectif est alors de fournir un estimateur des paramètres du modèle, ainsi que de la variance de cet estimateur (par exemple pour construire des intervalles de confiance). L'estimateur du maximum de vraisemblance (EMV) est un de ceux considérés classiquement dans ce contexte. Il est obtenu comme solution du problème d'optimisation en le paramètre de la vraisemblance observée, c'est-à-dire la vraisemblance marginale définie comme l'intégrale de la vraisemblance jointe sur les variables latentes. Dans des modèles complexes, cette intégrale n'admet généralement pas de forme analytique et le problème d'optimisation ne peut être résolu par un calcul direct. Dans ce cas, un algorithme d'optimisation peut permettre d'obtenir une approximation numérique de l'EMV.

Un des algorithmes les plus répandus est l'algorithme Expectation Maximization (EM) proposé par Dempster et al. [1977]. Il s'agit d'un algorithme déterministe itératif, chaque itération comportant deux étapes. La première consiste à évaluer l'espérance de la log-vraisemblance complète conditionnellement aux observations et à la valeur courante du paramètre (étape E) ; dans la seconde, la valeur du paramètre est mise à jour en maximisant cette quantité (étape M). Pour des modèles exponentiels, cet algorithme converge vers un maximum local de la vraisemblance observée sous des hypothèses générales de régularité du modèle. Cependant, dans de nombreux modèles, l'espérance conditionnelle de la log-vraisemblance complète n'admet pas d'expression analytique et l'algorithme EM ne peut être implémenté. Des algorithmes alternatifs ont été proposés, soit en approximant la vraisemblance, soit en faisant intervenir une étape de simulation des variables latentes. La plupart de ces algorithmes ne possèdent pas de propriétés théoriques de convergence, ou requièrent des hypothèses de convergence peu réalistes, ou encore nécessitent de très longs temps de calcul.

L'algorithme Stochastic Approximation Expectation Maximization (SAEM) proposé par Delyon et al. [1999] est une version stochastique de l'algorithme EM qui possède d'intéressantes propriétés combinant celles de l'approximation stochastique et celles de l'algorithme EM. L'étape E de l'algorithme EM est remplacée par deux étapes. Dans la première, une réalisation des variables latentes est simulée selon la loi conditionnelle, dans la seconde, cette réalisation est utilisée pour calculer une quantité auxiliaire approximant l'espérance conditionnelle de la log-vraisemblance complète par un schéma d'approximation stochastique. La convergence presque sûre de cet algorithme vers un maximum local de la vraisemblance observée a été établie sous des hypothèses générales de régularité du modèle (cf. Delyon et al. [1999]). L'algorithme SAEM est facile à implémenter et nécessite de faibles temps de calcul. Il peut être appliqué à des modèles complexes sous réserve de savoir simuler des réalisations des variables latentes selon la loi conditionnelle aux observations. Cette condition restreint drastiquement le champ des applications possibles. De façon plus générale, l'étape de simulation des variables latentes requiert une attention spécifique. En effet, des comportements numériques atypiques peuvent apparaître lorsque les variables latentes sont de grande dimension, car leur distribution est souvent multimodale dans ce contexte.

Mes thèmes de recherche principaux concernent le développement, l'analyse et l'implémentation de nouveaux algorithmes stochastiques dérivant de l'algorithme déterministe EM et permettant d'obtenir une approximation numérique d'un estimateur des paramètres pour des modèles probabilistes complexes à variables latentes. Les applications font partie intégrante de mes travaux de recherche, les nouvelles applications posant de nouvelles problématiques, à la fois de modélisation et computationnelles, et peuvent elles mêmes conduire à de nouveaux développements mathématiques et algorithmiques. Cette symbiose entre recherche statistique et applications est un élément clé de la dynamique de mes activités de recherche et passe par des collaborations étroites avec des scientifiques d'autres domaines.

Organisation du manuscrit

L'essentiel de mes travaux de recherche est présenté dans ce manuscrit. J'ai choisi de les regrouper en trois parties comme suit.

Dans la première partie, après une brève description du problème de l'estimation dans des modèles à variables latentes et des algorithmes d'optimisation existants, je présente des développements autour de l'algorithme SAEM et leurs analyses. Au cours de ma thèse dirigée par Marc Lavielle, nous avons proposé d'introduire dans l'étape de simulation de l'algorithme SAEM une méthode de Monte Carlo par chaines de Markov (MCMC) permettant ainsi d'appliquer cet algorithme sans avoir besoin de simuler des réalisations des variables latentes selon la loi conditionnelle [A1]. Cet algorithme, noté SAEM-MCMC dans la suite, conserve toutes les propriétés intéressantes de l'algorithme SAEM initial tout en étant facile à mettre en oeuvre. Nous avons démontré sa convergence presque sûre vers un point critique de la vraisemblance observée sous des hypothèses générales de régularité du modèle et sous une hypothèse forte pour la loi des variables latentes, à savoir être à support compact. Par la suite, en collaboration avec Alain Trouvé et Stéphanie Allassonnière, nous avons relâché cette hypothèse en introduisant une étape supplémentaire de troncature à chaque itération de l'algorithme [A3]. Nous avons obtenu le même résultat de convergence pour cet algorithme avec troncature sous des hypothèses générales de régularité seulement. Par ailleurs, l'algorithme proposé est facile à implémenter et rapide. Nous obtenons également un estimateur de la matrice d'information de Fisher observée [A1].

Dans le cas de variables latentes de grande dimension, la performance des méthodes MCMC classiques décroit rapidement lorsque cette dimension augmente. Motivée par des applications en analyse d'images, je me suis intéressée aux techniques de simulation de variables aléatoires de grande dimension. En collaboration avec Stéphanie Allassonnière, nous avons proposé une version anisotrope de l'algorithme Metropolis Adjusted Langevin (AMALA). Nous avons démontré son ergodicité uniforme. Nous l'avons utilisé comme échantillonneur dans l'algorithme SAEM-MCMC et avons prouvé la convergence presque sûre de cet algorithme vers un point critique de la vraisemblance observée, ainsi qu'un théorème de la limite centrale [A12]. En collaboration avec Gersende Fort, Benjamin Jourdain, Tony Lelièvre et Gabriel Stolz, nous avons étudié l'algorithme de Wang Landau, bien adapté pour simuler des variables en grande dimension. Nous avons démontré son ergodicité uniforme ainsi qu'un théorème de la limite centrale [A11]. Nous avons également analysé son comportement via une étude de simulations [A8].

Dans la deuxième partie, je présente mes travaux relatifs à la modélisation et à l'estimation paramétrique dans des modèles à effets mixtes, en particulier dans les modèles déformables en analyse d'image, ainsi que mes travaux liés à des problématiques de tests en régression gaussienne. En collaboration avec Marc Lavielle, nous avons proposé d'utiliser l'algorithme SAEM-MCMC pour obtenir une approximation numérique de l'EMV dans des modèles à effets mixtes. Ces modèles sont fréquemment utilisés pour l'analyse de mesures longitudinales répétées. Lorsque le modèle est linéaire en les effets aléatoires et que l'erreur résiduelle est gaussienne, la vraisemblance est explicite et l'estimateur du maximum de vraisemblance peut être calculé par une optimisation directe. En revanche, hors de ce contexte spécifique, la vraisemblance n'admet pas forcément de forme analytique, rendant le calcul de l'EMV difficile. Il faut alors recourir à des procédures numériques plus complexes, souvent longues en temps de calcul, et pour lesquelles il n'existe pas toujours de propriétés théoriques garantissant leur convergence. L'algorithme SAEM-MCMC est une solution efficace, rapide et convergente, pour calculer l'EMV dans des modèles à effets mixtes, en particulier non linéaires. Nous l'avons appliqué pour des modèles de courbes de croissance et de pharmacocinétique [A2]. En collaboration avec Alain Trouvé et Stéphanie Allassonnière, nous avons appliqué l'algorithme SAEM-MCMC avec l'étape de troncature à un modèle déformable d'images dans un cadre bayésien, motivé par la petite taille des échantillons disponibles en imagerie médicale [A3]. Un modèle déformable permet de représenter un échantillon d'images d'un objet d'intérêt par une image de référence et des déformations géométriques, de sorte que chaque image de l'échantillon est obtenue comme résultat d'une déformation géométrique de l'image de référence, à une petite erreur près. Les déformations géométriques sont les variables latentes du modèle. Nous avons considéré dans un premier temps des petites déformations linéaires et avons utilisé un échantillonneur de type Gibbs hybride comme méthode MCMC pour des variables multivariées de dimension raisonnable. Cette application est à la source des développements liés aux variables latentes de grande dimension présentés dans la première partie, l'algorithme de Gibbs hybride devenant très gourmand en temps de calcul lorsque la dimension des variables latentes augmente. En collaboration avec Stéphanie Allassonnière, nous avons considéré une extension de ce modèle à un modèle de mélanges. Dans de nombreuses applications, il est nécessaire de contraindre le type de déformations géométriques considérées. Par exemple, les déformations difféomorphes n'autorisent pas les recouvrements dans l'image et empêchent des changements de topologies. Un modèle de mélanges permet de modéliser l'échantillon d'images comme étant issu de plusieurs populations ayant des images de référence topologiquement différentes (par exemple avec ou sans "boucle" pour le chiffre manuscrit 2). Dans le cas de mélange de modèles déformables, la mise en oeuvre de l'algorithme SAEM-MCMC donnent lieu à des comportements numériques instables liés aux états absorbants. Nous avons proposé un algorithme stochastique d'estimation spécifique que nous avons mis en oeuvre et pour lequel nous avons établi la convergence vers un point critique de la vraisemblance [A5]. Par la suite, en collaboration avec Stéphanie Allassonnière et Stanley Durrlemann, nous avons appliqué l'algorithme SAEM-MCMC avec l'échantillonneur AMALA au modèle déformable à grandes déformations difféomorphes [A13]. Le nouvel algorithme révèle tout son potentiel dans ce contexte de très grande dimension. Nous avons également développé une extension du modèle permettant d'optimiser la position des points de contrôle de la déformation, simultanément aux autres paramètres du modèle. Nous avons aussi proposé un critère empirique pour sélectionner un nombre optimal de points de contrôle, dans le but d'optimiser la dimension du modèle. Les applications, réalisées en image sur la base US POSTAL et sur des données d'imagerie médicale 2D et 3D, sont regroupées à la fin de cette partie.

En collaboration avec France Mentré et Adeline Samson, nous avons comparé la pertinence de traitements effectués sur deux groupes de patients. Dans ce cadre, j'ai appliqué l'algorithme SAEM-MCMC à des modèles de pharmacologie. Nous avons appliqué l'algorithme SAEM-MCMC pour effectuer l'estimation de paramètres sous l'hypothèse nulle et sous l'alternative et avons ensuite appliqué un test du rapport de vraisemblance. En collaboration avec Mohamed Sahmoudi, Karim Abed-Meraim, Philippe Ciblat et Marc Lavielle, nous avons appliqué l'algorithme SAEM-MCMC pour estimer les paramètres d'un modèle de séparation de sources pour des signaux à queue lourde. Ces travaux détaillés au chapitre 2 de ma thèse de doctorat [T] et dans le proceeding [P2] ne sont pas présentés dans ce manuscrit.

Motivée au départ par la question du choix de modèle pour les modèles à effets mixtes, j'ai proposé en collaboration avec Sylvie Huet un test d'adéquation de modèles pour tester une hypothèse de type linéaire sur l'espérance d'un vecteur gaussien à erreurs bloc corrélées à structure de covariance connue aux paramètres près [A10]. Nous avons démontré que notre procédure de test est asymptotiquement consistante et puissante sur une large classe d'alternatives. Nous avons également proposé une version bootstrap de notre test et montré sa consistance. Nous avons évalué via une étude de simulation les propriétés non asymptotiques de nos procédures sur des échantillons de taille finie et les avons appliquées à un jeu de données de couverture forestière de Galicie.

J'ai également effectué un travail d'appui statistique pour la mise en oeuvre d'une procédure de test multiple en régression gaussienne, réalisé en collaboration avec Renaud Rincent et Alain Charcosset, généticiens à l'INRA, pour étudier la puissance de détection de Quantitative Trait Loci le long du génome dans un modèle à effets mixtes linéaire faisant intervenir la matrice d'apparentement. Nous avons étudié les performances de la procédure en fonction de l'estimateur considéré pour cette matrice [A9]. Ce travail n'est pas présenté dans ce manuscrit.

Dans la troisième partie de ce document, j'ai regroupé mes travaux relatifs à la modélisation et l'estimation en analyse de survie. En collaboration avec Charles El-Nouty, j'ai étudié les modèles de fragilité proposés par Vaupel et al. [1979] qui sont une extension du modèle de Cox et permettent de prendre en compte l'hétérogénéité présente dans les données de survie en introduisant des variables latentes. Notre intérêt pour ces modèles a été motivé par de nombreux échanges avec des chercheurs de l'Unité de Recherche en Epidémiologie Nutritionnelle (UREN) de l'Université Paris 13 qui ont fréquemment besoin d'analyser ce type de jeux de données. Très riches du point de vue de la modélisation, ces modèles sont relativement complexes du point de vue mathématique et l'estimation des paramètres y est souvent difficile. Dans de nombreux cas, les algorithmes existants n'apportent pas de solutions satisfaisantes. Nous avons proposé d'appliquer aux modèles de fragilité l'algorithme SAEM-MCMC pour obtenir une approximation numérique de l'estimateur du maximum de vraisemblance [A6]. Nous avons démontré que sous des hypothèses générales de régularité vérifiées par les modèles de fragilité classiques, l'algorithme SAEM-MCMC était presque sûrement convergent vers un point critique de la vraisemblance observée. Nous avons comparé les performances de cet algorithme à celles d'autres algorithmes sur des données simulées et réelles. Les résultats numériques montrent un net avantage à l'utilisation de l'algorithme SAEM-MCMC, tant du point de vue de la vitesse d'exécution que de celui de la précision.

En collaboration avec Luc Duchateau et Klaartje Goethals, chercheurs en biométrie à la faculté vétérinaire de l'Université de Ghent, nous avons analysé un jeu de données d'épidémie de mastitis en mettant en oeuvre des modèles de fragilité avec des variables de fragilité multivariées possédant différentes structures de corrélation [R1]. Pouvoir considérer ces modèles complexes permet d'étudier le risque de propagation de la maladie en fonction de la position locale de l'infection. L'estimation des paramètres qui était jusqu'alors impossible en un temps raisonnable dans de tels modèles est réalisée par l'algorithme

SAEM-MCMC. Nous comparons quatre modèles emboités en utilisant des tests du rapport de vraisemblance. Nous justifions a posteriori la validité à distance finie de ce test par une étude de simulation.

Dans le cadre d'une collaboration avec Catherine Picon-Cochard, biologiste à l'INRA, j'ai effectué un travail d'appui statistique sur l'utilisation du modèle de Cox pour l'analyse de données morphométriques de racines en fonction de covariables environnementales [A7]. Ce travail n'est pas présenté dans ce manuscrit.

Pour finir, mes conclusions et perspectives de recherche sont rassemblées dans la dernière partie du manuscrit.

Introduction

Scientific path

My first experience in statistical research took place during my master's internship supervised by Marc Lavielle (INRIA Saclay Ile de France). We addressed, in *close collaboration* with a geophysicist an *estimation problem* in tomography in a *missing data* setting by *implementing a numerical algorithm.* This project covered the principal components of my future research activities.

Then, I did a Phd in statistics at University Paris Sud supervised by Marc Lavielle. I have proposed, studied and implemented algorithms for maximum likelihood estimation in non linear latent variables models. I have applied them in pharmacology and signal processing in collaboration with scientists of these fields. After my Phd, I worked as assistant professor at University Paris Nord at the "Laboratoire d'Analyse, de Géométrie et de leurs applications" during four years. I continued my research activities in statistics and developed new collaborations motivated by applications in image processing and nutrition. Since 2009, I am researcher at INRA in the departement "Mathématiques et Informatique Appliquées" (MIA) in the team DYNENVIE of the unit MIA in Jouy-en-Josas. I am interested in new applications as the study of population dynamics in epidemiology or agronomy, and as the study of plant growth in collaboration with scientists of INRA and of others instituts. I position and develop my theoretical research activities with the objective of providing practical answers to the questions raised by these applications.

Scientific context

I am interested in statistical problems raised by the analysis of concrete complex phenomena in which several quantities are involved: some are observed, more or less directly; others are not. A classical example is the deconvolution problem where the signal of interest is observed up to a noise term. The objective is to extract information on the non-observed or partially observed quantity of interest from the observations. The usual probabilistic tools used in such cases are latent variable models. The quantities of interest that are inaccessible are modeled with unobserved random variables, also referred to as latent or missing variables. The observations are modeled with random variables that are observed. The dependence between observations and latent variables is specified by the joint distribution. One of the main goals for the statistician is therefore to characterize this distribution on the basis of observations alone. In this manuscript, I will assume that this joint distribution is parametric. The objective is then to provide an estimate for the model parameter, as well as for the variance of this estimate (for example, to build confidence intervals). The maximum likelihood estimate (MLE) is one of the estimates usually considered in this context. It is obtained by solving the optimization in the parameter of the observed likelihood, i.e., the marginal likelihood defined as the integral of the joint likelihood over the latent variables. In complex models, this integral admits in general no analytical expression and the optimization cannot be directly solved. In such cases, a numerical value of the MLE may be obtained by using an optimization algorithm.

One of the most common algorithms is the Expectation Maximization (EM) algorithm proposed by Dempster et al. [1977]. It is a deterministic iterative algorithm, where each iteration is composed of two steps. The first consists in evaluating the conditional expectation of the complete log-likelihood given the current parameter estimate value (step E); in the second one, the parameter value is updated by maximizing this quantity (step M). For models belonging to the curved exponential family, this algorithm converges toward a local maximum of the observed likelihood under general regularity assumptions on the model. However, in numerous models, the conditional expectation of the complete log-likelihood has no analytical expression and the EM algorithm cannot be implemented. Alternative algorithms have been proposed, either based on approximations of the likelihood or by introducing a simulation step of the latent variable. Most of them have no established theoretical convergence property, require unrealistic assumptions for theoretical convergence, or require very long computation times.

The Stochastic Approximation Expectation Maximization (SAEM) algorithm proposed by Delyon et al. [1999] is a stochastic version of the EM algorithm that has interesting properties, combining those of stochastic approximation procedures and those of the EM algorithm. Step E of the EM algorithm is divided into two steps. In the first one, a realization of the latent variable is drawn from the conditional distribution; in the second one, this realization serves to calculate an auxiliary quantity that approximates the conditional expectation of the complete log-likelihood through a stochastic approximation procedure. The almost sure convergence of this algorithm toward a local maximum of the observed likelihood is established under general regularity assumptions on the model (cf. Delyon et al. [1999]). It is easy to implement and requires only short computation times. It can be implemented in complex models as soon as it is possible to draw a realization of the latent variable from the conditional distribution. This condition drastically limits the application possibilities. More generally, the simulation step of the latent variable requires close attention since atypical numerical behavior can result from its implementation, in particular, for multimodal latent variable distributions or high dimensional ones.

My main research subjects deal with the development, the analysis and the implementation of stochastic algorithms derived from the deterministic EM algorithm that makes it possible to calculate a numerical value of a parameter estimate for complex latent variable models. Applications are an integral part of my research activities. New problematics in modeling and in computation arise from applications and raise new theoretical and algorithmic developments. This symbiosis between statistical research and applications is a key element of the dynamics of my research activities and is based on close collaboration with scientists in other fields.

Manuscript organization

In this manuscript, I describe my main contributions, broken down into three parts as follows.

In the first part, after a short description of the estimation problem in general latent variable models and the existing algorithms used to address it, I present developments and analysis of the SAEM algorithm. During my PhD which was supervised by Marc Lavielle, we proposed the introduction of a Monte

Carlo Markov Chain (MCMC) method into the simulation step of the SAEM algorithm. This makes it possible to implement the algorithm without being able to draw a realization of the latent variable from the conditional distribution [A1]. This algorithm, designated as SAEM-MCMC below, maintains the good theoretical and practical convergence properties of the SAEM algorithm and, at the same time, extends its application possibilities. We first proved the almost sure convergence of the generated estimate sequence toward the maximum likelihood estimate under usual regularity assumptions on the model and a strong assumption on the latent variable support, i.e., having a compact support. Later, in collaboration with Alain Trouvé (ENS Cachan, CMLA) and Stéphanie Allassionnière (Ecole Polytechnique, CMAP), we relaxed this strong assumption by adding a truncation step to each iteration of the algorithm [A3]. We obtained the same convergence result under general regularity assumptions alone. Moreover, the resulting algorithm is easy to implement and very fast. We also obtained an estimate of the observed Fisher information matrix.

However, considering latent variables of high dimension, the efficiency of classical MCMC methods quickly decreases as the dimension increases. Motivated by complex application settings in image analysis, I focused on sampling methods adapted to this context. In collaboration with Stéphanie Allassionnière, we proposed a new Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA). We proved its uniform ergodicity. We also used it as a sampler in the SAEM-MCMC and proved the almost sure convergence of the generated estimate sequence toward the maximum likelihood estimate under usual regularity assumptions, as well as a central limit theorem [A12]. In addition, in collaboration with Gersende Fort (CNRS, TELECOM ParisTechTelecom, LTCI), Benjamin Jourdain, Tony Lelièvre and Gabriel Stolz, (Ecole des Ponts ParisTech, CERMICS), we studied another sampling algorithm that was well adapted to the high dimension setting, known as the Wang Landau algorithm. We proved its uniform ergodicity as well as a central limit theorem [A11]. We also studied its behavior by performing numerical simulations [A8].

The second part covers my contributions concerning with modeling and parametric estimations in mixed effects models, in particular, in the deformable template model in image analysis, and with the testing problematic in Gaussian regression models.

In collaboration with Marc Lavielle, we proposed using the SAEM-MCMC algorithm to obtain a numerical value of the MLE in mixed effects models. These models are particularly used to analyze repeated longitudinal data. When the model is linear in the random effects and the residual random error Gaussian, the likelihood has an explicit analytical expression and the maximum likelihood estimate can be calculated through some direct optimization. On the other hand, outside this specified context, the likelihood usually does not admit an explicit analytical form, making it difficult to evaluate the MLE. It is then necessary to appeal to more complex numerical procedures, often time-consuming, and not always provided with theoretical convergence properties. The SAEM-MCMC algorithm is an efficient solution, fast and convergent, to evaluate the MLE in mixed effects models, in particular, in non-linear ones. We used it for parameter estimation in growth curve models and in pharmacodynamic models [A2]. In collaboration with Alain Trouvé and Stéphanie Allassonnière,we implemented the SAEM-MCMC algorithm with the additional truncation step to estimate parameters in a deformable template model in a Bayesian setting useful in medical imaging where samples are usually small [A3]. Deformable template models allow us to represent a sample of images with a reference image called template and geometrical deformations, so that each image of the sample is obtained as the result of the geometrical deformation of the template, up to a small error term. The geometrical deformations are the latent variables of such a model. We first considered the case of small linear deformations and used a hybrid Gibbs sampler as the MCMC method. This application motivated the development of high dimensional latent variables presented in the first part of this manuscript, since the hybrid Gibbs sampler becomes very time-consuming as the dimension of the latent variable increases. Later, in collaboration with Stéphanie Allassonnière,we proposed a specific algorithm for parameter estimation in the multicomponent model which is a mixture of deformable template models motivated by a crucial modeling issue [A5]. In numerous applications, it is necessary to constrain the type of geometrical deformations considered, in particular, so that the diffeomorphic deformations do not allow overlapping and prevent topological changes from occuring between the template and the observation, thus creating the need of a mixture model. Since the SAEM-MCMC algorithm is sensitive to numerical phenomena such as trapping states in such a case, we proposed a specific stochastic estimation algorithm and implemented it. We also established its convergence property toward a local maximum of the observed likelihood. Finally, in collaboration with Stéphanie Allassonnière and Stanley Durrlemann, we applied the SAEM-MCMC algorithm provided with the AMALA sampler to the very complex deformable template model using large diffeomorphic deformations [A13]. The new algorithm takes on its full meaning in this high dimensional setting. We also developed an extension that allowed us to optimize the position of the control points of the deformation, simultaneously with the estimation of the other model parameters. We proposed an empirical criterion to select an optimal number of control points as well, leading to the optimization of the model dimension. All the applications related to deformable template models and the different estimation algorithms were performed on the US POSTAL handwritten digit database and 2D and 3D medical image databases. The corresponding experimental results are presented in Section 5.5.

I also implemented the SAEM-MCMC algorithm to carry out the parameter estimation in pharmacological models and signal processing. In collaboration with France Mentré and Adeline Samson,we compared the efficiency of treatments given to two patient groups. We used the SAEM-MCMC algorithm to carry out the parameter estimation under the null hypothesis and under the alternative, and subsequently applied the likelihood ratio test. In collaboration with Mohamed Sahmoudi, Karim Abed-Meraim, Philippe Ciblat and Marc Lavielle,we applied the SAEM-MCMC algorithm to carry out the parameter estimation in blind source separation for heavy tailed signals. These contributions are detailed in Chapter 2 of my PhD manuscript [T] and in the proceedings [P2], respectively, and are not presented in this manuscript.

Motivated at the beginning by the model choice issue for mixed effects models, we finally proposed, in collaboration with Sylvie Huet a goodness-of-fit test for testing a linear hypothesis on the expectation of a Gaussian vector with block correlated errors with a known covariance structure up to some parameters [A10]. We established that our test procedure was asymptotically of the nominal level and consistent over a large class of alternatives. We also proposed a bootstrap version of our procedure. Using a simulation study, we evaluated the finite sample size properties of our procedures and applied them to a forest cover dataset of Galicia.

I also did a statistical support study on a multiple test procedure in Gaussian regression, in collaboration with Renaud Rincent and Alain Charcosset, geneticians at INRA, to study the power of detection of Quantitative Trait Loci along the genome in a non linear mixed effects model involving the kinship matrix. We have studied its performances according to the estimate used for the kinship matrix [A9]. This work is not presented in this manuscript.

In the third part, I included my contributions dealing with modeling and estimation in survival analysis. In collaboration with Charles El-Nouty, we considered the frailty models introduced by Vaupel et al. [1979] which are an extension of the Cox model that takes the heterogeneity that exists in survival data into account by introducing latent variables. We were interested in these models since we have been in contact with researchers at the Unité de Recherche en Epidémiologie Nutritionnelle (UREN) at the University of Paris 13, who often need to analyze this type of dataset. Very rich from a modeling point of view, these models are very complex from a mathematical point of view, and the estimation task is often very difficult. In numerous cases, the existing algorithms do not provide satisfactory solutions. We proposed the application of the SAEM-MCMC algorithm to frailty models to evaluate the MLE [A6]. We proved that under general regularity assumptions fulfilled by classical frailty models, the SAEM-MCMC algorithm is almost surely convergent toward a local maximum of the observed likelihood. We compared the performances of this algorithm with others that exist in the literature on simulated data and on a real set of bladder cancer data. The numerical results showed a net advantage when using the SAEM-MCMC algorithm, both in terms of the accuracy of the limit as well as the computation time.

In collaboration with Luc Duchateau and Klaartje Goethals (Faculty of Veterinary Medicine, Ghent University), we have analyzed a mastitis epidemic dataset using frailty models with a frailty vector of size four with different covariance structures [R1]. Assessing the correlation structure in cow udder quarter infection times allows us to analyze the propagation risk of the disease as a function of the position of the infection. The parameter estimation that could not be performed in a reasonable time in such models until now, is done using the SAEM-MCMC algorithm. We compared four nested models using likelihood ratio tests. Using simulation studies, we justified *a posteriori* the use of likelihood ratio tests for finite sample size.

I also did a statistical support study in collaboration with Catherine Picon-Cochard on the application of the Cox model for analyzing morphometric plant root data as a function of environmental covariables [A7]. This work is not presented in this manuscript.

Finally, my conclusions and perspectives can be found in the last part of this manuscript.

Part I

Extending the Stochastic Approximation Expectation Maximization Algorithm

The first part is composed of my research dealing with the development, the study and the implementation of a new stochastic version of the Expectation Maximization (EM) algorithm. Related studies include [A1,A3,A8,A11,P1,P2,T,A12].

In the first part, after a short description of the estimation problem in general latent variable models and the existing algorithms used to address it, I present developments and analysis of the SAEM algorithm. During my PhD which was supervised by Marc Lavielle, we proposed the introduction of a Monte Carlo Markov Chain (MCMC) method into the simulation step of the SAEM algorithm. This makes it possible to implement the algorithm without being able to draw a realization of the latent variable from the conditional distribution [A1]. This algorithm, designated as SAEM-MCMC below, maintains the good theoretical and practical convergence properties of the SAEM algorithm and, at the same time, extends its application possibilities. We first proved the almost sure convergence of the generated estimate sequence toward the maximum likelihood estimate under usual regularity assumptions on the model and a strong assumption on the latent variable support, i.e., having a compact support. Later, in collaboration with Alain Trouvé (ENS Cachan, CMLA) and Stéphanie Allassonnière (Ecole Polytechnique, CMAP), we relaxed this strong assumption by adding a truncation step to each iteration of the algorithm [A3]. We obtained the same convergence result under general regularity assumptions alone. Moreover, the resulting algorithm is easy to implement and very fast. We also obtained an estimate of the observed Fisher information matrix.

However, considering latent variables of high dimension, the efficiency of classical MCMC methods quickly decreases as the dimension increases. Motivated by complex application settings in image analysis, I focused on sampling methods adapted to this context. In collaboration with Stéphanie Allassonnière, we proposed a new Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA). We proved its uniform ergodicity. We also used it as a sampler in the SAEM-MCMC and proved the almost sure convergence of the generated estimate sequence toward the maximum likelihood estimate under usual regularity assumptions, as well as a central limit theorem [A12]. In addition, in collaboration with Gersende Fort (CNRS, TELECOM ParisTechTelecom, LTCI), Benjamin Jourdain, Tony Lelièvre and Gabriel Stolz, (Ecole des Ponts ParisTech, CERMICS), we studied another sampling algorithm that was well adapted to the high dimension setting, known as the Wang Landau algorithm. We proved its uniform ergodicity as well as a central limit theorem [A11]. We also studied its behavior by performing numerical simulations [A8].

1. Estimation in the Latent Variable Models

1.1. Latent variable models

We consider a latent variable model defined as follows: the observed variable is denoted by y and the latent variable (also referred to as the missing or hidden variable) by z. The observed variables are related to the latent variables. This link is specified through the joint distribution of the complete variable defined by (y, z). We assume that this joint distribution has a density against a given σ-finite Borelian measure μ which belongs to a parametric family denoted by $\{f(y, z; \theta), \theta \in \Theta\}$, where the parameter θ takes its value in a subset of $\mathbb{R}^p$ denoted by Θ. This joint distribution defines the parametric probabilistic latent variable model. The most popular ones are hidden Markov models, hierarchical models and mixed effects models. Such models are widely used in many application fields such as tomography, signal processing, pharmacology, genetics, economics and image analysis.

Given the probabilistic latent variable model, the statistician is interested in estimating the parameter θ from the observed data y, the latent variable z being unobserved. In a frequentist estimation approach, a very popular estimator with good asymptotical properties in many classical models is the Maximum Likelihood Estimator (MLE) of the parameter θ, namely the value $\hat{\theta}$ of θ that maximizes the observed likelihood g defined by:

$$g(y; \theta) = \int f(y, z; \theta)\mu(dz). \tag{1.1}$$

We can also consider a Bayesian estimation approach, for example if only few observations are available or if a priori information is available and has to be introduced into the probabilistic model. The parameter θ is considered as a random variable that follows a given prior distribution denoted by q. A useful estimator would then be the Maximum A Posteriori (MAP) estimator, namely the value $\tilde{\theta}$ of θ that maximizes the posterior distribution denoted by p and defined by:

$$p(\theta|y) \propto g(y; \theta)q(\theta).$$

In the following, I consider a frequentist estimation approach and focus on the Maximum Likelihood Estimator (MLE). However all the considerations developed in this section can be directly transposed to the Bayesian context by replacing the MLE with the MAP estimator, as will be done in Section 5.

As soon as the joint distribution of the complete variable (y, z) is complex, it is often not possible to directly compute the numerical value of such estimators as the solution of an optimization problem. It is then necessary to use numerical algorithms to evaluate an estimated value of this estimator.

1.2. Numerical algorithms for estimation

A useful tool to compute a numerical value of the Maximum Likelihood Estimator $\hat{\theta}$ is the Expectation Maximization (EM) algorithm introduced by Dempster et al. [1977]. This is an iterative algorithm that generates a sequence of estimated values $(\theta_k)_k$ converging toward a stationary point of the observed likelihood g under some regularity assumptions of the model (see Dempster et al. [1977], Vaida [2005], Wu [1983]). The heuristic of this algorithm is based on the following strategy: the variable z being unobserved, instead of maximizing the observed log-likelihood $\log g(y;\theta)$ in θ, it is instead possible to maximize the expectation of the complete log-likelihood conditional to the observed variable y equal to $\mathbb{E}(\log f(y,z;\theta)|y;\theta)$, which might be a realistic approximation of $\log g(y;\theta)$ denoted by $l(\theta)$ for a given observation y in the sequel. The two steps of iteration k of the algorithm consist in alternatively updating, first, the conditional expectation:

$$Q(\theta|\theta_{k-1}) = \mathbb{E}(\log f(y,z;\theta)|y;\theta_{k-1}) \tag{1.2}$$

using the current parameter value θ_{k-1} and, second, the parameter by maximizing Q according to $\theta_k = \operatorname{argmax}_\theta Q(\theta|\theta_{k-1})$. The initial value θ_0 is chosen arbitrarily. Applying Jensen's inequality shows that the sequence $(\log g(y,\theta_k))_k$ is non-decreasing where the sequence $(\theta_k)_k$ is generated by the algorithm. The structural hypothesis always assumed to study its theoretical convergence is that the complete likelihood belongs to the curved exponential family, meaning that the complete likelihood $f(y,z;\theta)$ can be written as:

$$f(y,z;\theta) = \exp\left[-\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right] ,$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product, the sufficient statistics S is a function on $\mathbb{R}^l$, taking its values in a subset $\mathcal{S}$ of $\mathbb{R}^m$ and ψ, ϕ are two functions on Θ (note that S may also depend on y, but we omit this dependency for the sake of simplicity). This condition is usual within the framework of EM algorithm applications and is fulfilled by a large range of complex models.

Moreover, Wu [1983] exhibits some strong assumptions that ensure the convergence of the sequence $(\theta_k)_k$ toward a local maximum of the observed log-likelihood $\log g$.

Nevertheless, one or both steps of this algorithm may not always be easily carried out in complex latent variable models. Thus, alternative algorithms were developed to allow wider range of applications for the EM algorithm.

When the M-step cannot be done analytically, alternative satisfactory solutions have been proposed, e.g., based on the Newton algorithm (see Lange [1995]) or on conditional maximization (see Meng [1994]). On the contrary, overcoming the difficulty of an untractable E-step leads to alternative solutions that are not as satisfactory. When the computation of the conditional expectation defined in Equation (1.2) is not feasible, two types of alternatives, in particular, have been proposed. The first ones are based on an approximation of the underlying likelihood function, e.g., by its first or second order development. To the best of my knowledge, there is no theoretical convergence property established in that case, probably since it is difficult to control the propagation of the error made by such an approximation. Several authors proposed using a Laplace approximation for the computation of the conditional expectation defined in Equation (1.2). In the case of non-linear mixed effects models, it was proposed by Vonesh [1996] but very restrictive assumptions were required to prove the theoretical convergence of the generated sequence (θ_k). The second ones are based on the simulation of the unobserved variable. Let us mention the stochastic EM algorithm proposed by Celeux and Diebolt [1986]. The authors proved the almost sure convergence in mean of the generated sequence toward a stationary point of the observed likelihood. They extended their approach to mixture models by adding a simulated annealing step and proved the almost sure convergence toward a local maximum likelihood. Another algorithm was proposed by Wei and Tanner [1990] who used a Monte Carlo sum to approximate the conditional expectation defined in Equation (1.2) at each iteration leading to the so-called Monte Carlo EM (MCEM) algorithm. Fort and Moulines [2003] established its almost sure convergence toward a local maximum of the observed likelihood but this algorithm is time consuming since it is necessary to simulate a huge quantity of auxiliary variables at each iteration.

1.3. The Stochastic Approximation Expectation Maximization (SAEM) algorithm

The Stochastic Approximation Expectation Maximization (SAEM) algorithm was proposed by Delyon et al. [1999] and combined the idea of simulating the unobserved variable to the nice convergence property of the stochastic approximation. It is based on the construction of a sequence $(Q_k(\theta))_k$, which will asymptotically produce a good approximation of the conditional expectation defined in Equation (1.2). The E-step of the EM algorithm is divided into two steps. Consider the iteration k. First, a realization z_k of the latent variable is sampled from the conditional distribution $\pi_{\theta_{k-1}}(z|y)$ of z conditional to y using the current value of the parameter estimate θ_{k-1}. Second the quantity Q_k is updated through a stochastic approximation procedure using the realization z_k. Let $(\gamma_k)_k$ be a decreasing positive step size sequence. The algorithm remains to:

Initialization step: Initialize θ_0 in a fixed compact set.

Then, for all $k \geq 1$ the k^{th} iteration consists in three steps :

Simulation step: simulate z_k from the conditional distribution $\pi_{\theta_{k-1}}(\cdot|y)$.
Stochastic approximation step: compute the quantity

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_{k-1}(\log f(y, z_k; \theta) - Q_{k-1}(\theta)), \tag{1.3}$$

Maximization step: update the parameter value according to $\theta_k = \operatorname{argmax}_\theta Q_k(\theta)$.

Delyon et al. [1999] proved the almost sure convergence of the sequence $(\theta_k)_k$ toward a *local* maximum of the observed likelihood by using martingale theory results. This algorithm therefore has very interesting convergence properties and requires only short computation times. Nevertheless, it suffers from several drawbacks. Most of them are common to other algorithms of this type. First, since the limit is a local maximum, it may depend on the initialization point, making it necessary to run the algorithm several times with different initializations. Second, the nature of the limit point has to be investigated to be sure that it is a *global* maximum. Finally, the last drawback, specific to this algorithm, is the requirement of simulating the unobserved variable from the conditional distribution, significantly limiting its application.

2. Using the Monte Carlo Markov Chain method in the SAEM algorithm

During my PhD, which was supervised by Marc Lavielle, we proposed a new version of the classical SAEM algorithm that does not require the simulation of the unobserved variable from the conditional distribution.

I provide the details of the Monte Carlo Markov Chain method below and its insertion into the SAEM algorithm.

2.1. Monte Carlo Markov Chain (MCMC) method

Given some objective distribution π to be reached, a MCMC procedure consists in generating an ergodic Markov chain having it as limiting distribution (see Gilks et al. [1996], Robert [1996]). The most common type of MCMC algorithm is the Metropolis Hastings algorithm. Its transition probability consists in sampling a candidate from a proposal (also referred to as instrumental) distribution q and accepting it with a probability equal to the acceptance ratio defined as:

$$\alpha(z,z') = \min\left\{\frac{\pi(z'|y)q(z',z)}{\pi(z|y)q(z,z')}\ ,\ 1\right\} \tag{2.1}$$

This provides a transition kernel Π of this form: for any Borel set $A \in \mathcal{B}$:

$$\Pi(x,A) = \int_A \alpha(x,z)q(x,z)dz + 1_A(x)\int_{\mathcal{X}}(1-\alpha(x,z))q(x,z)dz\,. \tag{2.2}$$

This procedure always accepts the new value z' when the likelihood ratio $\frac{\pi(z'|y)}{q(z,z')}$ is larger than the previous one.

Another well-known MCMC algorithm for multivariate random variables is the Gibbs sampler (see Gilks et al. [1996], Robert [1996]). Its transition probability consists in updating each coordinate at a time by simulating it from its conditional distribution, conditional to all of the other coordinates. This algorithm is particularly well adapted to the Bayesian context when using conjugate laws, since simulation tasks are carried out directly. Otherwise, it is possible to use a Metropolis Hastings step in one simulation step of the Gibbs sampler, leading to the so-called hybrid Gibbs sampler. However, since the dimension of the variable becomes huge, the computation time drastically increases and numerical problems such as trapping states may appear. This often limits its implementation, as will be seen in Section 5.2.

2.2. The SAEM-MCMC algorithm

Let us assume that we are not able to simulate the missing data z_k at iteration k from the conditional distribution $\pi_{\theta_k}(\cdot|y)$. In such a case, we propose using a MCMC procedure to simulate the value z_k of the latent variable at iteration k. In practice only the simulation step (S-step) of the SAEM algorithm is modified; the other two remain unchanged. In fact, instead of simulating the unobserved variable from the conditional distribution, we simulate it from the transition kernel Π_θ of an ergodic Markov Chain that has this conditional distribution π_θ as stationary distribution. Since we consider only parametric models $\mathcal{P}$ in this manuscript that belong to the curved exponential family, the stochastic approximation can be made either on the complete log-likelihood or on the sufficient statistics S of the model using a positive step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$. Finally, we update the parameter in the M-step.

Thus the algorithm can be summarized as follows:

Initialization step: Initialize θ_0, s_0 and z_0 in fixed compact sets.

Then, for all $k \geq 1$ the k^{th} iteration consists in three steps :

Simulation step: simulate z_k from the transition probability $\Pi_{\theta_{k-1}}(z_{k-1},\cdot)$.
Stochastic approximation step: compute the quantity

$$s_k = s_{k-1} + \gamma_{k-1}[S(z_k) - s_{k-1}], \tag{2.3}$$

Maximization step: update the parameter value according to:

$$\theta_k = \hat{\theta}(s_k) \text{ where } \hat{\theta}(s) = \underset{\theta}{\operatorname{argmax}}\left[-\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right] . \tag{2.4}$$

This algorithm still has interesting theoretical properties and its implementation is fast and easy. Moreover, it can be applied to many complex latent variable models. Assuming that the latent variable z has a compact support, we prove the almost sure convergence of the sequence $(\theta_k)_k$ toward a stationary point of the observed likelihood under some usual regularity assumptions on the model, on the transition kernel of the MCMC method, ensuring its ergodicity, and on the step size sequence (see [A1]).

Theorem 2.1. *Assume that some usual regularity assumptions on the model, the transition kernel and the step size sequence are fulfilled. Assume in addition the assumption* **(C)***: the sequence $(s_k)_{k\geq 0}$ takes its values in a compact subset of $\mathcal{S}$. Then, w.p. 1, $\lim_{k\to+\infty} d(\theta_k, \mathcal{L}) = 0$ where $d(x, A)$ denotes the distance of x to the closed subset A and $\mathcal{L} = \{\theta \in \Theta, \partial_\theta l(y;\theta) = 0\}$ is the set of stationary points of l.*

We apply a result of Benveniste et al. [1990] to control small random Markovian perturbations.

To relax the restrictive assumption of bounded support for the latent variable, we proposed, in collaboration with Stéphanie Allassonnière and Alain Trouvé, to add a truncation step to the algorithm (see [A3]). In fact, to prove the convergence of the sequence generated through the stochastic approximation procedure, it is first necessary to prove its stability i.e. it stays in some given compact set. The assumption of bounded support for the latent variable allowed us to prove this in the algorithm without using the truncation procedure. Introducing a more general technique that involves truncation on random boundaries allows us to prove this without assuming that the support of the latent variable distribution is bounded. Our proof is based on the general stability and convergence results for stochastic algorithms with truncation on random boundaries given in Andrieu et al. [2005]. The main technical point is that in the presence of unbounded latent variables, the usual regularity conditions as a function of the parameters for the solutions of the Poisson equations for the Markovian dynamic cannot be verified and have to be relaxed.

The truncation on random boundaries can be formalized as follows. Let $(\mathcal{K}_q)_{q\geq 0}$ be a sequence of increasing compact subsets of $\mathcal{S}$ so that $\cup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$. Let $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ be a monotone non-increasing sequence of positive numbers and K a compact subset of $\mathbb{R}^N$. We construct a sequence $((z_k, s_k))_{k\geq 0}$ as follows. As long as the stochastic approximation does not wander outside the current compact set and is not too far from its previous value, we run the SAEM-MCMC algorithm. As soon as one of these conditions is not satisfied, we reinitialize the sequences of z and s using a projection, we increase the size of the compact set and continue the iterations until convergence (for more details see Andrieu et al. [2005]). This is detailed in the following steps:

Initialization step: Initialize z_0 and s_0 in two fixed compact sets.

Then, for all $k \geq 1$ the k^{th} iteration consists in four steps :

Simulation step: Draw one new element $\bar{z}$ of the non-homogeneous Markov Chain with respect to the kernel with the current parameters $\Pi_{\theta_{k-1}}$ and starting at z_{k-1}.

$$\bar{z} \sim \Pi_{\theta_{k-1}}(z_{k-1}, \cdot)\,.$$

Stochastic approximation step: . Compute

$$\bar{s} = s_{k-1} + \gamma_{\zeta_{k-1}}(S(\bar{z}) - s_{k-1})\,. \tag{2.5}$$

Truncation step: If $\bar{s}$ is outside the current compact set $\mathcal{K}_{\kappa_{k-1}}$ or too far from the previous value s_k, then restart the stochastic approximation in the initial compact set, extend the truncation boundary to $\mathcal{K}_{\kappa_k}$ and start again with a bounded value of the missing variable. Otherwise, set $(z_k, s_k) = (\bar{z}, \bar{s})$ and keep the truncation boundary to $\mathcal{K}_{\kappa_{k-1}}$. Update the sequence ζ_k and κ_k following Andrieu et al. [2005].

Maximization step: Update the parameters using (2.4).

The index κ denotes the current active truncation set, the index ζ is the current index in the sequences $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ and the index ν denotes the number of iterations since the last projection.

2.3. Estimation of the observed Fisher Information Matrix

An estimation procedure should generate a point estimate $\widehat{\theta}$ together with the covariance of the estimate (e.g., to enable construction of confidence sets for the true parameter value). The asymptotic theory for maximum likelihood estimation, when established, ensures that:

$$\sqrt{n}(\widehat{\theta} - \theta^\star) \rightarrow_{n\rightarrow\infty} \mathcal{N}(0, I(\theta^\star)^{-1})\,, \tag{2.6}$$

where $\theta^\star$ is the MLE and $I(\theta^\star)$ is the observed Fisher Information Matrix. Thus, an estimate of the asymptotic covariance of $\widehat{\theta}$ is the inverse of the observed Fisher information matrix $-\partial_\theta^2 l(\widehat{\theta})$.

Thanks to the maximum likelihood estimator obtained with the SAEM-MCMC algorithm, it is possible to simultaneously obtain an estimation of the Fisher Information Matrix. In [A1] we propose a method to estimate this matrix by using the fact that the gradient (the Fisher score function) and the Hessian (observed Fisher Information) of the log-likelihood l can be almost directly obtained from the simulated missing data z. Using the so-called Fisher identity, the Jacobian of the log-likelihood of the observed data $l(\theta)$ is equal to the conditional expectation of the complete data likelihood:

$$\partial_\theta l(\theta) \triangleq \mathrm{E}[\partial_\theta \log f(y,z;\theta)|y;\theta]$$

where ∂_θ denotes the differential with respect to θ. By analogy with the implementation of the SAEM algorithm, the following approximation scheme is proposed:

$$\boldsymbol{\Delta_k} = \boldsymbol{\Delta_{k-1}} + \gamma_k \left[\partial_\theta \log f(y, \boldsymbol{z^{(k)}}; \theta_k) - \boldsymbol{\Delta_{k-1}}\right].$$

Using Louis' missing information principle (Louis [1982]), the Hessian of l at θ is the observed Fisher Information matrix $\partial_\theta^2 l(\theta)$ that may be expressed as:

$$\partial_\theta^2 l(\theta) = \mathrm{E}_\theta[\partial_\theta^2 \log f(y,z;\theta)] + \mathrm{Cov}_\theta[\partial_\theta \log f(y,z;\theta)].$$

where $\mathrm{Cov}_\theta(\psi(z)) \triangleq \mathrm{E}_\theta[(\psi(z) - \mathrm{E}_\theta(\psi(z)))(\psi(z) - \mathrm{E}_\theta(\psi(z)))^t]$. Using this expression, it is possible to derive the following stochastic approximation procedure to approximate $\partial_\theta^2 l(\theta)$:

$$\begin{aligned}
\boldsymbol{G_k} &= \boldsymbol{G_{k-1}} + \gamma_k \left[\partial_\theta^2 \log f(y, \boldsymbol{z^{(k)}}; \theta_k) + \partial_\theta \log f(y, \boldsymbol{z^{(k)}}; \theta_k) \partial_\theta \log f(y, \boldsymbol{z^{(k)}}; \theta_k)^t - \boldsymbol{G_{k-1}}\right] \\
\boldsymbol{H_k} &= \boldsymbol{G_k} - \boldsymbol{\Delta_k}\boldsymbol{\Delta_k}^t.
\end{aligned}$$

Knowing that the algorithm proposed above converges to a limiting value $\theta^\star$ and that l is regular enough, $(-\boldsymbol{H_k})$ converges to the inverse of the observed Fisher Information Matrix $-\partial_\theta^2 l(\theta^\star)$ (see [A1]).

3. Toward an efficient sampling step in high dimension

3.1. A new Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA)

In collaboration with Stéphanie Allassonnière, we proposed a new anisotropic version of the well-known Metropolis Adjusted Langevin Algorithm (MALA) [A12].

Let us first recall the steps of the Metropolis Adjusted Langevin Algorithm (MALA), which is a particular case of the Metropolis Hastings algorithm (see Gilks et al. [1996]). The focus is on optimizing the proposal distribution.

Let us denote by π the pdf of the target distribution with respect to the Lebesgue measure on $\mathcal{X}$, an open subset of $\mathbb{R}^l$. We assume that π is positive continuously differentiable. At each iteration k of this algorithm, a candidate X_c is simulated with respect to the Gaussian distribution with expectation $X_k + \frac{\sigma^2}{2} D(X_k)$ and covariance $\sigma^2 Id_l$ where X_k is the current value,

$$D(x) = \frac{b}{\max(b, |\nabla \log \pi(x)|)} \nabla \log \pi(x)\,, \tag{3.1}$$

Id_l is the identity matrix in $\mathbb{R}^l$ and $b > 0$ is a fixed truncation threshold. Note that the truncation of the drift D was already suggested in Gilks et al. [1996] to provide more stability.

The Gaussian proposal of the MALA algorithm is optimized with respect to its expectation guided by the Langevin diffusion. One step further is to optimize also its covariance matrix. A first step in this direction was proposed in Atchadé [2006]. The covariance matrix of the proposal is given by a projection of a stochastic approximation of the empirical covariance matrix. It produces an adaptive Markov chain. This process involves some additional tuning parameters that have to be calibrated. Since our goal is to use this sampler in an estimation algorithm, the sampler has a different target distribution (depending on the current estimate of the parameter) at each iteration. Therefore, the optimal tuning parameter may be different along the iterations of the estimation process. Although we agree with the idea of using adaptive chains, we prefer taking advantage of the dynamic of the estimation algorithm. For these reasons, we propose a sampler in the spirit of Atchadé [2006], Girolami and Calderhead [2011] or Marshall and Roberts [2012] without providing an adaptive chain. The adaption will result from the dependency of the target distribution with respect to the parameters of the model that are updated throughout the estimation algorithm. The proposal remains a Gaussian distribution, but both the drift and the covariance matrix depend on the gradient of the target distribution. At the k^{th} iteration, we are provided with X_k. The candidate is sampled from the Gaussian distribution with expectation $X_k + \delta D(X_k)$, and the covariance matrix $\delta\Sigma(X_k)$ denoted in the sequel $\mathcal{N}\left(X_k + \delta D(X_k), \delta\Sigma(X_k)\right)$ where $\Sigma(x)$ is given by:

$$\Sigma(x) = \varepsilon Id_l + D(x)D(x)^T \, , \tag{3.2}$$

D is defined in Equation (3.1) and $\varepsilon > 0$ is a small regularization parameter. Note that the threshold parameter b leads to a symmetric positive definite covariance matrix with bounded non zero eigenvalues. We introduce the gradient of $\log \pi$ into the covariance matrix to provide an anisotropic covariance matrix depending on the amplitude of the drift at the current value. When the drift is large, the candidate is likely to be far from the current value. This big step may not be of the right amplitude and a large variance will allow for more flexibility. Moreover, this makes it possible to explore a larger area around these candidates, which would not be possible with a fixed variance. On the other hand, when the drift is small in a particular direction, it means that the current value is within a region of high probability for the next value of the Markov chain. Therefore, the candidate should not move too far neither with a large drift nor with a large variance. This makes it possible to extensively sample around large modes, which is of particular interest. This covariance also makes it possible to treat the directions of interest with different amplitudes of variance, as is already the case with the drift. It also provides dependencies between coordinates since the directions of large variances are likely to be different from the Euclidean axis. This is taken into account here by introducing the Gram matrix of the drift into the covariance matrix.

Since our purpose was to plug the AMALA sampler into the SAEM-MCMC algorithm, we have to consider a parametric family of target densities $(\pi_s)_s$ and the corresponding transition kernels $(\Pi_s)_s$. In particular, we exhibit new assumptions that enable us to prove that the AMALA sampler is uniformly geometrically ergodic when considering a parametric family of target densities $(\pi_s)_s$ and the corresponding transition kernels $(\Pi_s)_s$.

We require a usual assumption on the stationary distributions namely the so-called super-exponential property given by:

(B1) For all $s \in \mathcal{S}$, the density π_s is positive with continuous first derivative such that:

$$\lim_{|x|\to\infty} n(x).\nabla \log \pi_s(x) = -\infty \tag{3.3}$$

and

$$\limsup_{|x|\to\infty} n(x).m_s(x) < 0 \tag{3.4}$$

where ∇ is the gradient operator in $\mathbb{R}^l$, $n(x) = \frac{x}{|x|}$ is the unit vector pointing in the direction of x and $m_s(x) = \frac{\nabla \pi_s(x)}{|\nabla \pi_s(x)|}$ is the unit vector in the direction of the gradient of the stationary distribution at point x.

We assume also some regularity properties of the stationary distributions with respect to s.

(B2) For all $x \in \mathcal{X}$, the functions $s \mapsto \pi_s$ and $s \mapsto \nabla_x \log \pi_s$ are continuous on $\mathcal{S}$.

We now define for some $\beta \in]0,1[$, $V_s(x) = c_s \pi_s(x)^{-\beta}$ where c_s is a constant so that $V_s(x) \geq 1$ for all $x \in \mathcal{X}$. Let also $V_1(x) = \inf_{s\in\mathcal{S}} V_s(x)$ and $V_2(x) = \sup_{s\in\mathcal{S}} V_s(x)$.

Let us assume conditions on V_2:

(B3) There exists $b_0 > 0$ such that, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, $V_2^{b_0}$ is integrable against $\Pi_s(x,.)$ and

$$\limsup_{b\to 0} \sup_{s\in\mathcal{S}, x\in\mathcal{X}} \Pi_s V_2^b(x) = 1\,. \tag{3.5}$$

We obtain the following result:

Proposition 3.1. *Assume (**B1-B3**). Let $\mathcal{K}$ a compact subset of $\mathcal{S}$. There exist a function $V \geq 1$, a set $\mathcal{C} \subseteq \mathcal{X}$, a probability measure ν such that $\nu(\mathcal{C}) > 0$ and there exist constants $\lambda \in]0,1[$, $b \in [0,\infty[$ and $\varepsilon \in]0,1]$ such that for all $s \in \mathcal{K}$:*

$$\begin{aligned} \Pi_s(x,A) &\geq \varepsilon\nu(A) \quad \forall x \in \mathcal{C} \quad \forall A \, \textit{Borel set}\,, \qquad (3.6)\\ \Pi_s V(x) &\leq \lambda V(x) + b 1_{\mathcal{C}}(x)\,. \qquad (3.7)\end{aligned}$$

The proof is performed in three steps. We first prove the existence of small sets being any compact subset of $\mathbb{R}^l$. Then, we prove the Drift condition for each transition kernel Π_s with a function V_s for all $s \in \mathcal{S}$ following the lines of Jarner and Hansen [2000] and Atchadé [2006]. The fact that both the drift and the covariance matrix are bounded even depending on the gradient of $\log \pi_s$ enables partially similar proofs.

Finally, the most technical step consists in exhibiting a single function V built from the family of functions $\{V_s\}_s$ satisfaying the Drift condition for all kernels Π_s for $s \in \mathcal{S}$.

We highlight the efficiency of the AMALA sampler by comparing its mixing properties with those of the MALA sampler. We used both algorithms to sample from a 100 dimensional normal distribution with a zero mean and a non-diagonal covariance matrix. Its eigenvalues range from 1 to 10. The eigendirections are chosen randomly. Ten examples of autocorrelations of both chains are plotted in Figure 1 where we can see that there is a significant benefit of using the anisotropic sampler. To evaluate the weight of the anisotropic term $D(x)D(x)^T$ in the covariance matrix, we compute its amplitude (as its non zero eigenvalue since it is a rank one matrix). We see that it is of the same order as the diagonal part on average and increases up to 15 times more. This shows the importance of the anisotropic term.

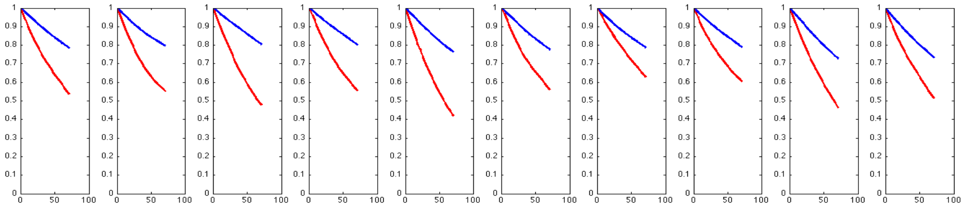


FIGURE 1. *Ten examples of autocorrelations of the MALA (blue) and AMALA (red) samplers to target the* 100 *dimensional normal distribution with anisotropic covariance matrix.*

3.2. Coupling the AMALA sampler and the SAEM algorithm

In collaboration with Stéphanie Allassonnière, we proposed using the AMALA sampler in the simulation step of the SAEM-MCMC algorithm presented in Section 2.2 [A12]. Thus, at each iteration k of the algorithm, simulated values of the missing data are drawn from the transition probability of the AMALA algorithm described in Section 3.1 with the current value of the parameters. The others steps remain unchanged.

We proved that the parameter estimate sequence generated by the AMALA-SAEM algorithm converges almost surely toward a stationary point of the likelihood under some regularity assumptions.

Theorem 3.1. *Assume some regularity conditions on the model and some usual conditions on the step size sequences. Assume that the family of conditional density probability functions* $\{\pi_{\hat{\theta}(s)}(\cdot|y),\ s \in \mathcal{S}\}$ *satisfies (**B1-B3**).*

Let K *be a compact subset of* $\mathcal{X}$ *and* $\mathcal{K}_0$ *a compact of* $\mathcal{S}$*. Then, for all* $z_0 \in \mathrm{K}$ *and* $s_0 \in \mathcal{K}_0$*, we have* $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$ *a.s. where* $(\theta_k)_k$ *is the sequence generated by the AMALA-SAEM Algorithm and* $\mathcal{L} \triangleq \{\theta \in \Theta, \partial_\theta l(\theta) = 0\}$.

The proof follows the same lines as the one given in [A3]. In particular, we first prove the sufficient usual Drift conditions (cf. Delyon et al. [1999]). We also prove in details that the transition kernel Π_s is Lipschitz in $s \in \mathcal{S}$. This proof extends the one proposed in Andrieu and Moulines [2006] to kernels and stationary distributions both depending on the parameter $s \in \mathcal{S}$.

The complete algorithm involves only three parameters: b, the threshold for the gradient that appears in the expectation as well as in the covariance matrix; δ, the scale on this gradient; and ε, a small regularization parameter to ensure a positive definite covariance matrix. The scale δ can be easily optimized in

terms of the data we are dealing with to adapt to the range of the drift. The value of the threshold b is, in practice, never reached.
We also established a Central Limit Theorem for the parameter estimate sequence generated by the AMALA-SAEM algorithm.

Theorem 3.1 ensures that the number of re-initializations of the sequence of stochastic approximation of the AMALA-SAEM Algorithm is finite almost surely. We can therefore consider only the non truncated sequence when we are interested in its asymptotic behavior. Moreover there are a priori multiple possible limiting points so we need to restrict our attention to the set of trajectories that converge to a given limiting point $\theta^* = \hat{\theta}(s^*)$.

Let us introduce some usual assumptions in the spirit of these of Delyon [2000].

(N1) The function h is C^1 in some neighborhood of s^* with first derivatives Lipschitz and J the Jacobean matrix of the mean field h in s^* has all its eigenvalues with negative real part.

(N2) Let $g_{\hat{\theta}(s)}$ be a solution of the Poisson equation $g - \Pi_{\hat{\theta}(s)} g = H_s - p_{\hat{\theta}(s)}(H_s)$ for any $s \in \mathcal{S}$. There exists a bounded function w such that

$$w - \Pi_{\hat{\theta}(s^*)} w = g_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}^T - \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)} (\Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)})^T - U \tag{3.8}$$

where the deterministic matrix U is given by :

$$U = \mathbb{E}_{\hat{\theta}(s^*)} \left[g_{\hat{\theta}(s^*)}(z) g_{\hat{\theta}(s^*)}(z)^T - \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}(z) \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}(z)^T \right] . \tag{3.9}$$

(N3) The step size sequence (γ_k) is decreasing and satisfies $\gamma_k = 1/k^\alpha$ with $2/3 < \alpha < 1$.

Theorem 3.2. *Under the assumptions of Theorem 3.1 and under **(N1)-(N3)**, the sequence $(s_k - s^*)/\sqrt{\gamma_k}$ converges in distribution to a Gaussian random vector with zero mean and covariance matrix Γ where Γ is the solution of the following Lyapunov equation:*

$$U + J\Gamma + \Gamma J^T = 0.$$

Moreover, denoting $\theta^* = \hat{\theta}(s^*)$, we have:

$$\frac{1}{\sqrt{\gamma_k}} (\theta_k - \theta^*) \to_{\mathcal{L}} \mathcal{N}(0, \partial_s \hat{\theta}(s^*) \Gamma \partial_s \hat{\theta}(s^*)^T).$$

This proof follows the lines of the proof of Theorem 25 of Delyon [2000]. However several of its assumptions are not satisfied by our stochastic approximation. Therefore we exhibit lighter assumptions leading to the same final result when combined with Drift and Hölder conditions.

3.3. Convergence study of the Wang Landau Algorithm

In collaboration with Gersende Fort, Benjamin Jourdain, Tony Lelièvre and Gabriel Stoltz, we study the Wang Landau algorithm, from a theoretical point of view as well as from a practical one [A11,A8].

I was interested in this study, having in mind to use the Wang Landau algorithm as sampler in the SAEM-MCMC algorithm. However I do not investigate further this possibility for two main reasons. First, the AMALA sampler developed in Section 3.1 gave very satisfaying theoretical and practical results when used as sampler in the SAEM-MCMC algorithm; second, using adaptive sampler such as the MALA version proposed by Atchadé [2006] into the estimation algorithm let to less performant results. One possible interpretation is that the dynamics of the parameter update in the estimation algorithm plays the role of the adaption process in the sampling step. Therefore I do not consider the coupling of the Wang Landau algorithm with the SAEM algorithm.

3.3.1. Introduction

The Wang-Landau algorithm belongs to the class of *free energy biasing techniques* (see Lelièvre et al. [2007]) which have been introduced in computational statistical physics to efficiently sample thermodynamic ensembles and to compute free energy differences. These algorithms can be seen as *adaptive importance sampling techniques*, the biasing factor being adapted on-the-fly in order to flatten the target probability measure along a given direction. Let us explain this with more details.

Let π be a multimodal probability measure over a high-dimensional space $\mathsf{X} \subseteq \mathbb{R}^D$. Classical algorithms to sample π such as a Metropolis-Hastings procedure with local proposal moves typically converge very slowly to equilibrium since high probability regions are separated by low probability regions. Averages have to be taken over very long trajectories in order to visit all the modes of the target probability measure π.

The idea of free energy biasing techniques is to *flatten the target probability along a well-chosen direction* through an importance sampling procedure in order to more easily sample π. More precisely, assume that we are given a measurable function $\mathcal{O}$ defined on X and with values in a low dimensional compact space, or in a discrete space. Let us introduce $\mathcal{O} * \pi$ the image of the measure π by $\mathcal{O}$: for any test function φ on the image $\mathcal{O}(\mathsf{X})$ of X by $\mathcal{O}$, $\int_{\mathcal{O}(\mathsf{X})} \varphi(y) \mathcal{O} * \pi(dy) = \int_{\mathsf{X}} \varphi(\mathcal{O}(x))\pi(dx)$. The free energy biased probability measure $\pi^\star$ is defined by the two following properties:

(i) the image $\mathcal{O} * \pi^\star$ of $\pi^\star$ by $\mathcal{O}$ is the uniform measure on $\mathcal{O}(\mathsf{X})$.

(ii) for each $y \in \mathcal{O}(\mathsf{X})$, the conditional distributions of x given $\mathcal{O}(x) = y$ under $\pi(dx)$ and $\pi^\star(dx)$ coincide i.e. there exists a measurable function $h : \mathcal{O}(\mathsf{X}) \to \mathbb{R}_+$ such that $\pi^\star(dx) = h(\mathcal{O}(x))\pi(dx)$.

The bottom line of free energy biasing techniques is that it should be easier to sample $\pi^\star$ than to sample π since, by construction, $\mathcal{O} * \pi^\star$ is the uniform probability measure. Then, sampling from π could be obtained by importance sampling from $\pi^\star$. The fact that $\pi^\star$ is indeed much easier to sample than π actually depends on the choice of $\mathcal{O}$. It is not an easy task to define and to design in practice a good choice for $\mathcal{O}$ and we do not discuss further these aspects here. This is related to the choice of a “good” reaction coordinate in the physics literature, which is a very debatable subject. We refer for example to Chopin et al. [2012] for such an analysis in the context of free energy biasing techniques used to sample posterior distributions in Bayesian statistics.

Of course, the difficulty is that in general, $\mathcal{O} * \pi$ is unknown so that it is not possible to sample from $\pi^\star$. The idea is then *to approximate* $\mathcal{O} * \pi$ *on the fly* in order to, in the longtime limit, sample from $\pi^\star$.

This is the adaptive feature of these algorithms: the importance sampling factor is computed as time goes, in order to penalize states (namely level sets of $\mathcal{O}$) which have already been visited. To approximate $\pi^\star$ at a given time, one could either use the occupation measure of the Markov chain up to the current time or one could use an approximation over many Markov chains running in parallel (see Lelièvre et al. [2007], Minoukadeh et al. [2010]).

In terms of mathematical analysis, approximations based on many replicas in parallel are typically easier to analyze, since they can be related in the limit of infinitely many replicas to mean field models for which powerful longtime convergence analysis techniques can be used. We refer for example to Lelièvre et al. [2008] and Lelièvre and Minoukadeh [2011] for such an analysis. The convergence analysis and, more importantly, the study of the efficiency of free energy biased techniques for approximations based on the occupation measure are much more involved since correlations in time of the Markov process play a crucial role. Our aim is to propose a convergence analysis for the Wang-Landau algorithm.

3.3.2. Description of the Wang-Landau algorithm

The Wang-Landau algorithm both computes a penalty sequence $\{\theta_n, n \geq 0\}$ approximating in the longtime limit the probability measure $\mathcal{O} * \pi$ and samples draws $\{X_n, n \geq 0\}$ distributed in the longtime limit according to $\pi^\star$. The update of the penalty sequence follows a Stochastic Approximation algorithm (see Benveniste et al. [1990], Robbins and Monro [1951]) and is of the form

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \mathcal{H}_n(X_{n+1}, \theta_n) \; .$$

Different strategies about the field $\mathcal{H}_n$ and the adaption schedule $\{\gamma_n, n \geq 1\}$ have been proposed in the literature. In the original paper of Wang and Landau [2001], the authors came up with a stochastic adaption schedule hereafter called *flat histogram Wang-Landau*. In this procedure, the updating parameter γ_n remains constant up to the random time when the sampling along the chosen order parameter $\mathcal{O}$ is approximately uniform, the “amount of uniformity” being measured according to the current value of γ_n. Then γ_n is lowered and a new updating procedure of the weights starts with a constant stepsize. Another strategy consists in a deterministic update of the adaption sequence $\{\gamma_n, n \geq 1\}$.

We now describe the Wang landau algorithm we studied. Let us consider a partition $\mathsf{X}_1, \ldots, \mathsf{X}_d$ of X in $d \geq 2$ elements, and define, for any $i \in \{1, \ldots, d\}$,

$$\theta_\star(i) \overset{\text{def}}{=} \int_{\mathsf{X}_i} \pi(x) \lambda(dx) \; . \tag{3.10}$$

In the following, X_i will be called the i-th *stratum*. Each weight $\theta_\star(i)$, which is assumed to be positive, gives the relative likelihood of the stratum $\mathsf{X}_i \subset \mathsf{X}$. In practice, the partitioning could be obtained by considering some smooth function $\xi : \mathsf{X} \to [a, b]$ and defining, for $i = 1, \ldots, d-1$,

$$\mathsf{X}_i = \xi^{-1}\Big([\alpha_{i-1}, \alpha_i)\Big) \; , \tag{3.11}$$

and $\mathsf{X}_d = \xi^{-1}\left([\alpha_{d-1}, \alpha_d]\right)$, with $a = \alpha_0 < \alpha_1 < \ldots \alpha_d = b$ (possibly, $a = -\infty$ and/or $b = +\infty$). In our previous notation, the order parameter is thus the discrete function $\mathcal{O}$ defined by

$$\forall x \in \mathsf{X},\ \mathcal{O}(x) = i \text{ if and only if } x \in \mathsf{X}_i\ . \tag{3.12}$$

In the following we consider a function $\mathcal{O}$ with values in a discrete finite set $\{1, \ldots, d\}$. Then, we have

$$\pi^\star(dx) = \frac{1}{d}\sum_{i=1}^{d} \frac{1_{\mathcal{O}(x)=i}}{\theta_\star(i)} \pi(dx)\ , \tag{3.13}$$

where $\theta_\star(i) = \pi\left(\{x \in \mathsf{X},\ \mathcal{O}(x) = i\}\right) = \mathcal{O} * \pi(i)$ for $i \in \{1, \ldots, d\}$.

The above discussion motivates the fact that the weights $\theta_\star(i)$ typically span several orders of magnitude, some sets X_i having very large weights, and other ones being very unlikely under π. Besides, trajectories bridging two very likely states may need to go through unlikely regions. To efficiently explore the configuration space, and sample numerous configurations in all the strata X_i, it is therefore a natural idea to resort to importance sampling strategies and reweight appropriately each subset X_i. A possible way to do so is the following. Let Θ be the subset of (non-degenerate) probability measures on $\{1, \ldots, d\}$ given by

$$\Theta = \left\{ \theta = (\theta(1), \ldots, \theta(d)) \ \middle|\ 0 < \theta(i) < 1 \text{ for all } i \in \{1, \ldots, d\} \text{ and } \sum_{i=1}^{d} \theta(i) = 1 \right\}\ .$$

For any $\theta \in \Theta$, we define the probability density π_θ on $(\mathsf{X}, \mathcal{X})$ (endowed with the reference measure λ) as

$$\pi_\theta(x) = \left(\sum_{i=1}^{d} \frac{\theta_\star(i)}{\theta(i)}\right)^{-1} \sum_{i=1}^{d} \frac{\pi(x)}{\theta(i)}\ 1_{\mathsf{X}_i}(x)\ . \tag{3.14}$$

This measure is such that the weight of the set X_i under π_θ is proportional to $\theta_\star(i)/\theta(i)$. In particular, all the strata X_i have the same weight under $\pi_{\theta_\star}$. Unfortunately, $\theta_\star$ is unknown and sampling under $\pi_{\theta_\star}$ is typically unfeasible.

The Wang-Landau algorithm precisely is a way to overcome these difficulties: at each iteration of the algorithm, a weight vector $\theta_n = (\theta_n(1), \ldots, \theta_n(d))$ is updated based on the past behavior of the algorithm and a point is drawn from a Markov kernel P_{θ_n} with invariant density π_{θ_n}. The intuition for the convergence of this algorithm is that if $\{\theta_n, n \geq 0\}$ converges to $\theta_\star$ then the draws are asymptotically distributed according to the density $\pi_{\theta_\star}$. Conversely, if the draws are under $\pi_{\theta_\star}$, then the update of $\{\theta_n, n \geq 0\}$ is chosen such that it converges to $\theta_\star$.

We now describe precisely the algorithm we considered. Let $\{\gamma_n, n \geq 1\}$ be a $[0, 1)$-valued deterministic sequence. For any $\theta \in \Theta$, denote by P_θ a Markov transition kernel onto $(\mathsf{X}, \mathcal{X})$ with unique stationary distribution $\pi_\theta(x)\lambda(dx)$; for example, P_θ is one step of a Metropolis-Hastings algorithm with target probability measure $\pi_\theta(x)\lambda(dx)$.

Consider an initial value $X_0 \in \mathsf{X}$ and an initial set of weights $\theta_0 \in \Theta$ (typically, in absence of any prior information, $\theta_0(i) = 1/d$). Define the process $\{(X_n, \theta_n), n \geq 0\}$ as follows: given the current value (X_n, θ_n),

- Draw X_{n+1} under the conditional distribution $P_{\theta_n}(X_n, \cdot)$;
- Set $i = \mathcal{O}(X_{n+1})$ where $\mathcal{O}$ is given by (3.12). The weights are then updated as

$$\begin{cases} \theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1}\ \theta_n(i)\,(1-\theta_n(i))\,, & \\ \theta_{n+1}(k) = \theta_n(k) - \gamma_{n+1}\ \theta_n(k)\ \theta_n(i) & \text{for } k \neq i. \end{cases} \tag{3.15}$$

Note that since $\gamma_n \in [0,1)$, $\theta_n \in \Theta$ for any $n \geq 0$. The update of the probability vector θ_n can be recast equivalently into the stochastic approximation framework upon writing

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\, H(X_{n+1}, \theta_n)\ , \tag{3.16}$$

where $H : \mathsf{X} \times \Theta \to [-1,1]^d$ is defined componentwise by

$$H_i(x,\theta) = \theta(i)\,(1_{\mathsf{X}_i}(x) - \theta(\mathcal{O}(x)))\ . \tag{3.17}$$

The updating strategy (3.15) (or equivalently (3.16)) is a modification of the original Wang-Landau algorithm obtained by (i) using a deterministic schedule for the evolution of the step-sizes used to modify the values of the weights (instead of reducing the value of these step-sizes at random times when the empirical frequencies of the strata are sufficiently uniform: this is the flat histogram version of the Wang-Landau algorithm) and (ii) linearizing at first order in γ_n the update of the weight θ_n.

Concerning this second point, the standard Wang-Landau update is

$$\theta_{n+1}(i) = \theta_n(i)\frac{1+\gamma_{n+1}1_{\mathsf{X}_i}(X_{n+1})}{1+\gamma_{n+1}\theta_n(\mathcal{O}(X_{n+1}))}\ . \tag{3.18}$$

The update (3.15) is obtained from (3.18) in the limit of small γ_n.

3.3.3. Convergence of the Wang-Landau algorithm

Despite the Wang-Landau algorithm has been successfully applied for many problems of practical interest, there are many open questions about its longtime behavior and its efficiency. Such a longtime behavior study relies on the convergence of stochastic approximation algorithms with Markovian inputs (see Andrieu et al. [2005], Benveniste et al. [1990]) combined with the convergence of adaptive Markov chain Monte Carlo samplers Fort et al. [2012]; for both parts, the stability of the sequence $\{\theta_n, n \geq 0\}$ is a fundamental property. Stability here means that the sequence $\{\theta_n, n \geq 0\}$ remains in a compact subset of the probability measures on $\{1,\ldots,d\}$ with support equal to the support of $\mathcal{O} * \pi$.

We consider here the Wang-Landau algorithm with a deterministic adaption sequence $\{\gamma_n, n \geq 1\}$ for a precise definition of the algorithm) and address both the convergence of $\{\theta_n, n \geq 0\}$ to $\mathcal{O} * \pi$ and the convergence of $\{X_n, n \geq 0\}$ to $\pi^\star$. More precisely, we prove first that the sequence $\{\theta_n, n \geq 0\}$ is stable, which is a crucial point for applications: no *ad hoc* stabilization techniques (such as truncation at randomly varying bounds Chen et al. [1988]) is required. We also prove the almost-sure convergence of $\{\theta_n, n \geq 0\}$ as well as a Central Limit Theorem. We then prove the ergodicity and a strong law of large numbers for the draws $\{X_n, n \geq 0\}$.

We adopt this linear update (3.15) for the stability and the convergence analysis. The main advantage is that it makes the proof of convergence simpler; nevertheless, since γ_n converges to zero, these stability and convergence results are unchanged and could be proved along the same lines for the standard Wang-Landau update (3.18).

The proof of the convergence of the Wang-Landau algorithm relies on its reformulation (3.16) as a stochastic approximation procedure. Since the draws $\{X_n, n \geq 1\}$ satisfy for any measurable non-negative function f:

$$\mathbb{E}\left[f(X_{n+1})|\mathcal{F}_n\right] = P_{\theta_n} f(X_n) \ , \tag{3.19}$$

where $\mathcal{F}_n$ denotes the σ-field $\sigma(\theta_0, X_0, X_1, \ldots, X_n)$.

The main difficulty, when proving the almost-sure convergence of such algorithms, is the stability, namely how to ensure that the sequence $\{\theta_n, n \geq 0\}$ remains in a compact subset of Θ. We use a traditional approach to answer this question: we first prove that our algorithm satisfies a recurrence property *i.e.* the sequence $\{\theta_n, n \geq 0\}$ visits infinitely often a compact subset of Θ; we then show that there exists a Lyapunov function with respect to the *mean-field* function $h : \Theta \to [-1, 1]^d$

$$h(\theta) = \int_{\mathsf{X}} H(x, \theta)\, \pi_\theta(x)\, \lambda(dx) = \left(\sum_{j=1}^{d} \frac{\theta_\star(j)}{\theta(j)} \right)^{-1} (\theta_\star - \theta) \ , \tag{3.20}$$

with strong enough properties so that the recurrence property implies stability. Different strategies based on truncations are proposed in the literature to circumvent the stability problem (see *e.g.* Kushner and Yin [1997]). The most popular technique is the truncation to a fixed compact set but this is not a satisfactory solution since the choice of this compact is delicate: a necessary condition for convergence is that the compact contains the unknown desired limit. An adaptive truncation has been proposed by Chen et al. [1988] which avoids the main drawbacks of the deterministic truncation approach.

We first prove that, under conditions on the target density π and the step-size sequence $\{\gamma_n, n \geq 1\}$, the algorithm (3.16) is recurrent, so that such truncation techniques are not required.

We detailed here the assumptions on the Metropolis dynamics and on the adaption rate. Our conditions fall into three categories: conditions on the equilibrium measure (see A1), on the transition kernels $\{P_\theta, \theta \in \Theta\}$ (see A2) and conditions on the step-size sequence $\{\gamma_n, n \geq 1\}$ (see A3). It is assumed that

A1 The probability density π with respect to the measure λ is such that $0 < \inf_{\mathsf{X}} \pi \leq \sup_{\mathsf{X}} \pi < \infty$. In addition, $\inf_{1 \leq i \leq d} \theta_\star(i) > 0$ where $\theta_\star$ is given by (3.10).

The minorization condition on π certainly is the most restrictive assumption: it is introduced in order to prove the recurrence of the algorithm (3.16). This condition can be removed by adding a stabilization step to (3.16) (such as a truncation technique at random varying bounds Chen et al. [1988], Kushner and Yin [1997]) in order to ensure the recurrence.

A2 For any $\theta \in \Theta$, P_θ is a Metropolis-Hastings transition kernel with invariant distribution $\pi_\theta \; d\lambda$, where π_θ is given by (3.14), and with symmetric proposal kernel $q(x, y)\lambda(dy)$ satisfying $\inf_{\mathsf{X}^2} q > 0$.

The minorization condition on q implies that the transition kernels $\{P_\theta, \theta \in \Theta\}$ are uniformly (geometrically) ergodic.This property allows a simple presentation of the main ingredients for the limiting behavior analysis of the algorithm. Extensions to a more general case could be done by using the same tools as in Fort et al. [2012] (see also [Andrieu et al., 2005, Section 3]) and controlling the dependence upon θ of the ergodic behavior.

A3 The sequence $\{\gamma_n, n \geq 1\}$ is a $[0, 1)$-valued deterministic sequence such that

1. $\{\gamma_n, n \geq 1\}$ is a non-increasing sequence and $\lim_n \gamma_n = 0$;
2. $\sum_n \gamma_n = \infty$;
3. $\sum_n \gamma_n^2 < \infty$.

We first proved the following result.

Proposition 3.2. *Under A1 and A2, there exists $\rho \in (0, 1)$ such that for all $\theta \in \Theta$, for all $x \in \mathsf{X}$ and for all $A \in \mathcal{X}$, it holds:*

$$P_\theta(x, A) \geq \rho \int_A \pi_\theta(x)\, \lambda(dx) \,, \tag{3.21}$$

$$\sup_{\theta \in \Theta} \sup_{x \in \mathsf{X}} \|P_\theta^n(x, \cdot) - \pi_\theta \, d\lambda\|_{\mathrm{TV}} \leq 2(1 - \rho)^n, \tag{3.22}$$

where for a signed measure μ, the total variation norm is defined as

$$\|\mu\|_{\mathrm{TV}} = \sup_{\{f \,:\, \sup_{\mathsf{X}} |f| \leq 1\}} |\mu(f)| \,.$$

We state that, almost surely, there exists a compact subset of Θ such that θ_n belongs to this compact subset for infinitely many n. For any $n \geq 0$, set

$$\underline{\theta}_n = \min_{1 \leq j \leq d} \theta_n(j) \,. \tag{3.23}$$

We prove the following theorem:

Theorem 3.1. *Assume A1, A2 and A31. Then,*

$$\mathbb{P}\left(\limsup_{n \to \infty} \underline{\theta}_n > 0\right) = 1 \,. \tag{3.24}$$

The proof is based on the following consideration. The value of the smallest weight increases when the chain goes into the corresponding stratum (see the updating formula (3.15)). Under the stated assumptions, we prove that the chain $\{X_n, n \geq 0\}$ returns in the strata of smallest weights often enough for the smallest weight to remain isolated from 0.

Then we addressed the almost-sure convergence of the sequence $\{\theta_n, n \geq 0\}$ to $\theta_\star$.

Theorem 3.2. *Assume A1, A2 and A3. Then,* $\mathbb{P}\left(\lim_{n \to \infty} \theta_n = \theta_\star\right) = 1$.

The proof relies on Andrieu et al. [2005] which provides sufficient conditions for convergence of stochastic approximation techniques. The first step consists in rewriting the weight update (3.16) as

$$\theta_{n+1} = \theta_n + \gamma_{n+1} h(\theta_n) + \gamma_{n+1} \Big(H(X_{n+1}, \theta_n) - h(\theta_n) \Big) \;, \tag{3.25}$$

where h is given by (3.20). The heuristic idea is that, if the step-size quickly is sufficiently small, and the Metropolis dynamics converges sufficiently fast to equilibrium for θ fixed (a result given by Proposition 3.2), the update of θ_n is indeed close to an update with the averaged drift $h(\theta_n)$. However, in order for the updates of the weights to be non-negligible, the step-sizes should not be too small. The balance between these two opposite effects is encoded in the conditions A32-3.

From a technical viewpoint, the proof of the theorem relies on two main tools. The first one is to show that the function $V : \Theta \to \mathbb{R}_+$ given by

$$V(\theta) \overset{\text{def}}{=} \sum_{i=1}^{d} \theta_\star(i) \log\left(\frac{\theta_\star(i)}{\theta(i)} \right) \tag{3.26}$$

is a Lyapunov function with respect to the mean-field h, namely $\langle \nabla V(\theta), h(\theta) \rangle < 0$ for $\theta \neq \theta_\star$ and $\langle \nabla V(\theta_\star), h(\theta_\star) \rangle = 0$ (here, $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^d$). This motivates the fact that $\{\theta_n, n \geq 0\}$ may converge to $\theta_\star$. The second important result establishes that the remainder term $\gamma_{n+1} \left(H(X_{n+1}, \theta_n) - h(\theta_n) \right)$ in (3.25) vanishes in some sense. This step is quite technical and requires regularity-in-θ of the transition kernels P_θ and the invariant distributions π_θ. The conclusion then follows from [Andrieu et al., 2005, Theorem 2.3] and Theorem 3.1.

Finally we studied the asymptotic behavior of the chain $\{X_k, k \geq 0\}$. We established the following result.

Theorem 3.3. *Assume A1, A2 and A3. Then, for any bounded measurable function f,*

$$\lim_{n \to \infty} \mathbb{E}\left[f(X_n) \right] = \int_{\mathsf{X}} f(x)\, \pi_{\theta_\star}(x)\, \lambda(dx) \;, \tag{3.27}$$

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \xrightarrow{\text{a.s.}} \int_{\mathsf{X}} f(x)\, \pi_{\theta_\star}(x)\, \lambda(dx) \;. \tag{3.28}$$

This theorem shows that the distribution of the sample X_n converges to $\pi_{\theta_\star}(x)\lambda(dx)$, where, we recall

$$\pi_{\theta_\star}(x) = \frac{1}{d} \sum_{i=1}^{d} \frac{\pi(x)}{\theta_\star(i)} 1_{\mathsf{X}_i}(x) \;.$$

Moreover, the empirical mean of the samples $\{f(X_k), k \geq 0\}$ converges to $\int f\, \pi_{\theta_\star}\, d\lambda$. Hence, although the weights θ_n evolve in the adaptive algorithm, ergodic averages can be thought of as averages with fixed weights $\theta_\star$.

In many practical cases, averages with respect to π are of interest. In this case, the Wang-Landau procedure is used as some adaptive importance sampling strategy. In order to obtain averages according to π along a trajectory of the algorithm, some reweighting has to be considered. A natural strategy is to use some stratified-type weighted sum of the samples $\{X_k,\, k \geq 1\}$:

$$\mathcal{I}_n(f) \overset{\text{def}}{=} d \sum_{i=1}^{d} \theta_n(i) \left(\frac{1}{n} \sum_{k=1}^{n} f(X_k) 1_{\mathsf{X}_i}(X_k) \right) .$$

We also prove the following result:

Theorem 3.4. *Assume A1, A2 and A3. Then for any bounded measurable function f,*

$$\lim_{n\to\infty} d\, \mathbb{E} \left[\sum_{i=1}^{d} \theta_n(i)\, f(X_n)\, 1_{\mathsf{X}_i}(X_n) \right] = \int_{\mathsf{X}} f(x)\, \pi(x)\, \lambda(dx) \; , \tag{3.29}$$

$$\mathcal{I}_n(f) \xrightarrow{\text{a.s.}} \int_{\mathsf{X}} f(x)\, \pi(x)\, \lambda(dx) \; . \tag{3.30}$$

There are of course many other reweighting strategies. We have discussed only one possible choice. May be the above estimator is not the best one.

We also state a Central Limit Theorem on the error $(\theta_n - \theta_\star)$. We show that the rate of convergence depends upon the step-size sequence $\{\gamma_n, n \geq 1\}$ and discuss an averaging strategy in order to reach the optimal rate of convergence. An additional assumption is required on the sequence $\{\gamma_n, n \geq 1\}$:

A4 $\lim_n \gamma_n \sqrt{n} = 0$, and one of the following condition holds:

1. $\log(\gamma_n/\gamma_{n+1}) = \mathrm{o}(\gamma_n)$;
2. $\log(\gamma_n/\gamma_{n+1}) \sim \gamma_n/\gamma_\star$ with $\gamma_\star > d/2$.

Theorem 3.5. *Assume that A1, A2, A3 and A4 hold. Then $\{\gamma_n^{-1/2} (\theta_n - \theta_\star)\, ,\, n \geq 1\}$ converges in distribution to a centered Gaussian distribution with variance-covariance matrix $\sigma^2 U_\star$ where $\sigma^2 = d/2$ in case A4(1) and $\sigma^2 = \gamma_\star d/(2\gamma_\star - d)$ in case A4(2),*

$$U_\star \overset{\text{def}}{=} \int_{\mathsf{X}} \left\{ \widehat{H}_{\theta_\star}(x) \widehat{H}_{\theta_\star}^T(x) - P_{\theta_\star} \widehat{H}_{\theta_\star}(x)\; P_{\theta_\star} \widehat{H}_{\theta_\star}^T(x) \right\} \pi_{\theta_\star}(x)\, \lambda(dx) \; , \tag{3.31}$$

and

$$\widehat{H}_{\theta_\star} \overset{\text{def}}{=} \sum_{n\geq 0} P_{\theta_\star}^n \left(I - \pi_{\theta_\star} \right) H(\cdot, \theta_\star) = \sum_{n \geq 0} P_{\theta_\star}^n \left(H(\cdot, \theta_\star) - h(\theta_\star) \right) \; .$$

Notice that $\widehat{H}_{\theta_\star}$ is the Poisson solution associated to the pair $(P_{\theta_\star}, H(\cdot, \theta_\star))$, namely $\widehat{H}_{\theta_\star}$ is a solution to: find $g : \mathsf{X} \to \mathbb{R}$ such that

$$g - P_{\theta_\star} g = H(\cdot, \theta_\star) - \int_{\mathsf{X}} H(x, \theta_\star)\, \pi_{\theta_\star}(x)\, \lambda(dx) \; .$$

By Proposition 3.2 and the results of [Meyn and Tweedie, 2009, Chapter 17], such a function exists and is unique up to an additive constant.

Theorem 3.5 shows that the rate of convergence depends upon the step-size sequence $\{\gamma_n, n \geq 1\}$: when $\gamma_n = \gamma_\star/n^\alpha$ for $\alpha \in (1/2, 1]$, the maximal rate of convergence is reached with $\alpha = 1$ and the rate is $O(n^{-1/2})$. When $\gamma_n = \gamma_\star/n$, one could be interested in optimizing the variance-covariance matrix: introducing a gain matrix Γ in the algorithm (3.16) yields the update

$$\check{\theta}_{n+1} = \check{\theta}_n + \gamma_{n+1}\Gamma \; H(X_{n+1}, \check{\theta}_n) \; .$$

It is proved in [Benveniste et al., 1990, Proposition 4 p.112] that for a large family of gain matrix (so-called "admissible gains") a Central Limit Theorem still holds for the sequence of random variables $\{\sqrt{n}(\theta_n - \theta_\star), n \geq 0\}$, the minimal variance-covariance is equal to $d^2 U_\star$ and is reached with $\Gamma = d\gamma_\star^{-1}\mathrm{Id}$. Since Γ is a scalar matrix, this discussion evidences that the minimal variance-covariance matrix $d^2 U_\star$ is reached when choosing $\gamma_n = d/n$.

From a practical point of view, it is known that stochastic approximation algorithms are more efficient when the step-size sequence decreases at a slow rate: in the polynomial schedule, this means that $\gamma_n = \gamma_\star/n^\alpha$ with α close to $1/2$. As shown by Theorem 3.5, this yields a slower rate of convergence. Nevertheless, combining Wang-Landau update with an *averaging technique* allows to reach the optimal rate of convergence and the optimal variance-covariance matrix: by applying [Fort, 2014, Theorem 1.4], it can be proved that $\{\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^n \theta_k - \theta_\star\right), n \geq 1\}$ converges in distribution to a centered Gaussian distribution with variance-covariance matrix $d^2 U_\star$.

3.3.4. Numerical studies

We complete our theoretical convergence results by a simulation study to discuss the efficiency of the Wang-Landau procedure.

This algorithm is actually known to be useful in metastable situations, namely when the original Markov chain (with transition kernel P_{θ_0}) remains trapped for very long times in some regions (called the metastable states). Metastability is one of the major bottleneck of standard Markov Chain Monte Carlo techniques, since ergodic averages should be considered over very long times in order to obtain accurate results. Our aim is to show that in such a metastable situation, the Wang-Landau algorithm indeed is an efficient sampling procedure. We consider a toy model composed of only three strata: two large probability strata (the metastable states) separated by a low probability stratum (the transition state). We analyze theoretically the first exit times out of a metastable state.

We show that the Wang-Landau algorithm allows to rapidly escape from a metastable state, namely from a large probability stratum surrounded by small probability strata. We are able to precisely quantify the time the system needs to go from the first metastable state to the second one. We show in particular that the exit time is dramatically reduced with the Wang-Landau dynamics compared to the corresponding non-adaptive dynamics.

Finally we show that (most of) the results obtained for the very simple three-state model are still valid for a less simple example inspired by target measures used in computational statistical physics.

Part II

Inference in mixed effects models, in deformable template models and in Gaussian regressions

The second part included my contributions dealing with modeling and parametric estimation in mixed effects models, in particular in the deformable template model in image analysis, and with the testing problematic in Gaussian regression models. Related studies include [A2,A3,A4,A5,A10,A12,A13,P3,P4,P5,T].

In collaboration with Marc Lavielle, we proposed using the SAEM-MCMC algorithm to obtain a numerical value of the MLE in mixed effects models. These models are particularly used to analyze repeated longitudinal data. When the model is linear in the random effects and the residual random error Gaussian, the likelihood has an explicit analytical expression and the maximum likelihood estimate can be calculated through some direct optimization. On the other hand, outside this specified context, the likelihood usually does not admit an explicit analytical form, making it difficult to evaluate the MLE. It is then necessary to appeal to more complex numerical procedures, often time-consuming, and not always provided with theoretical convergence properties. The SAEM-MCMC algorithm is an efficient solution, fast and convergent, to evaluate the MLE in mixed effects models, in particular, in non-linear ones. We used it for parameter estimation in growth curve models and in pharmacodynamic models [A2]. In collaboration with Alain Trouvé and Stéphanie Allassonnière,we implemented the SAEM-MCMC algorithm with the additional truncation step to estimate parameters in a deformable template model in a Bayesian setting useful in medical imaging where samples are usually small [A3]. Deformable template models allow us to represent a sample of images with a reference image called template and geometrical deformations, so that each image of the sample is obtained as the result of the geometrical deformation of the template, up to a small error term. The geometrical deformations are the latent variables of such a model. We first considered the case of small linear deformations and used a hybrid Gibbs sampler as the MCMC method. This application motivated the development on high dimensional latent variables presented in the first part of this manuscript, since the hybrid Gibbs sampler becomes very time-consuming as the dimension of the latent variable increases. Later, in collaboration with Stéphanie Allassonnière,we proposed a specific algorithm for parameter estimation in the multicomponent model which is a mixture of deformable template models motivated by a crucial modeling issue [A5]. In numerous applications, it is necessary to constrain the type of geometrical deformations considered, in particular, so that the diffeomorphic deformations do not allow overlapping and prevent topological changes from occuring between the template and the observation, thus creating the need of a mixture model. Since the SAEM-MCMC algorithm is sensitive to numerical phenomena such as trapping states in such a case, we proposed a specific stochastic estimation algorithm and implemented it. We also established its convergence property toward a local maximum of the observed likelihood. Finally, in collaboration with Stéphanie Allassonnière and Stanley Durrlemann, we applied the SAEM-MCMC algorithm provided with the AMALA sampler to the very complex deformable template model using large diffeomorphic deformations [A12,A13]. The new algorithm takes on its full meaning in this high dimensional setting. We also developed an extension that allowed us to optimize the position of the control points of the deformation, simultaneously with the estimation of the other model parameters. We proposed an empirical criterion to select an optimal number of control points as well, leading to the optimization of the model dimension. All the applications related to deformable template models and the different estimation algorithms were performed on the US POSTAL handwritten digit database and 2D and 3D medical image databases. The corresponding experimental results are presented in Section 5.5.

Motivated at the beginning by the model choice issue for mixed effects models, we finally proposed, in collaboration with Sylvie Huet a goodness-of-fit test for testing a linear hypothesis on the expectation of a Gaussian vector with block correlated errors with a known covariance structure up to some parameters [A10]. We established that our test procedure was asymptotically of the nominal level and consistent over a large class of alternatives. We also proposed a bootstrap version of our procedure. Using a simulation study, we evaluated the finite sample size properties of our procedures and applied them to a forest cover dataset of Galicia.

4. Maximum Likelihood Estimation in Mixed Effects Models

4.1. Mixed effects models

Mixed effects models were introduced mainly for modeling repeated longitudinal data (see Davidian and Giltinan [1995]). Such models allow us to analyze responses of a population of individuals that share a global behavior with the exception of some individual variations. Some of them are shared by all the individuals of the population, whereas the others are random, depending on the individuals or possibly on sub-groups of the population. Thus, the model has two types of parameters: global parameters defined as the fixed effects, and parameters that vary among the population defined as random effects. These kinds of observations are usually the result of repeated measurements of some individuals that are repeatedly observed under different experimental conditions. Such settings are very common in practice, for example, in the fields of pharmacokinetics, biological growth, epidemiology and econometry. We refer to Pinheiro and Bates [2000] for more details on mixed effects models.

Let us consider here the following general mixed effects model defined as:

$$y_{ij} = g(t_{ij}, \phi_i, \beta) + h(t_{ij}, \phi_i, \beta)\varepsilon_{ij} \quad , \ 1 \leq i \leq n \quad , \quad 1 \leq j \leq m_i \tag{4.1}$$

where y_{ij} is the jth observation of the ith individual, at some known instant t_{ij}. Here, n is the number of individuals and m_i is the number of observations of individual i. The random effects (ϕ_i) are assumed to be independent identically distributed. We denote by η the parameter of their common distribution. The vector β also denotes unknown population parameters, that do not appear in the distribution of the random effects ϕ_i. The within-group errors (ε_{ij}) are assumed to be *i.i.d.* Gaussian random variables with a mean zero and unknown variance σ^2. We assume that the (ε_{ij}) and the (ϕ_i) are mutually independent. The model is said to be nonlinear if the functions g or h are nonlinear in the random effects ϕ_i.

Note that mixed effects models are a particular type of latent variable model introduced in Section 1 since the responses (y_{ij}) are observed and the random effects (ϕ_i) are unobserved.

4.2. Maximum Likelihood Estimation

We choose a frequentist approach and consider the maximum likelihood estimator of the unknown parameter vector $\theta = (\beta, \eta, \sigma^2)$.

The theoretical study of the MLE has been addressed by Nie and Yang [2005]. It is a difficult task. The authors studied different asymptotics, as the number of individuals and/or the number of observations per individual tend to infinity. Under some regularity assumptions on the model, they proved the consistency and the asymptotic normality of the MLE.

Let us now focus on the computation of the MLE in practice. For linear mixed effects models, the estimation of the unknown parameters can be treated with the usual EM algorithm (Dempster et al. [1977]), or with a Newton-Raphson algorithm (Pinheiro and Bates [2000]). However, nonlinear functions are often more suitable for modeling considerations. Estimating the parameters by maximizing the observed likelihood then requires a specific approach. Different methods, based generally on linearization of the log-likelihood, were suggested for dealing with nonlinear models. A Laplace approximation was proposed by Vonesh [1996], and a Bayesian approach was proposed by Racine-Poon [1985], Wakefield et al. [1994], Wakefield [1996]. Walker [1996] uses a Monte-Carlo EM algorithm, whereas a simulated pseudo maximum likelihood estimator for these specific models was developed by Concordet and Nunez [2002]. These methods are either not proven to be convergent or are very time-consuming. Therefore, applying the SAEM-MCMC algorithm presented in Section 2 is a powerful tool for the estimation task in mixed effects models.

4.3. Applications to a growth curve model and to a pharmacodynamic model

4.3.1. The orange tree dataset

We consider the example of orange trees to illustrate our algorithm. This data was studied by Pinheiro and Bates [2000] and is available on S-plus. The data consist in seven measurements of the trunk circumference of each of five orange trees. Pinheiro and Bates [2000] uses a logistic curve to model the trunk circumference y_{ij} of tree i at age x_j:

$$\begin{aligned} y_{ij} &= g(x_j, \phi_i; \beta_1, \beta_2) + \varepsilon_{ij} \quad 1 \le i \le n \ , \ 1 \le j \le m, && (4.2) \\ g(x_j, \phi_i; \beta_1, \beta_2) &= \frac{\phi_i}{1 + \exp\left(-\frac{x_j - \beta_1}{\beta_2}\right)} \ . && (4.3) \end{aligned}$$

We assume here that the error terms ε_{ij} are independent Gaussian centered variables of variance σ^2. On the one hand, the asymptotic trunk circumference $\boldsymbol{\phi_i}$ is treated as a random effect and is assumed to be

TABLE 1
Comparison of EM and SAEM estimates after 100 and 1000 iterations.

Parameters	β_1	β_2	μ	τ^2	σ^2
$\boldsymbol{\theta_0}$	650	250	100	50	10
θ_{100}^{EM}	727.89	348.06	192.05	1001.45	61.51
θ_{1000}^{EM}	727.91	348.07	192.05	1001.49	61.51
θ_{100}^{SAEM}	725.34	346.11	191.46	1020.26	60.56
θ_{1000}^{SAEM}	727.36	347.67	191.93	1003.76	61.52

Gaussian with mean μ and variance τ^2. On the other hand, the age at which the tree attains half of its asymptotic trunk circumference β_1 and the growth scale β_2 are treated as two fixed effects. Setting

$$g_1(\phi_i) = \phi_i \quad \text{and} \quad g_2(\beta_1, \beta_2, x_j) = \frac{1}{1 + \exp\left(-\frac{x_j - \beta_1}{\beta_2}\right)}, \tag{4.4}$$

the likelihood of the complete model has the form:

$$f(y, \boldsymbol{\phi}; \theta) = (2\pi\sigma^2)^{-\frac{nm}{2}} (2\pi\tau^2)^{-\frac{m}{2}} \times \exp\left[-\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - g_1(\phi_i) g_2(\beta_1, \beta_2, x_j))^2 - \frac{1}{2\tau^2} \sum_i (\phi_i - \mu)^2\right]$$

where $\theta = (\beta_1, \beta_2, \mu, \tau^2, \sigma^2)$.

Note that the EM algorithm can be implemented in this model. Thus, the value obtained with this algorithm may be considered as the maximum likelihood estimate of θ. The estimation of the parameters after 100 and 1000 iterations with EM and SAEM are displayed in Table 1.

We can observe that EM has almost converged after 100 iterations. In this example, the step size sequence (γ_k) used for SAEM was: $\gamma_k = 1$ for $1 \leq k \leq 100$ and $\gamma_k = (k - 99)^{-1}$ for $k \geq 100$. After some iterations, the SAEM algorithm has converged to a neighborhood of the MLE of θ. Since $\gamma_k = 1$ during the first iterations, no further stochastic approximation is performed. Thus, the behavior of the sequence (θ_k^{SAEM}) remains quite perturbed until iteration 100. After that, the introduction of a decreasing step size allows the almost sure convergence of the sequence (θ_k^{SAEM}) to $\widehat{\theta}^{MLE}$.

The Fisher Information of the MLE can also be estimated by using the stochastic approximation scheme presented in Section 2.3. In Table 2, we present the estimated standard deviation of each component of (θ^{EM}) and (θ^{SAEM}), obtained after 100 and 1000 iterations. We observe once again that the SAEM algorithm provides a good estimation in just a few iterations.

TABLE 2
Estimation of the standard deviation of θ^{EM} and θ^{SAEM} obtained after 100 and 1 000 iterations.

Parameters	β_1	β_2	μ	τ^2	σ^2
$\hat{\sigma}(\theta_{100}^{EM})$	13.51	13.04	14.15	633.39	14.70
$\hat{\sigma}(\theta_{1000}^{EM})$	13.51	13.04	14.15	633.40	14.70
$\hat{\sigma}(\theta_{100}^{SAEM})$	12.89	12.51	13.83	604.53	13.41
$\hat{\sigma}(\theta_{1000}^{SAEM})$	13.51	13.04	14.15	633.40	14.70

4.3.2. Extension to a heteroscedastic model

Let us now consider the following heteroscedastic model for this same example:

$$y_{ij} = \frac{\phi_i}{1+\exp\left(-\frac{x_j-\beta_1}{\beta_2}\right)}(1+\varepsilon_{ij}). \tag{4.5}$$

Note that we are outside the scope of the exponential model. One solution consists in regarding the fixed parameters (β_1, β_2) as the realization of a Gaussian random vector of mean (μ_1, μ_2) and a diagonal covariance matrix with diagonal terms (τ_1^2, τ_2^2). As before, $\boldsymbol{\phi} = (\phi_i)$ is a sequence of *i.i.d.* Gaussian random variables of mean μ and variance τ^2.

It is important to observe that we do not change the model by doing this. The fixed effects remain fixed effects, since we still consider only one vector (β_1, β_2) for the whole population.

4.3.3. Comparisons with other methods on a pharmacodynamic model

In this section, we consider the nonlinear population pharmacodynamic model used by Walker [1996] for comparing the MLEs obtained with the EM algorithm to approximate MLEs obtained from the NONMEM package.

Simulated data are given by:

$$y_{ij} = \phi_{1i} - \frac{\phi_{2i}x_j}{\phi_{3i}+x_j} + \varepsilon_{ij} \quad ; \quad 1 \leq i \leq n \ , \ 1 \leq j \leq m \tag{4.6}$$

where $n = 30$, $m = 6$, $x_1 = 0$, $x_2 = 5$, $x_3 = 10$, $x_4 = 20$, $x_5 = 40$ and $x_6 = 80$. The random effects and the additive noise are simulated with Gaussian distributions:

$$\phi_{1i} \sim_{iid} \mathcal{N}(105, 64) \ , \ \phi_{2i} \sim_{iid} \mathcal{N}(12, 36) \ , \ \phi_{3i} \sim_{iid} \mathcal{N}(10, 12.25) \ , \ \varepsilon_{ij} \sim_{iid} \mathcal{N}(0, 4).$$

According to Sheiner et al. [1991] and Walker [1996], this model can be used for the analysis of blood pressure y as a function of the dose d of an anti-hypertensive drug from a longitudinal study.

Walker [1996] compares different popular methods of estimation,such as FOCE (First-Order Conditional Estimation) and LAPLACIAN methods of NONMEM. He computes the means and the standard errors based on 50 simulations for these different estimators. Table 3 reproduces these values, with the estimates and standard errors obtained with the SAEM algorithm as well. We see that the EM algorithm of Walker and the SAEM algorithm give similar results, but it is important here to observe that only 300 iterations of the SAEM algorithm are performed for each single simulated dataset. Computing time for the SAEM algorithm is then very much reduced, in comparison to the Monte-Carlo EM algorithm that requires a sampling of 10,000 random variates at each iteration and converges very slowly. Table 3 also gives the estimation of the standard deviation of the MLE, using the approach proposed in Section 2.3. This method seems to be very accurate since these values are just below the empirical standard deviation computed from the 50 simulations.

Table 3
Pharmacodynamic model: comparison of parameter estimates. The means and the estimated square roots of the MSEs between parentheses, based on 50 simulations.

Parameters	Exact	FOCE		LAP		EM		SAEM		$\hat{\sigma}(\widehat{\boldsymbol{\theta}}^{MLE})$
μ_1	105	105.5	(1.8)	105.3	(1.6)	105.4	(1.7)	104.7	(1.5)	1.4
μ_2	12	12.2	(1.2)	12.4	(1.3)	12.3	(1.3)	11.8	(1.3)	1.0
μ_3	10	9.0	(2.8)	9.7	(2.4)	9.7	(1.6)	10.1	(0.9)	0.6
τ_1^2	64	59.7	(20.9)	58.4	(20.2)	60.0	(20.1)	62.1	(16.6)	14.5
τ_2^2	36	31.5	(11.0)	30.7	(11.9)	30.9	(10.7)	34.4	(10.8)	7.6
τ_3^2	12.25	6.6	(7.6)	13.3	(6.1)	10.1	(2.9)	11.2	(3.0)	2.8

It should also be recalled that the gap observed between the true value of the parameter and the limit of the parameter estimate sequence generated by the algorithm results from two phenomena: first, the convergence of the maximum likelihood estimate toward the true value of the parameter that occurs as the number of individuals n tends to infinity; second, the convergence of the parameter estimate sequence generated by the algorithm toward a local maximum of the observed likelihood that occurs as the number of iterations of the algorithm tends to infinity.

5. Statistical modeling and estimation for Deformable Template Models

In this section, I present my contributions to the analysis of deformable template models, from modeling to parameter estimation.

5.1. *The Bayesian Mixed Effects Template Model*

We considered the hierarchical Bayesian framework for dense deformable templates developed in Allassonnière et al. [2007] . Each image of a given population is assumed to be generated as a noisy and randomly deformed version of a common template drawn from a prior distribution on the set of templates. Individual deformations are *hidden variables* of the model or equivalently random effects in the mixed effects setting, whereas the template and the law of the deformations are parameters (or equivalently fixed effects) of interest.

5.1.1. *The observation model*

We considered gray level images $(y_i)_{1\leq i\leq n}$ observed on a grid of pixels $\{v_u \in D \subset \mathbb{R}^2, u \in \Lambda\}$ that is embedded in a continuous domain $D \subset \mathbb{R}^2$, (typically $D = [-1,1] \times [-1,1]$). Although the images are observed only at the pixels $(v_u)_u$, we are looking for a template image $I_0 : \mathbb{R}^2 \to \mathbb{R}$ defined on the plane (the extension to images on $\mathbb{R}^d$ is straightforward). Each observation y is assumed to be the discretization on a fixed pixel grid of a deformation of the template, plus independent noise. For each observation, there exists an *unobserved* deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that for $u \in \Lambda$:

$$y(u) = I_0(v_u - z(v_u)) + \epsilon(u) \ ,$$

where ϵ denotes an independent additive noise.

Considering the template and the deformations as continuous functions would lead to a dense problem. We use the same framework as chosen in Allassonnière et al. [2007] to describe both the templates I_0 and the deformation fields z. Our model takes two complementary aspects into account: photometric - indexed by p- corresponding to the templates and the noise variances, and geometric -indexed by g- corresponding to the deformations. We choose a representation of both the templates I_0 and the deformations z by finite linear combinations of the kernels centered at some fixed landmark points in the domain D: $(v_{p,j})_{1\leq j\leq k_p}$ (respectively, $(v_{g,j})_{1\leq j\leq k_g}$). They are therefore parameterized by the coefficients $\alpha\in\mathbb{R}^{k_p}$ and $\beta\in(\mathbb{R}^{k_g})^2$ that yield: $\forall v\in D$,

$$\begin{aligned} I_\alpha(v) &\triangleq (\mathbf{K_p}\alpha)(v)\triangleq\sum_{j=1}^{k_p}K_p(v,v_{p,j})\alpha^j\ , \\ z_\beta(v) &\triangleq (\mathbf{K_g}\beta)(v)\triangleq\sum_{j=1}^{k_g}K_g(v,v_{g,j})\beta^j . \end{aligned}$$

For the sake of clarity, we denote the collection of data and their corresponding deformation coefficients by $\mathbf{y}^t=(y_1^t,\dots,y_n^t)$ and $\boldsymbol{\beta}^t=(\beta_1^t,\dots,\beta_n^t)$, respectively. The statistical model of the observations we consider is a generative hierarchical one. We assume conditional normal distributions for $\mathbf{y}$ and $\boldsymbol{\beta}$:

$$\begin{cases} \boldsymbol{\beta}\sim\otimes_{i=1}^n\mathcal{N}_{2k_g}(0,\Gamma_g)\mid\Gamma_g\ , \\ \\ \mathbf{y}\sim\otimes_{i=1}^n\mathcal{N}_{|\Lambda|}(z_{\beta_i}I_\alpha,\sigma^2\mathrm{Id})\mid\boldsymbol{\beta},\alpha,\sigma^2\ , \end{cases} \tag{5.1}$$

where $\otimes$ denotes the product of distributions of independent variables and $zI_\alpha(u)=I_\alpha(v_u-z(v_u))$, for u in Λ denotes the action of the deformation on the template image. The parameters of interest are α which determines the template image; σ^2, the variance of the additive noise; and Γ_g, the covariance matrix of the variables β. We assume that $\theta=(\alpha,\sigma^2,\Gamma_g)$ belongs to an open parameter space Θ:

$$\Theta\triangleq\{\ \theta=(\alpha,\sigma^2,\Gamma_g)\mid\alpha\in\mathbb{R}^{k_p},\ \|\alpha\|<R\ |,\ \sigma>0,\ \Gamma_g\in\mathrm{Sym}_{2k_g}^+\ \}\,,$$

where $\|.\|$ is the Euclidean norm, $\mathrm{Sym}_{2k_g}^+$ is the cone of real positive $2k_g\times2k_g$ definite symmetric matrices, and R is an arbitrary positive constant.

5.1.2. The Bayesian Statistical Model

Even though the parameters are finite dimensional, the maximum likelihood estimator can yield degenerate estimates when the training sample is small. Introducing prior distributions on the parameters regularized the estimation with small samples. The analytical effect of such priors can be seen in the parameter update steps (cf. Allassonnière et al. [2007]). We use a generative model based on standard conjugate prior distributions for parameters $\theta=(\alpha,\sigma^2,\Gamma_g)$ with fixed hyper-parameters. Specifically, we assume a normal prior for α, an inverse-Wishart prior on σ^2 and an inverse-Wishart prior on Γ_g.

Furthermore, all priors are assumed to be independent. This yields $\theta = (\alpha, \sigma^2, \Gamma_g) \sim q_{para} \triangleq \nu_p \otimes \nu_g$ where

$$\begin{cases} \nu_p(d\alpha, d\sigma^2) \propto \exp\left(-\frac{1}{2}(\alpha-\mu_p)^t(\Sigma_p)^{-1}(\alpha-\mu_p)\right)\left(\exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right)\frac{1}{\sqrt{\sigma^2}}\right)^{a_p} d\sigma^2 d\alpha,\ a_p \geq 3\,, \\ \nu_g(d\Gamma_g) \propto \left(\exp(-\langle\Gamma_g^{-1}, \Sigma_g\rangle_F/2)\frac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g,\ a_g \geq 4k_g+1\,. \end{cases} \tag{5.2}$$

For two matrices A and B, we define $\langle A, B\rangle_F \triangleq tr(A^tB)$ as the Frobenius dot product on the set of matrices where tr denotes the trace of the matrix.

5.2. Maximum A Posteriori Estimation

In this Bayesian framework, the parameter estimation will be performed by Maximum A Posteriori (MAP) defined as:

$$\tilde{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(\theta|\mathbf{y})\,,$$

where p denotes the posterior likelihood of the parameters given the observations. The dependence on n refers to the sample size. The existence and consistency (as the number of observed images tends to infinity) has been proven (see Allassonnière et al. [2007]). This contrasts with earlier studies in Glasbey and Mardia [2001] because it uses a penalized likelihood or the more recent maximum description length approach in Marsland et al. [2007] for which consistency cannot be proved because the deformations are considered as *nuisance parameters* to be estimated.

However, the maximization problem of the posterior distribution has no closed form in our case, which prevents a direct computation of $\tilde{\theta}_n$. The EM algorithm, although quite natural for maximizing a likelihood under a hierarchical model with missing variables, is not adapted to the deformable template model. In fact, direct computation is unfortunately not tractable and we have to find a solution to overcome the problematic E step where we have to compute an expectation with respect to the conditional distribution of $\boldsymbol{\beta}$ given $\mathbf{y}$.

A first attempt was proposed in Allassonnière et al. [2007] where this conditional distribution is approximated by a Dirac distribution at its mode. The authors called their algorithm the Fast Approximation with Mode EM (FAM-EM). The results were very interesting, but the authors point out the lack of convergence of the FAM-EM algorithm when the quality of the input images is not good, and, typically, when they are noisy. Note that the FAM-EM algorithm mentioned in that paper corresponds analytically exactly to the EM-Laplace (see Vonesh [1996]). Indeed, the algorithm achieves its limits in several cases. Using the SAEM algorithm would make it necessary to sample the hidden variable from the conditional distribution. This sampling is not possible in this complex model.

We therefore apply the SAEM-MCMC algorithm with truncation on random boundaries proposed in [A3] to approximate the MAP estimator $\tilde{\theta}_n$.

But the setting of deformable template models deals with high-dimensional missing variables. This raises several issues. If we simulate candidates for the hidden variable as a complete vector, it appears that most of the candidates are rejected. This is a typical high dimensional concentration phenomenon: locally, around a current point, the proportion of the space occupied by acceptable moves becomes negligible when the space dimension grows. From a more practical point of view, even if the proposed candidate is drawn with respect to the current prior distribution, it creates a deformation that is very different from the current one and too large for the corresponding deformed template to fit the observations. This yields very few possible moves from the current missing variable value, and the algorithm is stuck in a non-optimal location or converges very slowly. As a consequence, the infinite support of the unobserved variables is not well covered by the simulated variables.

One solution is to update the chain one coordinate at a time, conditional on the others. This corresponds to a Gibbs sampler and leads to more relevant candidates that have a higher chance of being accepted (see Amit [1996]). From an image analysis point of view, this puts stronger conditions on the type of deformations that are produced when proposing a candidate for each coordinate. Knowing the tendency of the movement given by the other coordinates, the candidate will either confirm it or not depending on if this is a suitable movement. It will thus be accepted with a corresponding probability. Even if some coordinates remain unchanged, some others are updated, enabling the algorithm to visit a larger part of the missing variable support. However, as the dimension grows, the computation time becomes very long. We thus adopt another approach: instead of updating only one coordinate, we optimize the proposal distribution in a high dimension. This can be efficiently done by using the AMALA sampler presented in Sections 3.1 and 3.2.

5.3. Estimation in the Bayesian mixture model

In collaboration with Stéphanie Allassonnière, we considered the framework of the multicomponent model introduced in Allassonnière et al. [2007] and extended the estimation method developed for the deformable template model to this multicomponent model [A5].

Let us consider the statistical estimation of the component weights and of the image labels in a multicomponent model given a set of images. In the existing methods, the templates of each component and the label are estimated iteratively (for example in methods like K-means) but the geometry, and related to this the metric used to compute the distances between elements, is still fixed. Moreover, the label, which is not observed, is, as the deformations, considered as a parameter and not as a hidden random variable. These methods do not lead to a statistical coherent framework for the understanding of deformable template estimation and none of these iterative algorithms derived from those approaches have a statistical interpretation in terms of parameter optimisation in a generative model that describes the data.

We therefore considered the statistical framework for dense deformable templates developed in Allassonnière et al. [2007] in the generalized case of a mixture model for multicomponent estimation. Each image of a database is assumed to be generated from a noisy random deformation of a template image picked randomly among a given set of possible templates. All the templates are assumed to be drawn from a common prior distribution on the template image space. To propose a generative model, each deformation and each image label have to be considered as *hidden* variables. The templates, the parameters of the deformation laws and the components weights are the parameters of interest. This generative model allows us to automatically break down the database into components and, at the same time, to estimate the parameters corresponding to each component while increasing the likelihood of the observations.

We consider a mixture of the deformable template models, which allows a fixed number τ_m of components in each training set. This means that the data will be separated into τ_m (at most) different components by the algorithm. For each observation y_i, we consider the pair (β_i, τ_i) of unobserved variables that correspond to the deformation field and to the label of image i, respectively. We denote this below by $\mathbf{y}^t \triangleq (y_1^t, \ldots, y_n^t)$, by $\boldsymbol{\beta}^t \triangleq (\beta_1^t, \ldots, \beta_n^t)$ and by $\boldsymbol{\tau}^t \triangleq (\tau_1, \ldots, \tau_n)$. The generative model is:

$$\begin{cases} \boldsymbol{\tau} \sim \otimes_{i=1}^n \sum_{t=1}^{\tau_m} \rho_t \delta_t \mid (\rho_t)_{1 \le t \le \tau_m} \,, \\ \\ \boldsymbol{\beta} \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_{g,\tau_i}) \mid \ \boldsymbol{\tau}, \ (\Gamma_{g,t})_{1 \le t \le \tau_m} \,, \\ \\ \mathbf{y} \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_{\alpha_{\tau_i}}, \sigma_{\tau_i}^2 Id_{|\Lambda|}) \mid \boldsymbol{\beta}, \ \boldsymbol{\tau}, \ (\alpha_t, \sigma_t^2)_{1 \le t \le \tau_m} \,, \end{cases} \tag{5.3}$$

where $z_\beta I_\alpha(u) = I_\alpha(v_u - z_\beta(v_u))$ is the action of the deformation on the template I_α, for u in Λ, and δ_t is the Dirac function on t. The parameters of interest are the vectors $(\alpha_t)_{1 \le t \le \tau_m}$ coding the templates, the variances $(\sigma_t^2)_{1 \le t \le \tau_m}$ of the additive noises, the covariance matrices $(\Gamma_{g,t})_{1 \le t \le \tau_m}$ of the deformation fields and the component weights $(\rho_t)_{1 \le t \le \tau_m}$. We denote the parameters by $(\theta_t, \rho_t)_{1 \le t \le \tau_m}$ so that θ_t corresponds to the parameters composed of the photometric part (α_t, σ_t^2) and the geometric part $\Gamma_{g,t}$ for component t. We assume that for all $1 \le t \le \tau_m$, the parameter $\theta_t = (\alpha_t, \sigma_t^2, \Gamma_{g,t})$ belongs to the open space Θ defined as $\Theta = \{ \ (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p}, \ |\alpha| < R, \ \sigma > 0, \ \Gamma_g \in \mathrm{Sym}_{2k_g}^+ \ \}$, where R is an arbitrary positive constant and $\mathrm{Sym}_{2k_g}^+$ is the set of strictly positive symmetric matrices. Concerning the weights $(\rho_t)_{1 \le t \le \tau_m}$, we assume that they belong to the set $\varrho = \left\{ (\rho_t)_{1 \le t \le \tau_m} \in]0,1[^{\tau_m} \mid \sum_{t=1}^{\tau_m} \rho_t = 1 \right\}$.

We choose a normal distribution for the unobserved deformation variable because of the background we have in image analysis. In fact, the registration problem is an issue that has been studied in depth over the past two decades. The goal is, given two images, to find the best deformation that will match one image close to the other. Such methods require choosing the kind of deformations that are allowed (smooth, diffeomorphic, etc). These conditions are equivalent, for some of these methods, to choose a covariance matrix that enables to define an inner product between two deformations coded by a vector β (cf. Amit et al. [1989], Miller et al. [2002]). The regularization term of the matching energy in the small deformation framework treated in this paper can be written as: $\beta^t \Gamma_g^{-1} \beta$. This looks like the logarithm of the density of a Gaussian distribution on β with 0 mean and a covariance matrix Γ_g. The link between these two points of view has been given in Allassonnière et al. [2007]; the mode of the posterior distribution equals the solution of a general matching problem. This is why we set such a distribution on the deformation vector β. Moreover, many experiments have been run using a large variety of such a matrix that now gives us a good initial guess for our parameter. This leads us to consider a Bayesian approach with a weakly informative prior.

We use a generative model that includes natural standard conjugate prior distributions with *fixed* hyper-parameters. These distributions are an inverse-Wishart prior on each $\Gamma_{g,t}$ and σ_t^2 and a normal prior on each α_t, for all $1 \leq t \leq \tau_m$. All priors are assumed to be independent.

For the prior law ν_ρ, we choose the Dirichlet distribution, $\mathcal{D}(a_\rho)$, with density:

$$\nu_\rho(\rho) \propto \left(\prod_{t=1}^{\tau_m} \rho_t\right)^{a_\rho}, \text{ with fixed parameter } a_\rho\,.$$

For the sake of simplicity, let us denote this by $N \triangleq 2nk_g$ and by $\mathcal{T} \triangleq \{1, \ldots, \tau_m\}^n$ so that the missing deformation variables take their values in $\mathbb{R}^N$ and the missing labels in $\mathcal{T}$. We also introduce the following notations: $\eta = (\theta, \rho)$ with $\theta = (\theta_t)_{1 \leq t \leq \tau_m}$ and $\rho = (\rho_t)_{1 \leq t \leq \tau_m}$.

In our Bayesian framework, we choose the MAP estimator to estimate the parameters:

$$\tilde{\eta}_n = \underset{\eta}{\operatorname{argmax}}\, p(\eta|\mathbf{y})\,, \tag{5.4}$$

where $p(\eta|\mathbf{y})$ denotes the posterior distribution of η given the observations $\mathbf{y}$.

In practice, to reach this estimator, we maximize this posterior distribution using a Stochastic Approximation EM (SAEM) algorithm coupled with a Monte Carlo Markov Chain (MCMC) method as used for the one component model in [A3]. However in the multicomponent model we show that it cannot be driven numerically. In fact, the direct generalization of the algorithm presented in [A3] turns out to be of no use in practice because of some trapping state problems.

If we consider the full vector $(\boldsymbol{\beta}, \boldsymbol{\tau})$ as a single vector of missing data, we can use the hybrid Gibbs sampler on $\mathbb{R}^N \times \mathcal{T}$. Even if this procedure provides an estimated parameter sequence which would theoretically converge toward the MAP estimator, in practice, as mentioned in Robert [1996], it would take a quite long time to reach its limit because of the trapping state problem: when a small number of observations are assigned to a component, the estimation of the component parameters is hardly concentrated and the probability of changing the label of an image to this component or from this component to another is really small (most of the time under the computer precision).

We can interpret this from an image analysis viewpoint: the first iteration of the algorithm gives a random label to the training set and computes the corresponding maximiser $\eta = (\theta, \rho)$. Then, for each image, according to its current label, it simulates a deformation field which only takes into account the parameters of this given component. Indeed, the simulation of $\boldsymbol{\beta}$ through the Gibbs sampler involves a proposal whose corresponding Markov chain has $q(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{y}, \eta)$ as stationary distribution. Therefore, the deformation tries to match $\mathbf{y}$ to the deformed template of the given component $\boldsymbol{\tau}$. The deformation field tries to get a better connection between the component parameters and the observation, and there is only small probability that the observation given *this* deformation field will be closer to another component. The update of the label $\boldsymbol{\tau}$ is therefore conditional to this deformation which would not leave much chance to switch component.

To overcome the trapping state problem, we will simulate the optimal label, using as many Markov chains in $\boldsymbol{\beta}$ as the number of components so that each component has a corresponding deformation which “computes” its distance to the observation. Then we can simulate the optimal deformation corresponding to that optimal label.

Since we aim to simulate $(\boldsymbol{\beta}, \boldsymbol{\tau})$ through a transition kernel that has $q_{post}(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, \eta)$ as stationary distribution, we simulate $\boldsymbol{\tau}$ with a kernel whose stationary distribution is $q_{post}(\boldsymbol{\tau}|\mathbf{y}, \eta)$ and then $\boldsymbol{\beta}$ through a transition kernel that has $q_{post}(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{y}, \eta)$ as stationary distribution.

For the first step, we need to compute the weights $q(t|y_i, \eta) \propto q(t, y_i|\eta)$ for all $1 \leq t \leq \tau_m$ and all $1 \leq i \leq n$ which cannot be easily reached. However, for any density function f, for any image y_i and for any $1 \leq t \leq \tau_m$, we have

$$q(t, y_i|\eta) = \left(\mathbb{E}_{q_{post}(\beta|y_i,t,\eta)}\left[\frac{f(\beta)}{q(y_i, \beta, t|\eta)}\right]\right)^{-1} . \tag{5.5}$$

Obviously the computation of this expectation w.r.t. the posterior distribution is not tractable either but we can approximate it by a Monte Carlo sum. However, we cannot easily simulate variables through the posterior distribution $q_{post}(\cdot|y_i, t, \eta)$ as well, so we use some realisations of an ergodic Markov chain having $q_{post}(\cdot|y_i, t, \eta)$ as stationary distribution instead of some independent realisations of this distribution.

The solution we propose is the following: suppose we are at the k^{th} iteration of the algorithm and let η be the current parameters. Given any initial deformation field $\xi_0 \in \mathbb{R}^{2k_g}$, we run, for each component t, the hybrid Gibbs sampler $\Pi_{\eta,t}$ on $\mathbb{R}^{2k_g}$ J times so that we get J elements $\xi_{t,i} = (\xi_{t,i}^{(l)})_{1\leq l\leq J}$ of an ergodic homogeneous Markov chain whose stationary distribution is $q(\cdot|y_i, t, \eta)$. Let us denote by $\xi_i = (\xi_{t,i})_{1\leq t\leq\tau_m}$ the matrix of all the auxiliary variables. We then use these elements for the computation of the weights $p_J(t|\xi_i, y_i, \eta)$ through a Monte Carlo sum:

$$p_J(t|\xi_i, y_i, \eta) \propto \left(\frac{1}{J}\sum_{l=1}^{J}\left[\frac{f(\xi_{t,i}^{(l)})}{q(y_i, \xi_{t,i}^{(l)}, t|\eta)}\right]\right)^{-1} , \tag{5.6}$$

where the normalisation is done such that their sum over t equals one, involving the dependence on all the auxiliary variables ξ_i. The ergodic theorem ensures the convergence of our approximation toward the expected value. We then simulate $\boldsymbol{\tau}$ through $\otimes_{i=1}^{n} \sum\limits_{t=1}^{\tau_m} p_J(t|\xi_i, y_i, \eta)\delta_t$.

Concerning the second step, we update $\boldsymbol{\beta}$ by re-running J times the hybrid Gibbs sampler $\Pi_{\eta,\boldsymbol{\tau}}$ on $\mathbb{R}^N$ starting from a random initial point $\boldsymbol{\beta}_0$ in a compact subset of $\mathbb{R}^N$. The size of J will depend on the iteration k of the SAEM algorithm in a sense that will be precised later, thus we now index it by k.

The density function f involved in the Monte Carlo sum above needs to be specified to get the convergence result proved in the last section of this paper. We show that using the prior on the deformation field enables to get the sufficient conditions for convergence. This density is the Gaussian density function and depends on the component we are working with:

$$f_t(\xi) = \frac{1}{\sqrt{2\pi}^{2k_g}\sqrt{|\Gamma_{g,t}|}} \exp\left(-\frac{1}{2}\xi^t\Gamma_{g,t}^{-1}\xi\right) . \tag{5.7}$$

Let (η_k) be the sequence generated by the algorithm 2.2 using the sampler described above. We obtained the following convergence result:

Theorem 5.1. *Under the assumptions of Theorem 4.2 of [A5], we have for all $\boldsymbol{\beta}_0 \in \mathrm{K}$, $\boldsymbol{\tau}_0 \in \mathcal{T}, s_0 \in \mathcal{S}$ and $\eta_0 \in \Theta \times \varrho$,*

$$\lim_{k\to\infty} d(\eta_k, \mathcal{L}) = 0 \;\; \bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0}\text{-}a.s\,,$$

where $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0,\boldsymbol{\tau}_0,s_0,0,}$ is the probability measure associated with the chain $(Z_k = (\boldsymbol{\beta}_k, \boldsymbol{\tau}_k, s_k, \kappa_k))_{k\geq 0}$ starting at $(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0, s_0, 0)$ and $\mathcal{L} \triangleq \{\ \eta \in \hat{\eta}(\mathcal{S}),\ \frac{\partial l}{\partial \eta}(\eta) = 0\}$.

Proof. Even if the only algorithmic difference between our algorithm and the SAEM algorithm is the simulation of the missing data which is not done with respect to the conditional law $q_{post}(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, \eta)$ but through an approximation which can be arbitrarily close, this yields a very different proof. Indeed, whereas for the SAEM algorithm, the stochastic approximation leads to a Robbins-Monro type equation with no residual term r_k, our method induces one. The first difficulty is therefore to prove that this residual term tends to 0 while the number of iterations k tends to infinity. Our proof is decomposed into two part, the first one concerning the deformation variable $\boldsymbol{\beta}$ and the second one the label $\boldsymbol{\tau}$. The first term requires to prove the geometric ergodicity of the Markov chain in $\boldsymbol{\beta}$ generated through our kernel. For this purpose, we prove some typical sufficient conditions which include the existence of a small set for the transition kernel and a drift condition. Then, we use for the second term some concentration inequalities for non stationary Markov chains to prove that the kernel associated with the label distribution converges toward the conditional distribution $q_{post}(\boldsymbol{\tau}|\mathbf{y}, \eta)$.

The second difficulty is to prove the convergence of the excitation term e_k. This can be carried out as in Delyon et al. [1999] using the properties of our Markov chain and some martingale limits properties. □

5.4. Estimation in the Large Diffeomorphic Deformable Mapping Model

In collaboration with Stéphanie Allassonnière and Stanley Durrlemann, we applied the Anisotropic Metropolis Adjusted Langevin Algorithm as a sampler into the SAEM-MCMC algorithm to estimate the parameters in the Large Diffeomorphic Deformable Mapping Model [A13]. We also proposed an extension of the model to optimize the position of the control points as well as a criterion to select the number of control points.

5.4.1. Introduction

Formerly we consider a simple deformation model may be the so-called “linearized deformation”. A linearized deformation ϕ is defined by the displacement field v of each point in the domain $D \subset \mathbb{R}^d$: $\forall r \in D,\ \ \phi(r) = r + v(r)$. The main advantage of this class of deformations is its numerical simplicity as it parameterizes the deformation by a single vector field v. Nevertheless, even with regularity conditions on v, there is no guarantee that the deformation is invertible, meaning that the deformation may create holes or overlapping regions in the domain. To avoid such unrealistic behaviors, diffeomorphic maps that preserve the topology of the shapes in the image set should be considered. This means that we assume that every sample has the same topology or, equivalently, that differences within sample shapes do not rely on changes in topology.

Diffeomorphic deformations can be built on linearized deformations within the framework of the Large Diffeomorphic Deformation Metric Mapping (LDDMM), which was introduced by Trouvé [1998] and Christensen et al. [1996] and further developed by Allassonnière et al. [2005], Beg et al. [2005], Glaunès et al. [2003], Holm et al. [2004], Joshi and Miller [2000], Miller and Younes [2001]. In this framework, the above linearized deformations are considered to be infinitesimal deformations, and the vector field v is seen as an instantaneous velocity field. The composition of such deformations creates a flow of diffeomorphisms, which can be written at the limit as the solution of a differential equation. The set of such diffeomorphisms can be equipped with a group structure and a right-invariant metric, providing regularity on the driving velocity fields. It follows that the set of images is given the structure of an infinite-dimensional manifold, on which distances are computed as the geodesic length in the deformation group between the identity map and the diffeomorphism that maps one image on to another.

It was shown by Durrleman [2010] that this infinite dimensional deformation set can be efficiently approximated by a finite control point parameterization that carries momentum vectors. This finite dimension reduction is a key aspect for statistical analysis. Durrleman et al. [2012] have enforced the velocity fields that are defined everywhere in the domain to be parameterized by a finite set of control points. Positions of control points are not given as a prior but optimized as parameters of the statistical model. As a consequence, control points tend to move to the regions that show the greatest variability among samples while optimizing a least-square criterion. At the same time, this optimization makes it possible to reduce the number of control points for the same matching accuracy, compared to the case where control points are fixed as the nodes of a regular lattice.

Once the deformation model has been fixed, it is necessary to estimate the parameters of the associated statistical model including, in particular, the template image. Different algorithms have been proposed to solve the template estimation. Most of them are based on a deterministic gradient descent. In particular, Durrleman et al. [2012] managed simultaneously the optimization in control point positions and momentum vectors using a joint gradient descent. Although it provided visually interesting results in several practical cases, the nature of the limit is not identified. Moreover, this type of method fails in specific cases, in particular, when using noisy training data. Another point of view is to consider stochastic algorithms, e.g., Zhang et al. [2013] using a Hamiltonian Monte Carlo sampler into a Monte Carlo Expectation Maximization algorithm in the dense LDDMM setting, although there is no theoretical convergence property proved for this algorithm.

We considered now the LDDMM setting where the deformations are parameterized by a finite number of initial control point positions and momenta such as in Durrleman et al. [2012]. To do this, we extend the generative statistical model of Allassonnière et al. [2007]. In that model, the deformations are assumed to be linearized and are modeled as random variables that are not observed. This enables us to estimate the representative parameters of their distribution that will characterize the geometric variability. On the one hand, we extend this approach to the LDDMM framework. On the other, we introduce the control point positions as population parameters into the model so that they can be optimized in the estimation process. This enables us to better fit the deformation model leading to a more accurate estimation of the geometric parameters.

From an algorithmic point of view, we propose to use the Anisotropic Metropolis Adjusted Langevin Algorithm (AMALA) within the SAEM algorithm introduced in [A12]. This algorithm has shown itself to have very interesting theoretical and numerical properties. Indeed, the AMALA sampler enables us to better explore the target distribution support in very high dimensions, compared to other samplers. It also increases the speed of convergence of the estimation algorithm. Moreover, we take advantage in our sampler of the efficient computation used in the joint gradient descent by Durrleman et al. [2012] so that the optimization of control point positions is of no additional cost at each iteration.

Another interesting question is how to optimize the number of control points required to parameterize the deformations. Indeed, the number of control points essentially depends on the variability in the data: it should be estimated rather than fixed by the user. In the geometric approach [Durrleman et al., 2012], control points were automatically selected using a L^1 type penalty that tends to zero out momentum vectors of small magnitude. Numerically it is solved by an adapted gradient descent known as FISTA (see Beck and Teboulle [2009]). However, this penalty acts on each observation separately, meaning that a control point that is needed to match only a single observation will be kept in the final set of control points. From a statistical point of view, this control point can be thought of as an outlier that could preferably be removed from the basis. The L^1 penalty is also not suitable for statistical purposes, since its associated distribution, namely the Laplace prior, does not generate sparse variables. In other words, the criterion with the L^1 penalty that is minimized in Durrleman et al. [2012] could not be interpreted as the log likelihood of a statistical model that generates sparse solutions.

We propose to include a sparsity constraint in the parameter space of our statistical model through a thresholding step, borrowing ideas from the Group LASSO literature initiated in Bach [2008]. This has the advantage to select control points based on their importance for the description of the variability of the whole population, and not only of one single sample. The thresholding step is then included in the Maximization step, so that the same AMALA-SAEM algorithm can be used for the estimation process. We also exhibit a criterion to select an optimal threshold leading to an optimal number of control points.

5.4.2. The LDDMM model

The model of diffeomorphic deformations we choose is derived from the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework (see Dupuis et al. [1998], Miller et al. [2002], Trouvé [1998]), which generalizes the linearized deformation setting that has been used for the statistical estimation of atlases by Allassonnière et al. [2007]. In the linearized deformation setting, the deformation ϕ is given by :

$$\phi(r) = r + v(r), \quad \forall\, r \in D\,, \tag{5.8}$$

with $d = 2$ or 3, and v a vector field on $\mathbb{R}^d$.

To build a diffeomorphic map, we use the linearized deformations given in Equation (5.8) as infinitesimal steps, and consider the corresponding vector field as an instantaneous velocity field. More precisely, we consider time-dependent velocity fields $(v_t)_t$ for a time-parameter t varying in $[0, 1]$. The motion of a point r_0 in the domain of interest D describes a curve $t \to r(t)$ which is the integral curve of the following Ordinary Differential Equation (ODE) called flow equation :

$$\begin{cases} \dfrac{dr(t)}{dt} = v_t(r(t)) \\ r(0) = r_0 \,. \end{cases} \tag{5.9}$$

The deformation ϕ_1 is defined as follows:

$$\forall r_0 \in D, \quad \phi_1(r_0) = r(1) \,.$$

Conditions under which this map ϕ_1 is diffeomorphic can be found in Beg et al. [2005]. In particular, the existence, uniqueness and diffeomorphic property of the solution are satisfied if the velocity v_t belongs to a RKHS at all time t and is square integrable in time.

Under these conditions, the model builds a flow of diffeomorphic deformations $\phi_t : r_0 \longrightarrow r(t)$ for all $t \in [0, 1]$. The flow describes a curve in a sub-group of diffeomorphic deformations starting at the identity map. The RKHS V plays the role of the tangent space of such an infinite-dimensional Riemannian manifold at the identity map Id. We can provide this group of diffeomorphisms with a right-invariant metric, where the square distance between the identity map $Id = \phi_0$ and the final deformation ϕ_1 is given as the total kinetic energy used along the path: $d(Id, \phi_1)^2 = \int_0^1 \|v_t\|_V^2 dt$, where $\| \cdot \|_V$ is the norm in the RKHS. The existence and uniqueness of minimizing paths have been shown by Miller et al. [2002].

According to mechanical principles, one can show that the kinetic energy is preserved along the geodesic paths, namely for all $t \in [0, 1]$ $\|v_t\|_V = \|v_0\|_V$. Moreover, the velocity fields (v_t) along such paths satisfy Hamiltonian equations, meaning that the geodesic is fully parametrized by the initial velocity field v_0. This velocity field plays the role of the Riemannian logarithm of the final diffeomorphism ϕ_1. Therefore, it belongs to a vector space and allows the definition of tangent-space statistics in the spirit of Vaillant et al. [2004] and Pennec [2006].

Following Durrleman et al. [2011] and Durrleman et al. [2012], we further assume that v_0 is the interpolation of momentum vectors $(\alpha_{k,0})_k$ at control point positions $(v_{g,j}{}_{,0})_k$:

$$v_0(r) = \sum_{k=1}^{k_g} \mathbf{K_g}(r, v_{g,j}{}_{,0}) \alpha_{k,0} \,, \tag{5.10}$$

where $\mathbf{K_g}$ is the kernel associated to the RKHS V. In this context, it has been shown in Miller et al. [2006] that the velocity fields $(v_t)_t$ along the geodesic path starting at the identity map in the direction of v_0 keep the same form:

$$v_t(r) = \sum_{k=1}^{k_g} \mathbf{K_g}(r, v_{g,j}(t)) \alpha_k(t) \,, \tag{5.11}$$

where the control point positions $(v_{g,j}(t))_k$ and the momentum vectors $(\alpha_k(t))_k$ satisfy the Hamiltonian equations:

$$\left\{\begin{aligned} \frac{dv_{g,j}(t)}{dt} &= \sum_{l=1}^{k_g} \mathbf{K_g}(v_{g,j}(t), c_l(t))\alpha_l(t) \\ \frac{d\alpha_k(t)}{dt} &= -\left(\sum_{l=1}^{k_g} d_{c_k(t)}(\mathbf{K_g}(v_{g,j}(t), c_l(t))\alpha_l(t))\right)^t \alpha_k(t) \end{aligned}\right. \tag{5.12}$$

with initial conditions $v_{g,j}(0) = c_{0,k}$ and $\alpha_k(0) = \alpha_{0,k}$ for all $1 \leq k \leq k_g$. This is similar to the equations of motion of a set of k_g self-interacting particles, with $\mathbf{K_g}$ modeling the interactions. One can easily verify that the Hamiltonian defined as $H_t = \|v_t\|_V^2 = \sum_{k=1}^{k_g}\sum_{l=1}^{k_g} \alpha_l(t)^t \mathbf{K_g}(c_l(t), v_{g,j}(t))\alpha_k(t)$ is constant in time when control point positions and momentum vectors satisfy the system (5.12).

This model defines a finite dimensional subspace of the group of diffeomorphisms. For a given set of initial control points, the diffeomorphisms are parametrized by the momentum vectors attached to them. For one instance of the initial momentum vectors, one builds the motion of the control points and of the momentum vectors by integrating the Hamiltonian system (5.12). Then, they define a dense velocity field at each time t according to Equation (5.11). Finally, one can find the motion $\phi_t(r_0)$ of any point r_0 in the domain D by integrating the flow equation (5.9). In this framework, the tangent-space representation of the diffeomorphic deformation ϕ_1 is given by the initial velocity field v_0 parametrized by $z = ((c_{0,k}, \alpha_{0,k}))_k$, called the initial state of the particle system. The position $\phi_1(r)$ depends on the parameters $((c_{0,k}, \alpha_{0,k}))_k$ via the integration of two non-linear differential equations in Equation (5.12) and Equation (5.9).

Remark 5.1. *The LDDMM framework formulation involves a coupling on the control point and the momentum evolutions along the geodesic path which is not the case in the linearized deformation setting. This joint equation introduces more constraints reducing the dimension of the solution space. Therefore, the identifiability of the control point positions may be expected in our LDDMM framework. This property would most probably fail in the linearized deformation setting where the momenta and the control points are not coupled.*

5.4.3. Statistical model

As pointed out in Allassonnière et al. [2007], the gradient descent optimization with respect to the template together with the momenta does not necessarily converge if the training set is noisy. To solve this problem, we introduce here a statistical model where we consider the deformations as well as the control point positions as non-observed random variables, in the spirit of the BME template model [A3].

We choose to model our data by a generative hierarchical model. In this model, the distribution of the deformations in the diffeomorphism group is parametrized. In a statistical approach, these parameters are estimated from the data, thus providing a metric in the shape space which is adapted to the data and takes into account the deformation constraints. This is in contrast to geometric approaches that estimate the template using a fixed metric.

More precisely, let I_0 be a template image: $I_0: \mathbb{R}^d \to \mathbb{R}$. We consider an observation, namely an image y, as a noisy discretization on a regular grid Λ of a diffeomorphic deformation of the template image. Let ϕ_1^z be the solution of both the flow equation (5.9) and the Hamiltonian system (5.12) with initial condition $z = ((c_{0,k}, \alpha_{0,k}))_k$. Then, for all $s \in \Lambda$,

$$y(s) = I_0((\phi_1^z)^{-1}(r_s)) + \sigma\epsilon(s), \tag{5.13}$$

where $\sigma\epsilon$ denotes an additive centered Gaussian random noise on the grid Λ with variance σ^2, and r_s is the coordinate of the voxel s in the continuous domain D.

We are provided with n images $\mathbf{y} = (y_i)_{1\le i\le n}$ in a training set. We assume that each of them follows the probabilistic model (5.13) and that they are independent.

We consider the initial state of particles, namely the control point positions and the momentum vectors, as random variables and estimate their probabilistic distributions, restricting ourselves to the case of parametric distributions. We assume that control points live in the template domain D and that they are the same for all observations. By contrast, the momentum vectors attached to them are specific to each observation, as they parametrize the matching of the template with each sample image.

Therefore, we propose the following probabilistic model: we assume that the initial control point positions $\mathbf{c}_0 = (c_{0,k})_{1\le k\le k_g}$ are drawn through a Gaussian distribution with mean $\bar{\mathbf{c}}_\mathbf{0}$ and covariance $a_c Id$ where Id is the identity matrix of dimension dk_g. We define the initial momenta $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_0^i)_{1\le i\le n}$ with $\boldsymbol{\alpha}_0^i = (\alpha_{0,k}^i)_{1\le k\le k_g}$. We assume that the variables $(\boldsymbol{\alpha}_0^i)_{1\le i\le n}$ are independent identically distributed and follow a Gaussian distribution with mean 0 and covariance matrix Γ_g. Note that this covariance matrix depends on the initial control point positions as the momenta are attached to them. Moreover the momenta $\boldsymbol{\alpha}_0$ are assumed to be independent of the control point positions $\mathbf{c}_0$ given Γ_g.

Following the same lines as Allassonnière et al. [2007], we parametrize the template function I_0 as a linear combination of gray level values of fixed voxels $(v_{p,j})_{1\le k\le k_p}$ equidistributed on the domain D. The interpolation kernel is denoted by K_p and the combination weights are denoted by w: $\forall r \in D$,

$$I_0(r) = \sum_{k=1}^{k_p} K_p(r, v_{p,j})\mathrm{w}_k\,. \tag{5.14}$$

The action of a diffeomorphism on this template is the linear combination of the deformed kernel with the same weights: $\forall r \in D$,

$$\mathbf{K}_\mathbf{p}^{\boldsymbol{z}}\mathrm{w}(r) = I_0 \circ (\phi_1^z)^{-1}(r) = \sum_{k=1}^{k_p} K_p\left((\phi_1^z)^{-1}(r), v_{p,j}\right)\mathrm{w}_k\,. \tag{5.15}$$

The parameters of the model are $\theta = (\mathrm{w}, \sigma^2, \Gamma_g, \bar{\mathbf{c}}_\mathbf{0})$ and the random variables $(\boldsymbol{\alpha}_0, \mathbf{c}_0)$ are considered as hidden random variables. As we often deal with small sample size in practice, we restrict our inference to a Bayesian setting. Some of the priors can be informative as the one of Γ_g. Other priors may be non-informative as for the expectation of the control point positions for which no additional information is available. The complete model writes therefore:

$$\begin{cases} \theta = (\mathrm{w}, \sigma^2, \Gamma_g, \bar{\boldsymbol{c}}_{\mathbf{0}}) \sim \nu_p \otimes \nu_g \\ \\ \mathbf{c}_0 \sim \mathcal{N}_{dk_g}(\bar{\boldsymbol{c}}_{\mathbf{0}}, a_c Id) |\ \theta , \\ \\ \boldsymbol{\alpha}_0^i \sim \mathcal{N}_{dk_g}(0, \Gamma_g) |\ \theta,\ \forall 1 \le i \le n\,, \\ \\ y_i \sim \mathcal{N}_{|\Lambda|}(\mathbf{K}_{\mathbf{p}}^{(\mathbf{c_0}, \boldsymbol{\alpha}_{\mathbf{0}}^{\mathbf{i}})} \mathrm{w}, \sigma^2 Id)\ |\ (\mathbf{c}_0, \boldsymbol{\alpha}_0^i),\ \ \theta,\ \forall 1 \le i \le n\,. \end{cases} \tag{5.16}$$

We define the prior distributions as follows:

$$\begin{cases} \nu_g(d\Gamma_g, d\bar{\boldsymbol{c}}_{\mathbf{0}}) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g \rangle_F / 2) \frac{1}{\sqrt{|\det(\Gamma_g)|}} \right)^{a_g} \cdot \exp\big(-\frac{1}{2} (\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c)^t \Sigma_c^{-1} (\bar{\boldsymbol{c}}_{\mathbf{0}} - \mu_c) \big) d\Gamma_g d\bar{\boldsymbol{c}}_{\mathbf{0}}, \\ \nu_p(d\mathrm{w}, d\sigma^2) \propto \exp\big(-\frac{1}{2} \mathrm{w}^t \Sigma_p^{-1} \mathrm{w} \big) \cdot \big(\exp\left(-\frac{\sigma_0^2}{2\sigma^2} \right) \frac{1}{\sqrt{\sigma^2}} \big)^{a_p} d\mathrm{w} d\sigma^2\,, \end{cases}$$

where $\langle ., . \rangle_F$ designs the Frobenius scalar product and the hyper-parameters satisfy $a_g \ge 4k_g + 1$, $\Sigma_g = Id$, $\sigma_0^2 > 0$, $a_p \ge 3$ and Σ_p is derived from the interpolation kernel K_p and the photometric grid $(v_{p,j})_{1 \le k \le k_p}$ (see Allassonnière et al. [2007] for more details). Concerning the hyper-parameters of the control point prior (μ_c, Σ_c), we choose μ_c to be the vector of the equidistributed grid coordinates. The covariance matrix Σ_c is assumed non-informative. All priors are the natural conjugate priors and are assumed independent to ease derivations.

Remark 5.2. *From a modeling point of view, the positions of the control points could have been consider as parameters of our model since they are fixed effects of the whole population as well as the template. However considering control points as parameters does not lead to a model belonging to the exponential family. Thus, we could not benefit from the convergence properties and efficient implementation of the SAEM algorithm for this class of models. Therefore, we model the control point positions as random variables following a Gaussian distribution.*

5.4.4. Parameter estimation

Let us define $\mathbf{y} = (y_1, ..., y_n)$. We consider the Maximum A Posteriori (MAP) estimator denoted by $\hat{\theta}_n$ obtained by maximizing the posterior density of θ conditional to $\mathbf{y}$ as follows :

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\, p(\theta | \mathbf{y}). \tag{5.17}$$

We first show that for any finite sample the maximum a posteriori will lie in the parameter set Θ; this is non-trivial due to the highly non-linear relationship between parameters and observations in the model.

Theorem 5.2 (Existence of the MAP estimator)**.** *For any sample* $\mathbf{y}$*, there exists* $\hat{\theta}_n \in \Theta$ *such that* $q(\hat{\theta}_n|\mathbf{y}) = \sup_{\theta\in\Theta} q(\theta|\mathbf{y})\,.$

We are interested in the consistency properties of the MAP estimator without making strong assumptions on the distribution of the observations $\mathbf{y}$ denoted by P. We seek to prove the convergence of the MAP estimator to the set Θ_* defined by :

$$\Theta_* = \{\ \theta_* \in \Theta \mid E_P(\log q(y|\theta_*)) = \sup_{\theta\in\Theta} E_P(\log q(y|\theta))\}.$$

Theorem 5.3 (Consistency)**.** *Assume that* Θ_* *is non empty. Then, for any compact set* $K \subset \Theta$*, for all* $\varepsilon > 0$*,*

$$\lim_{n\to+\infty} P(\ \delta(\hat{\theta}_n, \Theta_*) \geq \varepsilon\ \wedge \hat{\theta}_n \in K\ ) = 0\,,$$

where δ *is any metric compatible with the usual topology on* Θ*.*

The proof follows the lines of Allassonnière et al. [2007]. Indeed, the observed likelihood of our diffeomorphic BME template model has the same regularity properties and asymptotic behaviors in the parameters as the linearized one.

To compute the MAP, we use the AMALA-SAEM algorithm proposed in [A12]. Indeed in our applications, the missing variables composed of the initial momenta and positions of control points $\mathbf{z} = (\mathbf{c}_0, \boldsymbol{\alpha}_0)$ are of very high dimension.

5.4.5. Extension toward sparse representation of the geometric variability

Obviously, the number of degrees of freedom needed to describe the variability of a given shape should be adapted to this shape. Therefore, the number of control points in our model should be estimated as a parameter of the model and not fixed by the user. This leads to automatically optimize the dimension of the deformation model. We propose here to simultaneously optimize the positions of the control points and select a subset of the most relevant ones for the description of the variability.

In Durrleman et al. [2012], the control point selection is done adding an L^1 penalty on the momenta to the energy E_θ and performing an adapted gradient descent called FISTA (see Beck and Teboulle [2009]). The effect of this penalty is to zero out momenta of small magnitude and to slightly decrease the magnitude of the other ones. A control point which does not contribute to *at least* one of the template-to-observation deformations at the convergence of the algorithm is called inactive. Note that since control points move in the domain, inactive control points may become active during the optimization process, and vice-versa.

This method suffers from three main limitations. First, the Laplace prior associated to the L^1 penalty does not generate sparse observations. Second, the method keeps active control points that may contribute to only few template-to-observation deformations. Lastly, L^1 penalty implies a soft thresholding step on the momentum vectors, thus reducing the norm of these vectors keeping the direction and therefore the local curvature. As a consequence, important momenta for the description of the variability will also be penalized. In the following, we propose to select control points given their importance to describe the variability of the *whole* population, and not of outliers. The idea is to inactivate a control point if the distribution of the momenta attached to it is not strongly correlated with the momentum distribution of other control points. Therefore our procedure selects control point positions and their number, relevant with regards to the whole population.

This constraint on the momenta is taken into account in the model by assuming that the geometric covariance matrix Γ_g is of the form $\Gamma_g = A_g + \varepsilon_g Id$, where ε_g is a small positive real number and A_g is a sparse symmetric positive matrix. We introduced a positive threshold λ which plays an equivalent role as the weight of the L^1 penalty in the criterion optimized in Durrleman et al. [2012]. The larger, the sparser the solution. We do not consider a control point whose contribution to A_g is lower than the threshold in a given sense defined through the parameter update equation [A13]

This modified update is performed at each iteration of the estimation algorithm in the M-step.

To go one step further, we propose to automatically select an optimal threshold λ. We consider a criterion based on two relevant quantities namely the data attachment residual over the training images and the number of active control points. Indeed, the larger the threshold, the larger the residual and the lower the number of active control points. These quantities are computed for different values of the threshold. These sequences are then normalized to 1. The optimal threshold is chosen to be the point where the two normalized sequences intersect.

5.5. Experiments

We apply all the proposed estimation algorithms in the BME template models on different datasets. The first one is the USPS hand-written digit base as used in Allassonnière et al. [2007]. The other two are medical images of 2D corpus callosum [A4] and 3D murine dendrite spine excrescences [A12]. We present in this manuscript only the results obtained for the template estimation. The corresponding results for the geometrical parameter and the residual variance are given in the corresponding papers.

We begin with presenting the experiments on the USPS database. In order to make comparison, we estimate the parameters in the same conditions as in the previous mentioned works that is to say using the same 20 images per digit. Each image has grey level between 0 (background) and 2 (bright white). These images are presented on the left panel of Figure 2. We also use a noisy training dataset generated by adding a standardized independent Gaussian noise. These images are presented on the right panel of Figure 2. We test five algorithms: the deterministic approximation of the EM algorithm (FAM-EM) presented in Allassonnière et al. [2007], four SAEM-MCMC where the sampler is either the MALA, the adaptive MALA proposed in Atchadé [2006], the hybrid Gibbs sampler presented in [A3] and our AMALA algorithm.

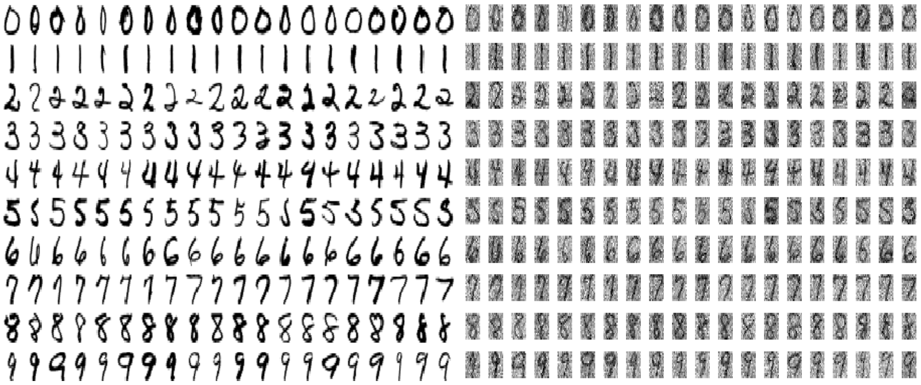


Figure 2. *Left: twenty images per digit of the training set used for the estimation of the model parameters (inverse video). Right: same images with additive noise of variance* 1.

5.5.1. *Computational performances*

The computational time is smaller for the three MCMC-SAEM algorithms using "MALA-like" samplers compared to the FAM. Comparing to the hybrid Gibbs-SAEM, the computational time is 8 times lower with the AMALA-SAEM in this particular case of application. Indeed, the hybrid Gibbs sampler requires no computation of the gradient. However, it includes a loop over the coordinates of the hidden variable, here the deformation vector of size $2k_g = 72$. At each of these iterations, the candidate is straightforward to sample whereas the computational cost lies into the acceptance rate. When this becomes heavy, the less times you calculate it, the better. In the AMALA-SAEM, this acceptance rate only has to be calculated once for each image. Therefore, even when the dimension of the hidden variable increases, this is of constant cost. The main price to pay is the computation of the gradient. Therefore, a tradeoff has to be found between the computation of either one gradient or dk_g acceptance rates in order to select the algorithm to use.

5.5.2. *Results on the template estimation*

All the estimated templates obtained with the five algorithms and noise-free and noisy training data are presented in Figure 3. As noticed in [A3], the FAM-EM estimation is sharp when the training set is noise-free and is deteriorated while adding noise. This behavior is not surprising with regard to the theoretical bound established in Bigot and Charlier [2011] in the particular case of compact deformation group. Considering the adaptive sampler, it does not reach a good estimation of the templates which are still very blurry and noisy in both cases. The problem seems to come from the very low acceptation rate already at the beginning of the estimation. The bad initial guess we have about the covariance matrix of the proposal seems to block the chain. Moreover, the tuning parameters are difficult to calibrate along the iterations of the estimation algorithm. Concerning the estimated templates using the Gibbs, MALA and AMALA samplers, they look very similar to each other using the noise-free data as well as the noisy ones. This similarity confirms the convergence of all these algorithms toward the MAP estimator. In this case, the templates are as expected: noise free and sharp.

Nevertheless, when the dimension of the hidden variable increases, both the Gibbs and the MALA samplers show limitations. We run the estimation on the same noisy USPS database, increasing the number k_g of geometrical control points. We choose the dimension of the deformation vector equal to 72, 128 and 200. The Gibbs-SAEM would produce sharp estimations but explodes the computational time. For this reason, we did not run this algorithm on higher dimension experiments. The results are presented in Figure 4. Concerning the MALA sampler, it does not seem to capture the whole variability of the population in such high dimension. This yields a poorly estimation of the templates. This phenomenon does not appear using our AMALA-SAEM algorithm. The templates still look sharp and the acceptation rates remain reasonable.

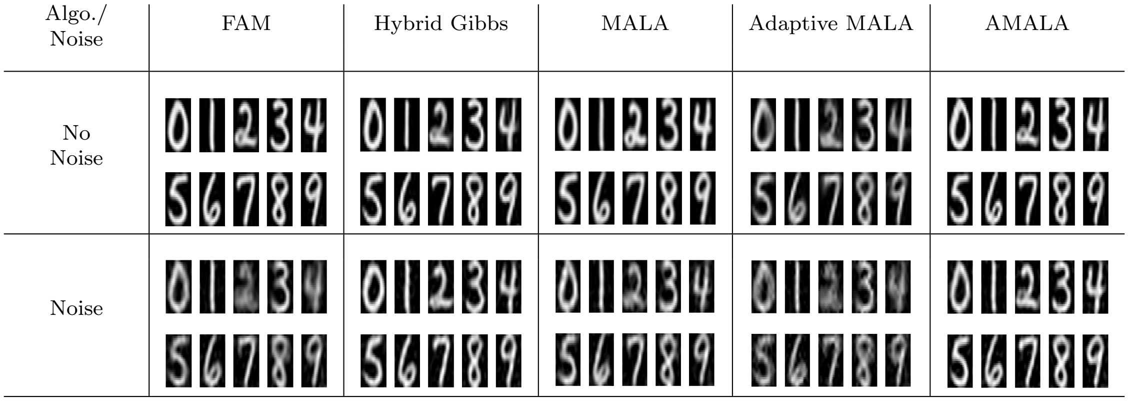


FIGURE 3. *Estimated templates using the five algorithms on noise free and noisy data. The training set includes* 20 *images per digit. The dimension of the hidden variable is* 72.

5.5.3. 2D medical image template estimation

A second database is used to illustrate our algorithm. As before, in order to make comparisons with existing algorithms, we use the same database presented in [A4]. It consists of 47 medical images, each of them is a $2D$ square zone around the end point of the corpus callosum. This box contains a part of this corpus callosum as well as a part of the cerebellum. Ten exemplars are presented in the top rows of Figure 5.

The estimations are compared with these obtained with the FAM-EM and the hybrid Gibbs-SAEM algorithms and with the grey level mean image (bottom row of Figure 5). In this real situation, the Euclidean grey level mean image (a) is very blurry. The estimated template using the FAM-EM (b) provides a first amelioration in particular leading to a sharper corpus callosum. However, the cerebellum still looks blurry in particular when comparing it to the shape which appears in the template estimated using the hybrid Gibbs SAEM (c). The result of our AMALA-SAEM is given in image (d). This template

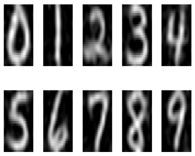
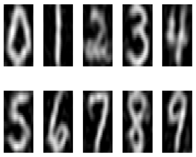
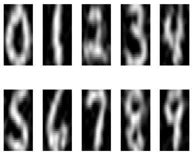
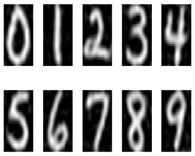
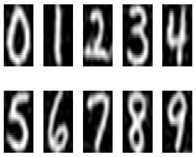
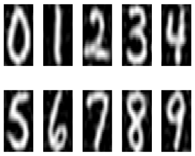
Dim. of deformation / Sampler	$2k_g = 72$	$2k_g = 128$	$2k_g = 200$
MALA			
AMALA			

FIGURE 4. *Estimated templates using MALA and AMALA samplers in the stochastic EM algorithm on noisy training data. The training set includes* 20 *images per digit. The dimension of the hidden variable increases from* 72 *to* 200.

is very close to (c) as we could expect at a convergence point. Nevertheless the AMALA-SAEM has much lower computational time than the hybrid Gibbs-SAEM. This shows the advantage of using AMALA-SAEM in real cases of high dimension.

5.5.4. Experiments in the LDDMM model

We consider the model with random control points presented in Section 5.4.3 as well as its simplified version where the control points are fixed. The number of control points is chosen equal to 4, 9 or 16 depending on the experiments. We infer the atlas of each digit independently using our stochastic estimation algorithm for the two models.

We present the estimated templates obtained with both models and varying number of control points in Figure 6. The first row shows the template images estimated with control points fixed. The second one provides the estimated templates together with the estimated control point positions.

As expected, the contours in the template image become sharper in both cases as the number of control points is increased. Moreover, the number of control points being fixed, the sharpness of the estimated template is improved by allowing the control points to move toward optimized positions. We can also note that the estimated control points are informative as they tend to move toward the contours of the digits, and in particular toward those that correspond to the regions of highest variability among samples. It is particularly noticeable on digits 5 and 6 for example.

Mouse mandible experiment in LDDMM model

We consider a second training set composed of 36 X-ray scans of mouse mandibles. Five of them are presented in Figure 7. The estimated template images resulting from three different experiments are shown in Figure 8. The image on the left shows the template estimated using 260 fixed equidistributed control points. The image on the middle (resp. right) shows the estimated template using 117 (resp. 70)

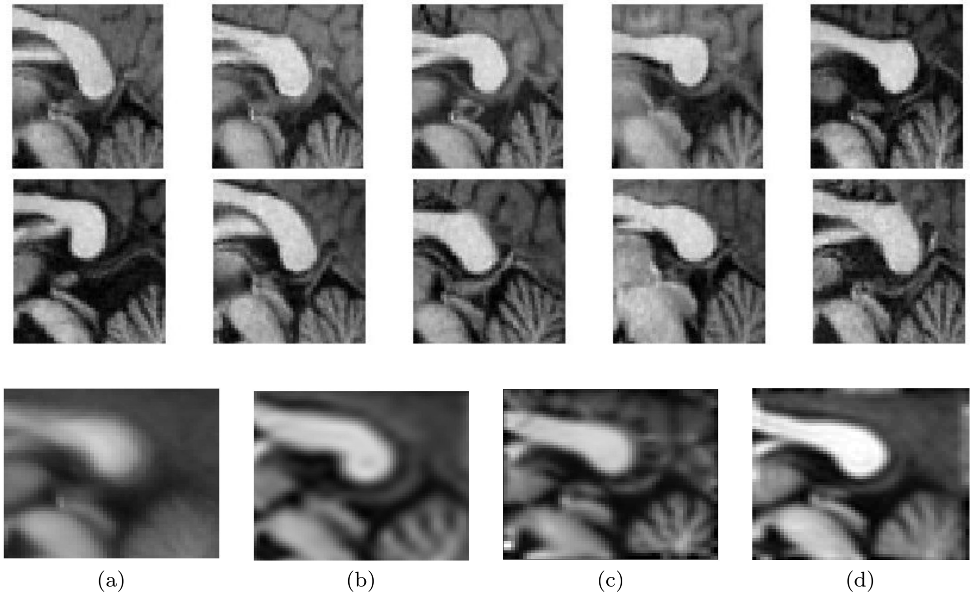

(a) (b) (c) (d)

Figure 5. *Medical image template estimation. Top rows : 10 Corpus callosum and cerebellum training images among the 47 available. Bottom row : (a) mean image. (b) FAM-EM estimated template. (c) Hybrid Gibbs - SAEM estimated template. (d) AMALA-SAEM estimated template.*

estimated control points. These templates look similar, thus showing that the same photometric invariants have been captured in each experiment. These invariants include the main bones of the mandibles (i.e. the brightest areas in the image). The decrease in number of control points is balanced by the optimization of their optimal positions. Control points in the right image are noticeably located on the edges of the shape in order to drive the dilation, contraction and opening of the mandible. Depending on the desired precision of the atlas, we can reduce even more the number of control points. This enables a faster estimation task at the cost of providing less information about the data.

Us Postal dataset experiment in LDDMM model

We conduct different experiments with different thresholds λ between 0.3 and 0.8 in order to see the evolution of the sparsity with respect to this parameter and also to capture the most interesting one (depending on the training digit). The initial number of control points is set to 16. The results of these experiments are presented in Figure 9.

As expected, increasing the threshold λ decreases the final number of selected control points, whose effects on template sharpness and description of variability have been presented in Figure 6. Using the modified parameter update equation to enforce sparsity allows to automatically select a subset of control points leading to estimation results of the same accuracy (see Figure 9). Contrary to the L^1 prior used by Durrleman et al. [2012], our sparsity prior selects a small number of control points without penalizing the magnitude of the momenta. Hence the variability of the model is not under-estimated. In this respect, our thresholding process has an effect which is closer to the expected L^0 norm than its surrogate L^1 norm.

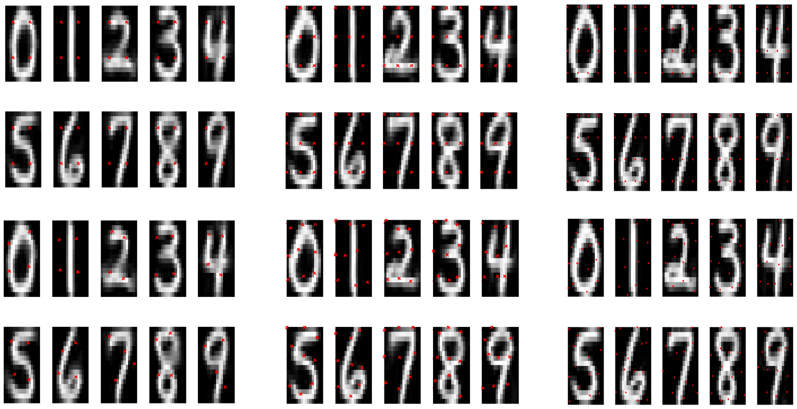


Figure 6. *Estimated templates with varying numbers of control points (Left:* 4*. Middle:* 9*. Right:* 16*). Top: fixed control point model. Bottom: estimated control point model.*

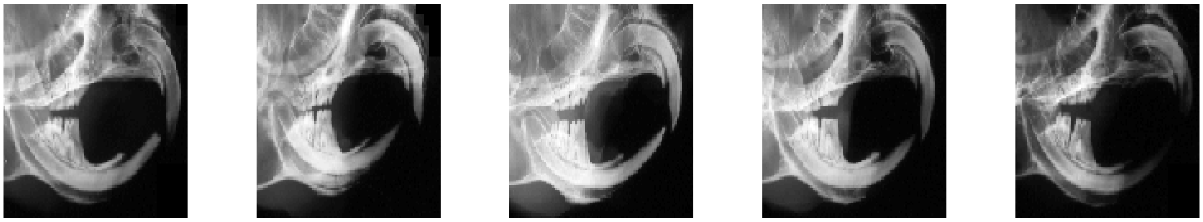


Figure 7. *Five training images from the mouse mandibles.*

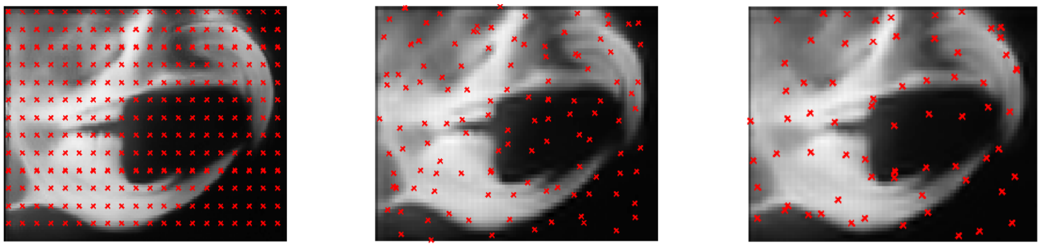


Figure 8. *Estimated templates of the mouse mandible images obtained with* 260 *fixed control points (left), with* 117 *(middle) and* 70 *(right) estimated control points.*

Independently of the threshold λ, control points move in areas where the shape is the most variable. This can be noticed in the loop of the digit 2 which is highly variable, especially in contrast to the loop of the digit 6 which is much more stable in shape across observations. This can be seen by a fastest decrease in number of control points when the threshold λ is increased for the digit 6 compared to digit 2. It is also interesting to notice how our model deals with a mixture of 2 that could be written with or without a loop. Such variability violates the hypothesis of our model, which assumes that observations derive from a *diffeomorphic* deformations of the template image. In this situation, the model estimates a template image that is fuzzy in the region of the loop: the non-diffeomorphic variability has been interpreted as a photometric variation. To overcome this problem, one may investigate the use of several template images in the atlas along the lines of [A5].

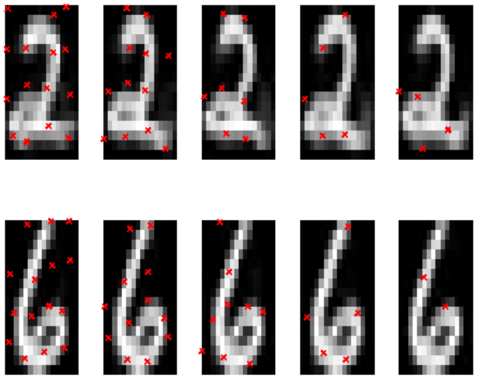


FIGURE 9. *Evolution of the estimated templates and of their number of active control points with respect to the threshold parameter. From left to right:* λ *equals to* $0.3, 0.45, 0.6, 0.75$ *and* 0.8.

The optimal threshold is chosen applying the proposed criterion. Figure 10 shows the estimated templates with their control points corresponding to the optimal threshold. The number of control points reflects the variability of the digits. In particular, very constrained shapes (see digits 1 and 9) require fewer control points than very complex irregular forms (see digits 3 and 8).

5.5.5. Experiments in the mixture model

We considered the multicomponent model decribed in Section 5.3. We ran the SAEM-MCMC algorithm presented in Section 5.3 on the US Postal dataset. In Figure 11, we showed the two estimated templates obtained in the mixture model by the algorithm with 40 training examples per class. It appeared that the two components reached were meaningful, such as the 2 with and without loop or American and European 7.

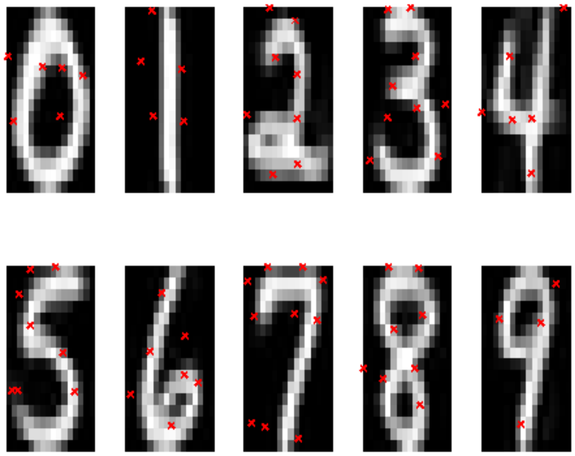


Figure 10. *Estimated templates with their optimal numbers and positions of control points.*

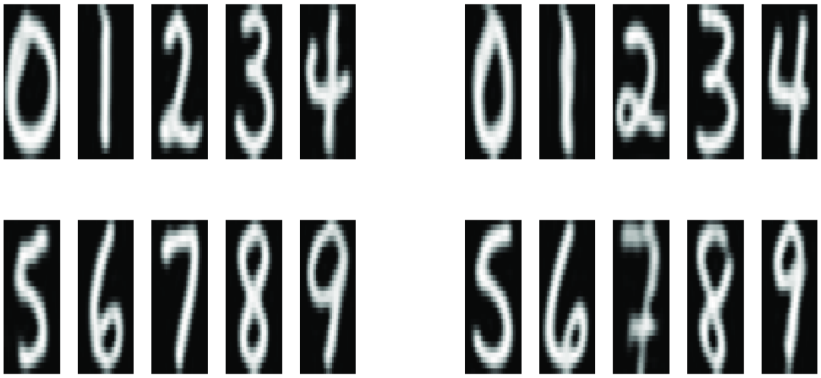


Figure 11. *Estimated prototypes of the two components model for each digit (40 images per class; 100 iterations; two components per class).*

6. Goodness-of-fit test for Gaussian regression with block correlated errors

In collaboration with Sylvie Huet, we proposed a goodness-of-fit test for testing linear hypothesis on the expectation of a Gaussian vector with block correlated errors [A10].

6.1. Introduction

We consider n independent Gaussian vectors $Y_i, i = 1, \dots, n$ with unknown expectation f_i and covariance matrix Σ_i known up to some unknown parameters. The covariance matrix W of the random variable $\mathbf{Y} = (Y_1^T, \dots, Y_n^T)^T$ has a block diagonal structure composed of n squared blocks. This assumption means that the n blocks behave independently but that a correlation exists among the observations within the same block. Many models in various application fields may be covered by this framework, e.g., Gaussian multiresponse regression models with heteroscedastic errors, autocorrelated error models and mixed-effects models.

For the sake of simplicity, we assume that all Y_i have the same size J. Let us denote the expectation of $\mathbf{Y}$ by $\boldsymbol{f} = (f_1^T, \dots, f_n^T)^T$. Our aim is to test the null hypothesis that $\boldsymbol{f}$ belongs to a specific linear subspace V of $\mathbb{R}^{nJ}$ against the alternative that it does not, without making any other assumption about its covariance matrix W than the block diagonal structure detailed above. For example, it would be possible to test the linearity of the relationship between the response and one covariate, or to test that a subset of covariates suffices to model the response.

Checking the adequacy of the model is of particular interest when analyzing data. Several graphical tools are proposed in regression models and, in particular, in linear mixed-effects models (see Diggle et al. [2002], for example). Our interest is in goodness-of-fit testing procedures, also referred to as lack-of-fit or specification tests. Our objective is to propose a procedure without any *a priori* knowledge about $\boldsymbol{f}$, which is adaptive rate-optimal and consistent against local alternatives.

A recent review on goodness-of-fit tests for regression models is provided by Gonzalez-Manteiga and Crujeiras [2013]. Several papers have considered the case where f_i takes the particular form $f_i = (\mu(X_{i,1}), \dots, \mu(X_{i,J}))^T$ for each $i = 1, \dots, n$, for some function μ and for some covariates $(X_{i,j}, j = 1, \dots, J)$, taking values in $\mathbb{R}^r$. Some tests are based on nonparametric function estimation methods since it appears logical to compare a nonparametric estimation of μ to a parametric one computed under the null hypothesis. We refer, for example, to Chen [1994], Eubank and LaRiccia [1993], Härdle and Mammen [1993], Härdle et al. [1998], Hart [1997], Müller [1992], Staniswalis and Severini [1991]. However, these methods present several drawbacks. Some of them require some *a priori* choice of the regularity of μ, and the result of the test depends on this choice. Moreover, a data driven choice of the smoothing parameter involved in the non-parametric estimation, like the bandwidth in a kernel estimator, for example, may affect the level of the test. Another procedure proposed by Dette and Munk [1998] is based on the estimation of the empirical $\mathbb{L}^2$ distance between μ and the null hypothesis, and does not depend on the choice of such a smoothing parameter. Horowitz and Spokoiny [2001] proposed a procedure that rejects the null hypothesis if the distance between the nonparametric kernel estimator and the kernel-smoothed parametric estimator of μ under the null hypothesis is large for some bandwidth within a grid. This test asymptotically achieves the desired level. It is rate-optimal among adaptive procedures over Hölder classes of alternatives and almost achieves the parametric rate of testing for directional alternatives. A similar approach was proposed by Guerre and Lavergne [2005].

An alternative approach is based on the cumulative sums of residuals over covariates or predicted values of the response variable. Due to their construction, such procedures do not depend on any smoothing parameter. They have shown themselves to be of an asymptotic level equal to the nominal level. They are consistent and optimal for testing the null hypothesis versus directional alternatives at the rate of $n^{-1/2}$ (see Diebolt and Zuber [2001], Stute [1997], Su and Wei [1991]).

A different approach proposed by Crainiceanu and Ruppert [2004] for testing a polynomial function versus a parametric alternative consists in testing for the presence of a random effect, the alternative being characterized by splines with random coefficients. This approach requires the choice of the degree and knots of the splines. The test procedure is based on likelihood ratio test statistics and on their exact distributions under the null hypothesis. A similar procedure based on wavelets was proposed by Claeskens et al. [2011].

Baraud et al. [2003] proposed a multiple testing procedure for the expectation of the response variable $\mathbf{Y}$ in the case of a homoscedastic one-dimensional Gaussian regression with an unknown variance parameter. It is based on a large collection of Fisher tests, each of them testing that $\boldsymbol{f}$ belongs to V against a parametric alternative of possibly high dimension. The size of the test is exactly equal to the nominal level. The authors showed that the procedure is simultaneously consistent and rate-optimal over various classes of alternatives.

In the case of mixed-effects models, many papers have dealt with the problem of testing the random effects, but only a few papers have focused on the problem of testing the fixed effects. A review of methods for testing polynomial covariate effects in linear and generalized linear mixed models is given by Huang and Zhang [2008].

In the case of generalized linear mixed-effects models, following the approach of Härdle et al. [1998], Lombardía and Sperlich [2008] proposed a procedure that combines the fully parametric likelihood approach for random effects models with a semiparametric regression profiled likelihood. The test statistic is based on the distance between estimates of the function μ under the null hypothesis and under the alternative. The asymptotic properties under the null hypothesis are established. A parametric bootstrap procedure is proposed for estimating the quantiles of the statistics distribution.

Pan and Lin [2005] generalized the work of Su and Wei [1991] and Stute [1997], and proposed several procedures based on the cumulative sums of residuals. As in the case of homoscedastic regression, their procedures do not depend on the choice of a smoothing parameter. They show that their tests are asymptotically of the nominal level. In the particular case of linear mixed-effects models, they show the consistency of their procedure when the marginal mean is misspecified.

Scheipl et al. [2008] generalized the procedure proposed by Crainiceanu and Ruppert [2004] for linear mixed-effects models. Following the work of Greven et al. [2008], they proposed to compute a pseudo-likelihood by plugging an estimator of the random part of the mixed-effects model into the complete likelihood. To our knowledge, no theoretical property has been established in that context. Recently, Greven and Crainiceanu [2013] established theoretical results for the score test statistic under the null hypothesis, the dimension of the spline basis being allowed to increase with the sample size. Nevertheless, the asymptotic distribution under the null hypothesis is not available in the case of mixed models.

6.2. Model and testing procedure

We consider the Gaussian regression model with block correlated errors. Let $Y_1, \dots, Y_n$ be n Gaussian independent random vectors so that for each $i = 1, \dots, n$, $\mathrm{E}(Y_i) = f_i$ and $\mathrm{Var}(Y_i) = \Sigma_i(\bar{\gamma})$. The vectors $(f_i)_{1 \le i \le n}$ are unknown vectors of $\mathbb{R}^J$, and the matrices $(\Sigma_i)_{1 \le i \le n}$ are symmetric positive matrices that depend on q unknown parameters designated as $\bar{\gamma}$. All the results presented in this paper remain true in the general case where the dimensions J_i are not all equal to J as soon as $\max_i J_i$ is smaller than some constant.

The random vector $\boldsymbol{Y}$ satisfies:

$$\boldsymbol{Y} = \boldsymbol{f} + W^{1/2}(\bar{\gamma})\boldsymbol{\varepsilon}, \tag{6.1}$$

where $\boldsymbol{\varepsilon}$ is a centered, standardized Gaussian vector of $\mathbb{R}^{nJ}$, and $W = W(\bar{\gamma})$ is the block diagonal matrix with $\Sigma_i(\bar{\gamma})$ in block i. Our aim is to test the null hypothesis that $\boldsymbol{f}$ belongs to V, where V is a linear subspace of $\mathbb{R}^{nJ}$ of dimension p. Many regression models in which it is of interest to test goodness-of-fit may satisfy Equation (6.1). Such models have been widely studied in the literature (e.g., Davidian and Giltinan [1995], Demidenko [2004], Diggle et al. [2002], Jones [1993], Pinheiro and Bates [2000], Seber and Wild [1989], Vonesh and Chinchilli [1997]). Let us quote for example linear mixed effects models, hierarchical models, heteroscedastic regression models, regression models with autocorrelated errors, growth curves models.

We propose a goodness-of-fit test to test that the expectation of $\boldsymbol{Y}$ belongs to a specific linear subspace. We generalize the procedure developed by Baraud et al. [2003] to our framework as follows: we define a collection of alternative hypotheses whose cardinality may depend on n, and build on parametric models with low and high dimensions, diversified enough to cover a wide variety of possible alternatives. For each of these alternatives, we consider the likelihood ratio test statistic that would be obtained if the covariance matrix W were known. We then replace the unknown covariance matrix in this statistic by its maximum likelihood estimate. The distribution of each statistic under the null hypothesis is approximated by a χ^2 distribution. Using the Bonferroni adjustment, the null hypothesis is rejected as soon as it is rejected by one of the parametric tests.

The heuristic of our testing procedure can be described in the following way. Let us first assume that the matrix W is known, and consider the test of level $\alpha \in]0,1[$ of the hypothesis $H_0 : \boldsymbol{f} \in V$ against the alternative $H : \boldsymbol{f} \in V + S$ where S is a linear subspace of $\mathbb{R}^{nJ}$ such that the projection of S onto the space orthogonal to $W^{-1/2}V$ in $\mathbb{R}^{nJ}$ is non zero. The likelihood ratio test statistic (see Cox and Hinkley [1974]) denoted by $T_S(W, \boldsymbol{Y})$ is defined as:

$$T_S(W, \boldsymbol{Y}) = \left\| \Pi_S \Pi_{(W^{-1/2}V)^\perp} W^{-1/2} \boldsymbol{Y} \right\|^2 .$$

The hypothesis H_0 is rejected if $T_S(W, \boldsymbol{Y}) > \bar{\chi}_D^{-1}(\alpha)$, where D is the dimension of $\Pi_{(W^{-1/2}V)^\perp} S$ and where $\bar{\chi}_D^{-1}(\alpha)$ designates the $(1-\alpha)$-quantile of a χ^2 distribution with D degrees of freedom.

Since the matrix W is unknown, we propose to plug the test statistic into the maximum likelihood estimator $\widehat{W}$ of W under the alternative hypothesis H.

Following the idea proposed by Baraud et al. [2003] in the linear Gaussian regression model, the multiple testing procedure is constructed as follows. Let us consider a collection of linear subspaces of $\mathbb{R}^{nJ}$ denoted as $\{S_m, m \in \mathcal{M}\}$ where $\mathcal{M}$ is a set of indices that depends on n. We denote the hypothesis $\boldsymbol{f} \in V + S_m$ by H_m and the maximum likelihood estimator of W under H_m by $\widehat{W}_m$. We assume that for each $m \in \mathcal{M}$, the dimension D_m of $\Pi_{(\widehat{W}_m^{-1/2}V)^\perp} S_m$ is greater than 1. Following the Bonferroni procedure, we choose some sequence $\{\alpha_m, m \in \mathcal{M}\}$ of positive numbers satisfying $\sum_{m \in \mathcal{M}} \alpha_m = \alpha$. The statistic for testing that $\boldsymbol{f}$ belongs to V against that it does not is then defined as:

$$T(\alpha) = \sup_{m \in \mathcal{M}} \left\{ T_m(\widehat{W}_m, \boldsymbol{Y}) - \bar{\chi}_{D_m}^{-1}(\alpha_m) \right\}, \tag{6.2}$$

where T_m stands for T_{S_m}. The hypothesis H_0 is rejected when $T(\alpha)$ is positive.

6.3. Theoretical results

We evaluate the properties of the procedure when the sample size n tends to infinity, taking the fact that the cardinality of the collection as well as the dimensions of the alternatives are allowed to grow with the sample size into account. We show that the test is asymptotically of the desired level and that it is consistent over a large class of alternatives. In particular, the rates of testing are the same as those obtained in the homoscedastic case.

6.3.1. Asymptotic level and power

Consider the regression model as defined in Equation (6.1). Let V be a linear subspace of $\mathbb{R}^{nJ}$, and let $\{S_m, m \in \mathcal{M}\}$ be a collection of linear subspaces as defined in Section 6.2. Let $\{\alpha_m, m \in \mathcal{M}\}$ be a sequence of positive numbers such that $\sum_{m \in \mathcal{M}} \alpha_m = \alpha$.

Assuming some regularity assumptions on the model and on the collection of linear subspaces, we have the following theorem:

Theorem 6.1. *Suppose that D_m^4/n tends to 0 and $\sum_m (D_m/n)^{1/3}$ tends to 0. Consider the statistic $T(\alpha)$ as defined in Equation (6.2), then:*

$$\lim_{n \to +\infty} \mathrm{P}_{H_0} \left(T(\alpha) > 0 \right) \leq \alpha,$$

where P_{H_0} denotes the probability when $\boldsymbol{f} \in V$.

Following are some comments about the assumptions that make it possible to prove this theorem. Assuming that D_m^4/n tends to 0, it is possible to show the existence of the parameter estimators defined under H_m. Moreover, if θ_V denotes the p coefficients of $\Pi_V \boldsymbol{f}$ in V, it can be shown that the estimator of the parameters θ_V is $\sqrt{n/D_m}$-consistent. Because the modeling of the covariance matrix W does not depend on S_m, and as a result of the block diagonal structure of W, the estimator of $\bar{\gamma}$ is $\sqrt{n}$-consistent. Finally the statistic $T_m(\widehat{W}_m, \mathbf{Y})$ is asymptotically equal to a χ^2 variable with D_m degrees of freedom up to a remainder term that is of order $D_m/\sqrt{n}$. This result follows from the consistency of the parameter estimators and from the regularity of the density of a χ^2 variable. Assuming that $\sum_m (D_m/n)^{1/3}$ tends to 0, it is possible to control the discrepancies between $T_m(\widehat{W}_m, \mathbf{Y})$ and the χ^2 over all alternatives H_m.

Let us now consider the power of the test. Assuming some regularity assumptions on the model and on the collection of linear subspaces , we have the following result:

Theorem 6.2. *Suppose that for all $m \in \mathcal{M}$, D_m^4/n tends to 0. Let $\mathcal{E}_n$ be the set of $\boldsymbol{f} \in \mathbb{R}^{nJ}$ for which there exists $m \in \mathcal{M}$ such that:*

$$\begin{aligned} \frac{1}{n} \|\boldsymbol{f} - \Pi_{V+S_m} \boldsymbol{f}\|^2 \quad &= \quad O\left(\frac{1}{\sqrt{n}}\right), \qquad (6.3) \\ \frac{1}{n} \|\Pi_{(W^{-1/2}V)^\perp} W^{-1/2} \boldsymbol{f}\|^2 \quad &\geq \quad \frac{1}{n} \|\Pi_{S_m^\perp} \Pi_{(W^{-1/2}V)^\perp} W^{-1/2} \boldsymbol{f}\|^2 + v_{n,m}^2 \ , \end{aligned}$$

$$\text{with } \ v_{n,m}^2 = \frac{\kappa}{n} \left[\sqrt{D_m \log\left(\frac{\log(n)}{\alpha_m}\right)} + \log\left(\frac{\log(n)}{\alpha_m}\right) \right] (1 + o(1)) \, ,$$

for some positive constant κ. Then:

$$\lim_{n \to +\infty} \sup_{\boldsymbol{f} \in \mathcal{E}_n} \mathrm{P}_{\boldsymbol{f}} \left(T(\alpha) \leq 0 \right) = 0,$$

where $\mathrm{P}_{\boldsymbol{f}}$ denotes the probability under Model (6.1).

Following are some comments about the theorem. For each $m \in \mathcal{M}$, under the assumption that D_m^4/n tends to 0, the estimators of the parameters under H_m converge in probability under $\mathrm{P}_{\boldsymbol{f}}$. Their rates of convergence are the same under $\mathrm{P}_{\boldsymbol{f}}$ as under P_{H_0}.

The condition $\|\boldsymbol{f} - \Pi_{V+S_m} \boldsymbol{f}\|^2 = O\left(\sqrt{n}\right)$ results from remainder terms in the asymptotic expansion of $T_m(\widehat{W}_m, \mathbf{Y})$ of the following form: $\|\Pi_{S_m} \Pi_{(W^{-1/2}V)^\perp} (\widehat{W}_m^{-1/2} - W^{-1/2}) \boldsymbol{f}\|^2$. It can be shown that these terms are of the order $\|\boldsymbol{f} - \Pi_{V+S_m} \boldsymbol{f}\|/\sqrt{n}$. If this quantity does not vanish, then we are not able to give an expansion for $T_m(\widehat{W}_m, \mathbf{Y})$ under $\mathrm{P}_{\boldsymbol{f}}$.

Finally, it should be emphasized that the power of our procedure is comparable to the power of the exact procedure proposed by Baraud et al. [2003] in the case of Gaussian linear regression variance with independent and homoscedastic errors.

6.3.2. Detection of local alternatives

The aim of this section is to establish the power of the test for vectors $\boldsymbol{f}$ such that $\|\boldsymbol{f} - \Pi_V \boldsymbol{f}\|$ is of order $\sqrt{\log \log n}$.

We consider a linear space S such that the dimension D of $\Pi_{(W^{-1/2}V)^\perp} S$ is greater than 1, and we denote the maximum likelihood estimator of W under the assumption $''\boldsymbol{f} \in V + S''$ by $\widehat{W}$.

Let us denote the canonical basis of $\mathbb{R}^{nJ}$ by $\{\boldsymbol{e}_{i,a}, a = 1, \ldots, J, i = 1, \ldots, n\}$. The collection $\{S_m, m \in \mathcal{M}\}$ is composed here of all vectors of the canonical basis such that the dimension D_m of $\Pi_{(\widehat{W}^{-1/2}V)^\perp} S_m$ is equal to 1. The collection $\{\alpha_m, m \in \mathcal{M}\}$ is defined as before.

Finally, we consider the test statistic:

$$U(\alpha) = \sup_{m \in \mathcal{M}} \left\{ T_m(\widehat{W}, \boldsymbol{Y}) - \bar{\chi}_1^{-1}(\alpha_m) \right\} \ . \tag{6.4}$$

Assuming some regularity assumptions on the model and on the collection of linear subspaces, we have the following result.

Theorem 6.3. *Suppose that D^4/n tends to 0 and that for all $m \in \mathcal{M}$, $\alpha_m \geq \exp(-n/(\log n)^3)$. If we consider the statistic $U(\alpha)$ as defined in Equation (6.4), then:*

$$\lim_{n \to +\infty} \mathrm{P}_{H_0} \left(U(\alpha) > 0 \right) \leq \alpha,$$

where P_{H_0} denotes the probability when $\boldsymbol{f} \in V$.

For $\boldsymbol{g} \in \mathbb{R}^{nJ}$ such that $\|\boldsymbol{g}\|^2/n$ is bounded, let:

$$\mathcal{A}_n(\boldsymbol{g}) = \left\{ \boldsymbol{f} \in \mathbb{R}^{nJ} \ \textit{such that}\ \boldsymbol{f} = \Pi_V \boldsymbol{f} + \sqrt{\frac{\log\log n}{n}} \boldsymbol{g} \right\}.$$

If $\boldsymbol{f} \in \mathcal{A}_n(\boldsymbol{g})$, and if there exists m_0 such that:

$$\|\Pi_{S_{m_0}} \Pi_{(W^{-1/2}V)^\perp} W^{-1/2} \boldsymbol{g}\|^2 \geq C n \log \left(\frac{1}{\alpha_{m_0}} \right) \ , \tag{6.5}$$

then:

$$\lim_{n \to +\infty} P_{\boldsymbol{f}} \left(U(\alpha) \leq 0 \right) = 0 \ ,$$

where $\mathrm{P}_{\boldsymbol{f}}$ denotes the probability under Model (6.1).

6.3.3. The Bootstrap procedure

We also propose a bootstrap procedure for estimating the quantiles of the distribution of each test statistic and prove that this bootstrap procedure is asymptotically of the desired level.

In the preceding sections, we proposed a test procedure based on the approximation of the quantiles of $T_m(\widehat{W}_m, \mathbf{Y})$ by those of a χ^2 variable with D_m degrees of freedom. An alternative to this procedure is the bootstrap, where the quantiles of $T_m(\widehat{W}_m, \mathbf{Y})$ are approximated by the quantiles of their bootstrap distribution under H_0.

Let $\boldsymbol{\varepsilon}^\star$ be a centered, standardized Gaussian vector of $\mathbb{R}^{nJ}$ independent of $\boldsymbol{\varepsilon}$, and let:

$$\boldsymbol{Y}^\star = \Pi_V \boldsymbol{Y} + \widehat{W}_0^{1/2} \boldsymbol{\varepsilon}^\star, \tag{6.6}$$

where $\widehat{W}_0$ is the maximum likelihood estimator of W under H_0.

Let $T_m^\star = T_m(\widehat{W}_m^\star, \boldsymbol{Y}^\star)$, where $\widehat{W}_m^\star$ is the bootstrap version of $\widehat{W}_m$, and let $q_m^\star(\alpha)$ be the $(1-\alpha)$-quantile of the distribution of $T_m^\star$ conditional on $\boldsymbol{Y}$. We then consider the test statistic defined as:

$$T^\star(\alpha) = \sup_{m \in \mathcal{M}} \left\{ T_m(\widehat{W}_m, \boldsymbol{Y}) - q_m^\star(\alpha_m) \right\}. \tag{6.7}$$

The hypothesis H_0 is rejected when $T^\star(\alpha)$ is positive.

The proof that the asymptotic level of the bootstrap procedure is α is similar to the proof of Theorem 6.1. However, it needs to uniformly control the difference between the distribution function of T_m and its bootstrap version. This leads to an additional assumption about the collection $\mathcal{M}$, precisely, about the cardinality of the subset $\mathcal{M}_1$ defined as the set of $m \in \mathcal{M}$ such that $D_m = 1$.

Theorem 6.4. *Assume that the assumptions of Theorem 6.1 are fulfilled and that $|\mathcal{M}_1|/n^{1/5}$ tends to 0. Let the statistic $T^\star(\alpha)$ be defined as in Equation (6.7), then:*

$$\lim_{n \to +\infty} \mathrm{P}_{H_0}\left(T^\star(\alpha) > 0\right) \leq \alpha.$$

6.3.4. Numerical studies

We conducted a simulation study to to assess the performances of our procedure when n is fixed, and to compare it with the omnibus test proposed by Pan and Lin [2005] based on cumulative residuals. We consider the case of a functional regression model with random effects where we observe (Y_{ij}, X_{ij}) for $i = 1, \ldots n$ and $j = 1, \ldots J$ under the following model:

$$Y_{ij} = \mu(X_{ij}) + b_i + \varepsilon_{ij} \ ,$$

where the variables b_i and ε_{ij} are independent centered Gaussian variables with the variance denoted as σ_b^2 and σ^2, respectively. Our aim is to test that μ is constant. In that case, V is the linear space spanned by the column vector with all components equal to 1.

We consider three types of alternatives, the first with smooth variations, the second with heavy variations and the third with oscillations. We study the level and the power of the test, as well as the effect of the choice of the collection of linear subspaces.

The simulation study shows that the bootstrap procedure is particularly efficient in the case of a small sample size. Our tests give results on the same order as those of Pan and Lin in the case of regular deviations from the null hypothesis and outperform their method in the context of oscillating deviations from the null hypothesis.

We complete our experiments by considering a real dataset of forest coverage in Galicia already treated in Lombardía and Sperlich [2008]. This application highlights, in particular, the advantage of our procedure of being independent of the choice of the smoothing parameter.

6.3.5. Perspectives

It would be interesting to investigate further works in this area. First, one could consider a non linear hypothesis instead of a linear one. The test procedure could be extended possibly using an approximation function basis. Second, one could investigate the case where the size J of each Gaussian vector goes to infinity. This would lead to test functional hypothesis.

Part III

Modeling and Inference in Survival Analysis

In the third part, I included my contributions dealing with modeling and estimation in survival analysis. Related studies include [A6,A7,P6,R1].

In collaboration with Charles El-Nouty, we considered the frailty models introduced by Vaupel et al. [1979] which are an extension of the Cox model that takes the heterogeneity that exists in survival data into account by introducing latent variables. We were interested in these models since we have been in contact with researchers at the Unité de Recherche en Epidémiologie Nutritionnelle (UREN) at the University of Paris 13, who often need to analyze this type of dataset. Very rich from a modeling point of view, these models are very complex from a mathematical point of view, and the estimation task is often very difficult. In numerous cases, the existing algorithms do not provide satisfactory solutions. We proposed the application of the SAEM-MCMC algorithm to frailty models to evaluate the MLE [A6]. We proved that under general regularity assumptions fulfilled by classical frailty models, the SAEM-MCMC algorithm is almost surely convergent toward a local maximum of the observed likelihood. We compared the performances of this algorithm with others that exist in the literature on simulated data and on a real set of bladder cancer data. The numerical results showed a net advantage when using the SAEM-MCMC algorithm, both in terms of the accuracy of the limit as well as the computation time.

In collaboration with Luc Duchateau and Klaartje Goethals (Faculty of Veterinary Medicine, Ghent University), we have analyzed a mastitis epidemic dataset using frailty models with a frailty vector of size four with different covariance structures [R1]. Assessing the correlation structure in cow udder quarter infection times allows us to analyze the propagation risk of the disease as a function of the position of the infection. The parameter estimation that could not be performed in a reasonable time in such models until now, is done using the SAEM-MCMC algorithm. We compared four nested models using likelihood ratio tests. Using simulation studies, we justified *a posteriori* the use of likelihood ratio tests for finite sample size.

7. Some survival models

7.1. Context and notation

Survival analysis consists mainly in the study of event times of interest for different individuals of a given population. For example, this could be failure times such as the death of some patients from a given disease in epidemiology, the breakdown of some systems in reliability, or the appearance of some particular leaves of interest for a plant in agronomy. The analysis of event times attempts to answer many questions such as: which proportion of a population will survive past a certain time? How will the surviving individuals die or fail? Are there particular characteristics that increase or decrease the chance of survival? Are there multiple causes of death or failure?

Let us introduce some general notations useful in the context of survival analysis. We consider a population of N individuals. We denote by T_i the random variable corresponding to the event time of interest, also referred to as lifetime, of the i-th individual of the population for $1 \leq i \leq N$.

A crucial quantity of interest is the survival function of the i-th individual defined for $t > 0$ by $S_i(t) = P(T_i \geq t)$. The most popular estimator of the survival function based on the lifetime data is the Kaplan-Meier estimator (see Kaplan and Meier [1958]) that has been studied by many authors.

A second quantity of interest is the hazard rate λ_i of the i-th individual defined for $t > 0$ by:

$$\lambda_i(t) = \lim_{dt \to 0^+} \frac{P(t \leq T_i < t + dt | T_i \geq t)}{dt},$$

which is the probability for the i-th individual that the event of interest arises at time t given that it has not arisen before t. This quantity can be interpreted in epidemiology, for example, as the risk of dying from a given disease, a constant hazard rate corresponding to a chronic disease.

The survival function can be related to the hazard rate as follows. Let us assume that the random variable T_i has a probability density function denoted by f_i. Then, for $t > 0$, we have:

$$\lambda_i(t) = \lim_{dt \to 0^+} \frac{P(t \leq T_i < t + dt | T_i \geq t)}{dt} = \frac{1}{S_i(t)} \lim_{dt \to 0^+} \frac{S_i(t) - S_i(t + dt)}{dt} = \frac{f_i(t)}{S_i(t)} = -[\log S_i(t)]' ,$$

thus

$$S_i(t) = \exp\left(- \int_0^t \lambda_i(s) ds \right) .$$

Note that a constant hazard rate equal to λ will correspond to an exponential distribution of parameter λ for the survival time T_i.

As mentioned before, it could be interesting to link some particular characteristics of the individuals of the population to their behavior when faced with the event of interest. For example if we consider several treatments for a population of patients, then covariates such as the type of treatment, age and gender of the patient can be taken into account in the model. Some survival models make it possible to include covariates in the analysis of lifetimes. The most popular one is the Cox model (see Cox (1972)).

7.2. The Cox Proportional Hazards Model

The Cox Model, also known as the Proportional Hazards Model, states that covariates act multiplicatively on the hazard rate. Thus, the hazard rate is explained as follows:

$$\lambda_i(t) = \lambda_0(t) \exp(x_i' \beta), \tag{7.1}$$

where X_i are the covariates of individual i, λ_0 is an unknown baseline function, and β is a parameter effect vector. Note that two individuals sharing the same covariates have the same hazard rate. However, the lifetimes of two individuals sharing the same covariates differ due to the hazard of the lifetime distribution.

The main assumption of the Cox model is the so-called risk proportional assumption.

Indeed, as a consequence of Equation (7.1), we have:

$$\log S_i(t) = -\int_0^t \lambda_0(s)\exp(x_i'\beta)ds,$$

leading to:

$$\log(-\log S_i(t)) = x_i'\beta + \log\left(\int_0^t \lambda_0(s)ds\right).$$

Thus, if we consider two groups A and B of individuals with the same covariates within each group, the plot of the function $t \to \log(-\log(S_A(t))$ should be obtained as the vertical translation of the one of $t \to \log(-\log(S_B(t))$. This is illustrated on simulated data in Figure 12. However, the datasets do not often satisfy this assumption. For example, Figure 13 presents the plot of the survival function of real data that do not satisfy this assumption.

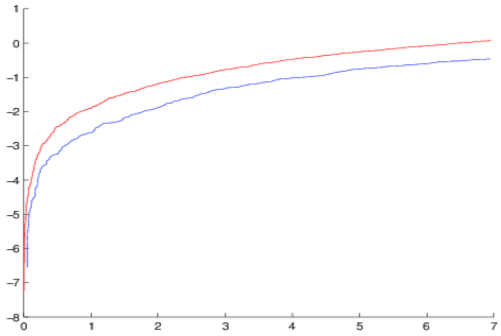


FIGURE 12. *Plots of* $t \to \log(-\log(S(t))$ *in red for group* A *and in blue for group* B.

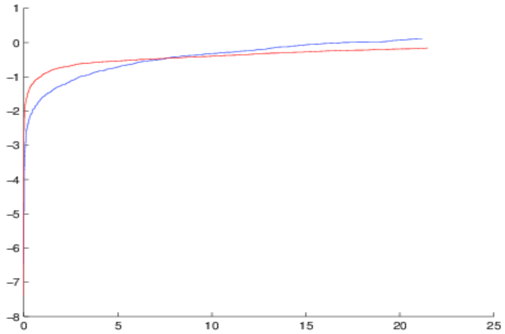


FIGURE 13. *Plots of* $t \to \log(-\log(S(t))$ *in red for the group without treatment and in blue for the group with treatment for a real dataset.*

A classical underlying assumption of the Cox model is that the observations are independent, at least conditionally on covariates. However, this assumption is often not fulfilled by the data because of the lack of homogeneity for the population of interest. For example it can happen that failure times are clustered into groups such as families or geographical areas. Illustrations on medical datasets are given in Aalen and Tretli [1999] where the authors studied the incidence of age on testiculars cancer, or in Gray [1994, 1995] where the author showed the effects of participating institutions in a multicenter lung cancer trial. The same frameworks are developed in economy (see Horowitz [2009]).

To face this lack of fit of the Cox model, one solution is to take the heterogeneity present in the data into account by introducing some random effects in the modeling of the hazard rate. This is the aim of frailty models introduced by Vaupel et al. [1979].

7.3. The Frailty Model

Let us introduce some notations used in the context of frailty models. We consider a population of individuals clustered into N groups. For $1 \leq i \leq N$, we denote by n_i the size of the i-th group. For $1 \leq i \leq N$ and $1 \leq j \leq n_i$, the event time for the subject j of the group i is modeled by a random variable denoted by T_{ij}. We also consider here the censoring time for the subject j of the group i which is modeled by a random variable denoted by C_{ij}. However, the event time (T_i) and the censored time (C_i) are generally not observed. Let us also define the random variables $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\Delta_{ij} = 1_{\{T_{ij} \leq C_{ij}\}}$, where 1_A denotes the indicator function of any set A. We observe the couples $(Y_{ij}, \Delta_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$.

Let b_i be the common frailty random vector for the i-th group for $1 \leq i \leq N$. The initial choice of Vaupel et al. [1979] of the Gamma distribution for the frailty was motivated by its mathematical convenience. It was extended by Clayton and Cuzick [1985] to other frailty distributions. We refer to Hougaard [2000] and to Duchateau and Janssen [2008] for further information on the choice of the frailty law.

Denote by β an unknown parameter effects vector and by λ_0 the unspecified baseline hazard function. Let x_{ij} and z_{ij} be design vectors of covariates associated with failure.

Consider $\lambda_{ij}(t|b_i)$ the conditional hazard function for the j-th individual of the i-th group at time t. The traditional frailty model is defined by:

$$\lambda_{ij}(t|b_i) = \lambda_0(t)\ \exp(x_{ij}^t \beta + z_{ij}^t b_i), \tag{7.2}$$

for $1 \leq i \leq N,\ 1 \leq j \leq n_i$ and $t \geq 0$, where ^{t} denotes the transposition operator.

The classical assumptions on the frailty model are given below:

- **F1 :** The censoring times $(C_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ are independent of the event times $(T_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ and of the frailties $(b_i)_{1 \leq i \leq N}$.
- **F2 :** Conditional to the frailties $(b_i)_{1 \leq i \leq N}$, the event times $(T_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ are independent.
- **F3 :** The frailties $(b_i)_{1 \leq i \leq N}$ are independent and identically distributed with probability density function f_η on $\mathbb{R}^q$, where η is an unknown vector.

We denote the vector of all unknown parameters by $\theta = (\alpha, \beta, \eta)$.

Recall that the frailties are unobserved random variables. The frailty model (7.2) can therefore be seen as a linear mixed effects model for the logarithm of the hazard function. In the terminology of mixed effects models, the frailties $(b_i)_{1\leq i\leq N}$ correspond to random effects and the vector β to fixed effects. Thus, the frailty model can also be viewed as a proportional hazards model with random effects, also referred to as proportional hazards mixed model (see Vaida and Xu [2000]).

7.3.1. Estimation in Frailty Model

We choose here a frequentist approach and consider the MLE for the parameter θ, namely the value $\hat{\theta}_N$ of θ which maximizes the marginal likelihood denoted by L_N^{obs} (see Duchateau and Janssen [2008]).

This quantity is obtained by integrating the complete likelihood L_N over the unobserved frailties, given by:

$$L_N(\mathbf{y}, \boldsymbol{\delta}, \mathbf{b}; \theta) = \prod_{i=1}^{N} f_\eta(b_i) \times \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\lambda_{ij}(y_{ij}|b_i)^{\delta_{ij}} \exp\left(- \int_0^{y_{ij}} \lambda_{ij}(u|b_i) du \right) \right), \tag{7.3}$$

where we denote the vectors corresponding to the realizations (y_{ij}), (δ_{ij}) and (b_i) by $\mathbf{y}, \boldsymbol{\delta}$ and $\mathbf{b}$, respectively.

Asymptotic theoretical properties for the Maximum Likelihood Estimator (MLE) were established by Murphy [1994], Murphy [1995] and Parner [1998]. The authors proved under general conditions that $\hat{\theta}_N$ exists and converges toward θ as N goes to infinity with probability one. Nevertheless it is often not possible in many practical cases to compute the MLE directly. To overcome this crucial difficulty, two main approaches were developed.

The first one consists in applying Cox's idea to the frailty models in order to obtain an approximated marginal likelihood. Examples where this approach is used can be found in McGilchrist and Aisbett [1991] (penalized likelihood), in Nielsen et al. [1992] (partial likelihood), in Therneau and Grambsch [2000] (penalized partial likelihood) and in Rondeau et al. [2003] (penalized full likelihood). This approach has been extended to a Bayesian model introduced by Ducrocq and Casella [1996] and developed by Legrand et al. [2005] and Legrand et al. [2009].

Consider now the second approach. It consists in a numerical approximation of the MLE instead of a direct computation. Since the frailties are not observed, the underlying model belongs to the family of models with hidden variables. Thus, a powerful tool to solve the MLE problem is the EM algorithm proposed by Dempster et al. [1977]. However, in many frailty models, the EM algorithm cannot be directly applied, more specifically, the expectation step (E-step). Thus, several authors suggest approximations of this algorithm. We mention two of them here that are commonly used. Ripatti et al. [2002] applied the Monte Carlo EM (MCEM) algorithm to frailty models. It was introduced by Wei and Tanner [1990], whereas its convergence in a general setting was established by Fort and Moulines [2003]. However, it requires intensive computation time. Cortiñas Abrahantes and Burzykowski [2005] then applied the deterministic EM-Laplace to frailty models. Unfortunately this method requires that the number of observations sharing the same frailty tends to infinity. In practice, this assumption is often not fulfilled, for example, in twin studies. Moreover, this method can induce some bias (see Cortiñas Abrahantes and Burzykowski (2005, pp. 853-859)). A well-known example of the lack of convergence is given in Allassonnière et al. [2007] (pp. 12-13). As a result, it must be used very carefully. As far as we know, there is no theoretical proof of the convergence of the other existing approximations.

8. Estimation in Frailty Model using the SAEM-MCMC Algorithm

8.1. Algorithmic method

In collaboration with Charles El-Nouty,we proposed to use the Stochastic Approximation Expectation Maximization with the Monte Carlo Markov Chain (SAEM-MCMC) algorithm introduced in (A1) for maximum likelihood estimation (MLE) in frailty models (see [A6]). We briefly recall here the characteristics of this method detailed in Section 2. The usual expectation step of the EM algorithm is divided into two new steps: the first one consists in simulating one realization of the non observed frailties, whereas the second one computes a stochastic approximation of the complete log-likelihood by using this simulated value of the frailties. The maximization step follows the same lines as those of the EM algorithm.

Moreover, this algorithm of convenient use has theoretical properties and requires less simulations than the MCEM algorithm. The proposed algorithm also makes it possible to deal with the multivariate frailty models without assumption on the frailty covariance structure. Another advantage of the SAEM-MCMC algorithm is that it generates the estimation of the observed Fisher information matrix at the same time.

8.2. Convergence property of the algorithm

I detail here the assumption required to ensure the convergence of the proposed estimation algorithm, as well as the related convergence result.

Let us introduce two additional assumptions on the frailty model denoted **(F4-F5)**. The first concerns the baseline function λ_0, whereas the second one deals with the frailty distribution f_η.

- **F4 : Regularity of the baseline function.** The function λ_0 belongs to the set of functions defined on $\mathbb{R}^+$ at values in $\mathbb{R}^+$ parameterized by the vector α taking values in an open subset $\mathcal{A}$ of $\mathbb{R}^a$, which are twice continuously differentiable on $\mathcal{A}$.
- **F5 : Exponential family for the frailty distribution.** The probability density function f_η of the frailties belongs to the set of exponential probability density functions where η takes values in an open subset $\mathcal{B}$ of $\mathbb{R}^b$, which are twice continuously differentiable on $\mathcal{B}$.

Extra assumptions denoted by **(H1-H2)** and by **(SAEM1-SAEM3)** in [A6] are made on the regularity of the model and on the stochastic approximation procedure, respectively (see [A6] for more details). Our convergence result is given in the following theorem.

Theorem 8.1. *Assume that* **(F1-F5)***,* **(H1-H2)** *and* **(SAEM1-SAEM3)** *are fulfilled. Let* (θ_k) *be the sequence generated by the SAEM-MCMC algorithm. We then have, with probability 1*

$$\lim_{k\to+\infty} d(\theta_k, \mathcal{L}) = 0,$$

where the distance of x *to the closed subset* A *is denoted by* $d(x, A)$ *and the set of stationary points of* $\log L_N^{obs}$ *by* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \log L_N^{obs}(\mathbf{y}, \boldsymbol{\delta}; \theta) = 0\}$.

Note that under additional regularity assumptions, the almost sure convergence of the sequence $(\theta_k)_k$ toward a local maximum of $\log L_N^{obs}$ is obtained (see [A1]).

The usual choices made for the parametric baseline λ_0 (respectively for the frailty distribution) satisfy the assumption **(F4)** (respectively **(F5)**). We refer to Duchateau and Janssen [2008] for further information about this field. For example, when the frailty follows a Gaussian, a Gamma or a Weibull distribution, the SAEM-MCMC algorithm converges almost surely toward the MLE under weak additional regularity conditions.

8.3. Experiments on a bladder cancer dataset and simulation studies

8.3.1. Experiments a the bladder cancer dataset

Let us briefly describe the bladder cancer dataset (EORTC trials 30781, 30782, 30791, 30831, 30832, 30845 and 30863 Genito-Urinary Tract Cancer Group) that we deal with. It is composed of 2596 eligible patients recruited by 63 medical centers. One hundred patients having missing values and 24 centers have less than 20 patients. In our analysis, we keep 2286 patients and 39 medical centers with more than 20 patients as in Abrahantes et al. [2007] and in Legrand et al. [2005].

The study of the dataset shows that it is a setting with approximately 51% of censuring and that approximately 80% of the individuals follow an intravesical treatment (see Sylvester et al. [2006]).

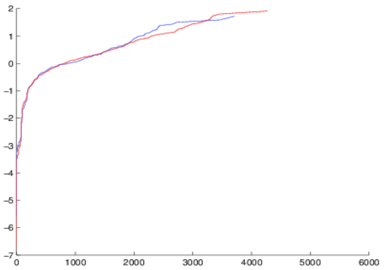


FIGURE 14. *Plot of* $t \to \log(-\log(S(t))$ *in red for the group without treatment and in blue for the group with treatment.*

Figure 8.3.1 shows that the proportional hazards model assumption is not fulfilled by this dataset. We therefore consider a Gaussian frailty model with two frailty terms, referred to as Model I and defined as follows:

$$\lambda_{ij}(t|b_i) = \lambda_0(t)\exp\Big(b_{0i} + x_{ij}^t(\beta + b_{1i})\Big),$$

for $1 \le i \le N,\ 1 \le j \le n_i$ and $t \ge 0$, where $b_i = (b_{0i}, b_{1i})^t$, the random variables $(b_{0i})_{1\le i\le N}$ (respectively, $(b_{1i})_{1\le i\le N}$) are independent and identically distributed Gaussian $\mathcal{N}(0,\sigma_0^2)$ (respectively, $\mathcal{N}(0,\sigma_1^2)$), and the sequences $(b_{0i})_{1\le i\le N}$ and $(b_{1i})_{1\le i\le N}$ are independent. Note that b_{0i} and b_{1i} do not have a symmetric role. Model I, already considered by Abrahantes et al. (2007) and by Legrand et al. [2005], is used to analyze the dataset of bladder cancer and also to carry out our simulation studies in the next section. The event times T_{ij} correspond to times from randomization to the date of the first bladder recurrence, censoring the patients without recurrence at the date of the last available follow-up cystoscopy. The covariate x_{ij} is equal to zero if the j-th patient of the i-th group receives no further intra-vesical treatment and equal to one otherwise. More generally, this covariate indicates whether the patient is classified in the good or poor prognosis group based on the particular prognostic index considered (see Legrand et al. [2009]). The variable b_{0i} can be understood as a random center effect, whereas the variable b_{1i} is viewed as a random treatment per center interaction. Finally, the parameter β is the fixed treatment effect. We refer to Sylvester et al. [2006] for detailed and complete information on the dataset.

Before starting the numerical studies, we have to choose the parametric baseline function λ_0. The SAEM-MCMC algorithm can handle a large class of functions. For example we may choose the Gompertz function $\lambda_0(t) = \lambda\exp(\gamma t)$, $\lambda > 0$, $\gamma \in \mathbb{R}$, or the Weibull function $\lambda_0(t) = \lambda\,\rho\,t^{\rho-1}$, $\lambda > 0$, $\rho > 0$ for $t \ge 0$ (see Duchateau and Janssen (2008, p 29)). Nevertheless, as in Abrahantes et al. [2007], we assume in this section that the baseline function is constant to compare our results with others found in the literature. Hence, the vector of unknown parameters becomes:

$$\theta = (\lambda_0, \beta, \sigma_0^2, \sigma_1^2).$$

We can easily verify that the regularity assumptions **(F1-F5)**, **(H1-H2)** and **(SAEM2)** are fulfilled by Model I. To satisfy the assumption **(SAEM1)**, we choose the sequence $(\gamma_k)_k$ so that if $1 \le k \le 50$, then $\gamma_k = 1$; otherwise: $\gamma_k = \frac{1}{(k-50)^{2/3}}$. Finally, we choose the transition probability Π_θ as a hybrid Gibbs sampler, also known as the Metropolis Hastings within Gibbs algorithm, in order to verify assumption **(SAEM3)**. The proposal distribution is chosen to be equal to the distribution of the frailty terms.

We apply three algorithms to estimate the parameters $\theta = (\lambda_0, \beta, \sigma_0^2, \sigma_1^2)$. We emphasize that we implement the MCEM algorithm (respectively, the EM-Laplace algorithm) according to the method given in Ripatti et al. [2002] (respectively, in Cortiñas Abrahantes and Burzykowski (2005)).

Different random initializations were tested. To investigate the nature of the limit points, we compute an estimate of the logarithm of the marginal likelihood for each limit point using a Monte Carlo sum. We present the relevant numerical results in Table 4 corresponding to the trajectory that gives the estimate of the logarithm of the marginal likelihood with the biggest value.

Let us first comment on the behavior of each algorithm, on the one hand, with respect to the convergence of the trajectories and, on the other, with respect to the computing time.

Consider the convergence problem first. When we apply the SAEM-MCMC algorithm or the MCEM algorithm, all the trajectories converge. The situation is totally different with the EM-Laplace algorithm.

Table 4

Model I. Estimation of the parameters $\theta = (\lambda_0, \beta, \sigma_0^2, \sigma_1^2)$, of the mean model-based standard error in parentheses and of the marginal log-likelihood for the bladder cancer data

Algorithm / Estimate	$\lambda_0(\times 10^{-4})$	β	σ_0^2	σ_1^2	$\log(L_N^{obs})$
EM-Laplace	10.563(0.6487)	−0.0167(0.0704)	0.4074(0.0923)	0.0032(0.0007)	−9490.47
MCEM	7.261(0.6441)	−0.2178(0.1224)	0.0915(0.0080)	0.1783(0.0156)	−9366.59
SAEM-MCMC	7.265(0.7186)	−0.2313(0.0880)	0.0840(0.0292)	0.1849(0.0637)	−9365.41

To obtain a satisfactory Laplace approximation, we have to choose a very sharp covering (or tolerance) of order 10^{-6} to find a global minimum and to assume regularity conditions on the conditional distribution of the frailty variables. Otherwise, many trajectories do not converge. Moreover, the nature of the trajectory is deeply affected by small variations of the initial values. This phenomenon is not really surprising and is also highlighted in [A3] where the EM-Laplace algorithm is termed as FAM-EM algorithm.

Consider now the computing time problem. As soon as the EM-Laplace converges, it requires the same computation time as the SAEM-MCMC algorithm. However, when the EM-Laplace does not converge, the computation time can increase considerably. Finally, the use of the MCEM algorithm requires a computation time ten times greater than the SAEM-MCMC, taking the number of Monte Carlo simulations equal to 5 during the first 50 iterations and equal to 1000 thereafter. This is the main difficulty to apply the MCEM algorithm and therefore justifies the use of the SAEM-MCMC algorithm.

Let us now comment on the numerical results.

As expected, the two stochastic algorithms give same order estimates of parameters and of the marginal log-likelihood, whereas those obtained by using the deterministic algorithm are very different. The EM-Laplace algorithm leads to an estimation of the marginal log-likelihood by -9490.47, which is strictly lower than the estimations of the marginal log-likelihood obtained with the MCEM algorithm and with the SAEM-MCMC algorithm, -9366.59 and -9365.41, respectively. Note that these two last values are local maxima of the marginal loglikelihood. The estimates of the parameters $(10.563 \times 10^{-4}, -0.0167, 0.4074, 0.0032)$ given by the EM-Laplace algorithm do not maximize the marginal log-likelihood and are consequently meaningless. The estimated standard errors given by the two stochastic algorithms have the same order for λ_0 and β, whereas there is a small gain for σ_0^2 and σ_1^2 when using the MCEM algorithm.

Some studies for the same real dataset suggest that for any $i = 1, \ldots, 39$, the center effect b_{0i} could be correlated with the treatment per center interaction b_{1i}. For this purpose, we consider a second model, referred to as Model II and defined as follows:

$$\lambda_{ij}(t|b_i) = \lambda_0(t) \exp\Big(b_{0i} + x_{ij}^t(\beta + b_{1i})\Big),$$

for $1 \leq i \leq N$, $1 \leq j \leq n_i$ and $t \geq 0$, where the N random vectors $(b_i^t)_{1 \leq i \leq N}$ are independent and

TABLE 5

Models I and II. Estimation of the parameters, of the mean model-based standard error in parentheses and of the marginal loglikelihood for the bladder cancer data

Model / Estimate	$\lambda_0(\times 10^{-4})$	β	σ_0^2	σ_1^2	σ_{01}	$\log(L_N^{obs})$
I	7.265(0.7186)	−0.2313(0.0880)	0.0840(0.0292)	0.1849(0.0637)	× × ××	−9365.41
II	7.243(0.4869)	−0.2540(0.0699)	0.0306(0.0002)	0.1077(0.0006)	0.0573(0.000003)	−9357.51

identically distributed multivariate Gaussian:

$$b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right)$$

We insist on the fact that we introduce a dependence structure on the two frailty terms. To study Model II, we limit ourselves to the SAEM-MCMC algorithm. In Table 5 we recall the results obtained for Model I in Table 4 and give the new numerical results for Model II. Note that the parameter σ_{01} only exists in Model II.

We would like to make some comments on these results at this point. Models I and II have the same order estimates for λ_0 and β, whereas the estimates of σ_0^2 and σ_1^2 are lower when considering Model II. This is not really surprising because there is a covariance term σ_{01}. Concerning standard error terms, there is a big gain when considering Model II, particularly for the parameters σ_0^2 and σ_1^2.

Based on the numerical results obtained for Model II, it is natural to determine if the covariance σ_{01} is significantly different from zero. Since Model I and Model II are nested, we do a likelihood ratio test where the null hypothesis is "$\sigma_{01} = 0$". The numerical value of the test statistic is $-2\left(-9365.41 + 9357.51\right) = 15.80$. Let Z be a $\chi_2(1)$ random variable. Since $15.80 > 6.63$ where $P(Z > 6.63) = 0.01$, we reject the null hypothesis at level 1%. Finally, we recommend the use of Model II to analyze this dataset.

8.3.2. Simulation studies

We conduct a simulation study to highlight the performance of our method regarding the other ones that exist in the literature in the same setting as the analysis of the bladder cancer dataset carried out above.

We consider two different settings of censoring namely a moderate one and a heavy one. The design vectors x_{ij} are equal to 1 for 70% of the individuals and 0 for 30% of the individuals in the moderate censoring setting with 40% of censoring, and equal to 1 for 50% of the individuals and 0 for 50% of the individuals in the heavy censoring setting with 60% of censoring. Although the parameter λ_0 could be estimated, we set $\lambda_0 = 0.077$ in order to keep the same choice as the one made in Abrahantes et al. [2007]. We generate $R = 250$ datasets for each setting. Let us denote by $\theta^{(r)} = (0.077, \beta^{(r)}, \sigma_0^{2(r)}, \sigma_1^{2(r)})$ the estimates obtained on the r-th simulated dataset for $1 \leq r \leq R$. Hence, the estimate $\hat{\theta}$ is the empirical mean of the $(\theta^{(r)})_{1 \leq r \leq R}$. The standard error of $\hat{\theta}$ is computed by two methods. The first one is the empirical standard error $\hat{\sigma}(\hat{\theta})$ of this estimate whereas the second one is the so-called mean model-based

Table 6
Model I in a moderate setting with true value $\theta = (0.077, 0.7, 0.4, 0.8)$ (where $\lambda_0 = 0.077$ is known). Estimation with the three algorithms of the parameters and of standard errors in parentheses (first number: mean model based; second number: empirical)

Algorithm / Estimate	β	σ_0^2	σ_1^2
EM-Laplace	0.849(0.031/0.296)	0.206(0.067/0.215)	0.765(0.204/0.481)
MCEM	0.765(0.179/0.151)	0.366(0.195/0.116)	0.778(0.289/0.195)
SAEM-MCMC	0.729(0.072/0.122)	0.392(0.128/0.098)	0.768(0.215/0.133)

Table 7
Model I in a heavy setting with true value $\theta = (0.077, -0.182, 0.4, 0.8)$ (where $\lambda_0 = 0.077$ is known). Estimation with the three algorithms of the parameters and of standard errors in parentheses (first number: mean model-based; second number: empirical)

Algorithm / Estimate	β	σ_0^2	σ_1^2
EM-Laplace	-0.155(0.129/0.156)	0.385(0.085/0.094)	0.766(0.165/0.200)
MCEM	-0.167(0.189/0.174)	0.391(0.199/0.113)	0.782(0.270/0.228)
SAEM-MCMC	-0.178(0.132/0.152)	0.392(0.120/0.109)	0.785(0.239/0.192)

standard error that corresponds to the square root of the empirical mean of the variance estimate given by the inversion of the Fisher Information Matrix.

Numerical results are given in Tables 6 and 7. For each algorithm and each parameter, we present three numbers: the first one corresponds to the estimation of the parameter, whereas the two last numbers correspond to the estimation of the standard error, computed by the two methods described above.

In both censoring settings the SAEM-MCMC algorithm provides a better numerical approximation of the true parameter value than the EM-Laplace and the MCEM algorithms in terms of bias and empirical standard error estimates. However, although the EM-Laplace algorithm often leads to smaller mean model-based standard errors estimates, it also leads to bigger differences between the empirical and the mean model-based estimates of the standard errors. Note also that we obtain good approximations using the MCEM algorithm. However, to reach the same accuracy as in the results obtained by the SAEM-MCMC algorithm, we have to consider large sample sizes in the Monte Carlo approximation that lead to longer computation times.

9. Assessing the correlation structure in cow udder quarter infection times through extensions of the correlated frailty model

In collaboration with Charles El-Nouty (LAGA, University Paris 13 Sorbonne Paris Cité), Luc Duchateau and Klaartje Goethals (Faculty of Veterinary Medicine, Ghent University), we propose to assess the correlation structure in cow udder quarter infection times through several correlated frailty models [R1]. The main contribution of this work is to consider frailty models with an unknown correlation structures on the frailty vector of size 4 which describe precisely the random effects of each of the 4 udder quarters. The usually correlated frailty models were only handling frailty vectors of size 2, because the parameter estimation could not be carry out efficiently. By using the convergent stochastic estimation algorithm presented in Section 8, the estimation task can now be achieve precisely in a reasonable time. We consider different possible correlation structures between all the frailty terms and analyze the performance of each corresponding model. We also compare these nested correlated frailty models by using the likelihood ratio test. We evaluate the performance of the estimation algorithm and of the finite sample size property of the likelihood ratio test on simulated data in this setting.

The existing correlated frailty model methodology is extended to clusters of size four and two different correlation structures. Wienke (2011) gives an excellent overview of the correlated frailty model. Most of the research on the correlated frailty model has been done for bivariate survival data, i.e., clusters of size two. Parameter estimates are in most cases obtained by rewriting the correlated frailty model in copula form, next estimate the marginal survival functions, and finally estimating the association parameter(s) after imputation of the marginal survival functions in the copula form. The cluster size of four allows us to investigate many more different correlation structures, and compare them to choose the most likely correlation structure generating the data , using the log likelihood ratio test. Furthermore, the used methodology, the SAEM-MCMC algorithm, is much more linked to the frailty model framework, and does not require the transformation to the copula format.

9.1. Mastitis dataset: times to infection in four clustered udder quarters

Mastitis or udder infection is economically the most important dairy cow disease in the western world (Seegers et al., 2003). Infections occur at udder quarter level, as the four udder quarters are fully separated, and are infected individually. From a point of view of controlling the disease in a herd, it is essential to know whether udder quarters have a higher risk of infection when one of the udder quarters is infected. With one of the udder quarters infected, it would be helpful to know as well whether each of the three other udder quarters have the same possibly increased risk, or whether the risk of infection differs according to the location of the udder quarter relative to the infected udder quarter.

In order to study such relationships between the four udder quarters, the individual udder quarters were followed up for one lactation period for infection, which generated clustered censored infection time data (Laevens et al., 1997). Obviously, the four udder quarters are clustered within a cow. Such multivariate data can be modeled in different ways to cope with the clustering structure (Duchateau and Janssen, 2008), but in our modeling approach it is important to capture and quantify the association between the infection times of the udder quarters within a cow. The copula model and the frailty model are two possible approaches.

The mastitis dataset consists in 1196 cows. The udder quarters of each cow are followed up individually for time to infection in a lactation period. The udder quarter in which no infection occured are right censored at the time of the end of the lactation period. Two different covariates are introduced in the frailty model. Parity is considered at two levels, primiparous and multiparous, and is a cow characteristic. Location is either front or rear, and changes obviously from one quarter to another within a cow.

9.2. Several correlation structures for the frailty model

We consider general frailty models as defined in Section 7.3 whose conditional hazard rate satisfayes Equation (7.2) to analyze this dataset.

Our purpose is to focus on the effect of the relative location of the infected udder quarter on the propagation of the infection. Thus we are interested in frailty models having a frailty vector of size 4 having a specific correlation structure on its components.

Let us denote the 4-dimensional frailty random vector by $b_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})'$ for the i-th cow for $1 \leq i \leq N$ and the vector of all frailty terms by $\mathbf{b} = (b_i)_{1\leq i\leq N}$. We make the classical independence assumptions for the frailty models: the censoring times $(C_{ij})_{1\leq i\leq N, 1\leq j\leq 4}$ are independent of the event times $(T_{ij})_{1\leq i\leq N, 1\leq j\leq 4}$ and of the frailties $(b_i)_{1\leq i\leq N}$; conditionally to the frailties $(b_i)_{1\leq i\leq N}$, the event times $(T_{ij})_{1\leq i\leq N, 1\leq j\leq 4}$ are independent.

Denote by $\lambda_0(t)$ the unspecified baseline hazard function at time t. In our setting the constant hazard rate is not realistic for describing the time to udder quarter infection, as it is mentioned in Goethals et al. [2009]. Therefore we assume that the baseline function has a Weibull parametric expression given by $\lambda_0(t) = \lambda_0 \gamma t^{\gamma-1}$ for $\lambda_0 > 0$ and $\gamma > 0$.

Denote by x_{ij} the 2-dimensional design vector of covariates associated with failure. This corresponds to the primiparous or multiparous status of the cow and the location of the udder quarter, namely front or rear. Let us introduce $\beta = (\beta_1, \beta_2)$ an unknown 2-dimensional vector corresponding to the two covariate effects. Note that the value of one covariate (location) changes within a cow whereas the other one (parity) changes between cows, i.e. the four udder quarters of a cow have the same parity.

Let us denote the conditional hazard function by $\lambda_{ij}(t|b_i)$ for the j-th individual of the i-th group at time t and the transposition operator by ^{t}. The frailty model $\mathcal{M}_1$, for $1 \leq i \leq N,\ 1 \leq j \leq 4$ and $t \geq 0$ is given by:

$$\lambda_{ij}(t|b_i) = \lambda_0 \gamma t^{\gamma-1}\ \exp(x_{ij}^t \beta + b_{ij}), \tag{9.1}$$

where the frailties $(b_i)_{1\leq i\leq N}$ are independent and identically multivariate Gaussian distributed $\mathcal{N}(0, \Sigma)$ with positive definite covariance matrix Σ given by:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_3 & \rho_2 \\ \rho_2 & \rho_3 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

with $\sigma^2 \geq 0$ and $(\rho_1, \rho_2, \rho_3) \in [-1, 1]^3$. This leads to the following explicit restrictions on the parameters:

$$\left\{ \begin{array}{ccc} |\rho_2 - \rho_3| & < & 1 - \rho_1 \\ |\rho_2 + \rho_3| & < & 1 + \rho_1 \\ 0 & < & \sigma^2 \end{array} \right. .$$

We insist on the fact that the rank of Σ equals 4. This ensures that the Gaussian law of the frailty random vector is non degenerate, i.e., its support is equal to the whole space $\mathbb{R}^4$.

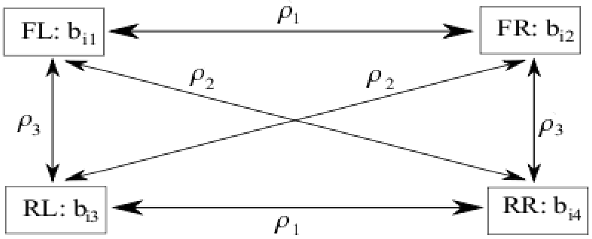


FIGURE 15. *The correlation structure between the random effects of the four udder quarters for model $\mathcal{M}_1$ with 'FL' front left, 'FR' front right, 'RL' rear left and 'RR' rear right udder quarter.*

Let us denote the vector of all unknown parameters of the frailty model by

$$\theta = (\lambda_0, \gamma, \beta_1, \beta_2, \sigma^2, \rho_1, \rho_2, \rho_3),$$

taking values in the open subset Θ of $\mathbb{R}^8$ defined by:

$$\Theta = \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^2 \times \mathbb{R}^{+*} \times [-1,1]^3.$$

Figure 15 shows the correlation structure between the random effects of the udder quarters corresponding to the covariance matrix Σ defined in the model $\mathcal{M}_1$. The correlation between the random effects of the front left and right quarters and the rear left and right quarters is denoted by ρ_1. The correlation between the random effects of the front left and rear right quarters and the front right and rear left quarters is denoted by ρ_2. The correlation between the random effects of the left front and rear quarters and the right front and rear quarters of the front left udder quarter is denoted by ρ_3.

We focus now on three particular models derived from the correlated frailty model $\mathcal{M}_1$, by considering more and more specific correlation structures imposed on the frailty random vector. Each of them corresponds to a practical situation for the interactions between udder quarters.

We define the sub-model $\mathcal{M}_2$ of model $\mathcal{M}_1$ by putting the correlation ρ_3 equal to the correlation ρ_2. The matrix Σ being positive definite, it follows:

$$\left\{ \begin{array}{ccc} |\rho_1| & < & 1 \\ |\rho_2| & < & \frac{1+\rho_1}{2} \\ 0 & < & \sigma^2 \end{array} \right. .$$

We define the sub-model $\mathcal{M}_3$ of model $\mathcal{M}_1$ by putting the correlations ρ_2 and ρ_3 all equal to the correlation ρ_1. Thus in model $\mathcal{M}_3$ the correlation between each pair of random effects is equal to ρ_1. Note that the model $\mathcal{M}_3$ is also a sub-model of model $\mathcal{M}_2$.

The matrix Σ being positive definite, it follows:

$$\begin{cases} -\frac{1}{3} & < & \rho_1 < 1 \\ 0 & < & \sigma^2 \end{cases} .$$

We define the sub-model $\mathcal{M}_4$ of model $\mathcal{M}_1$ by assuming that there is no correlation between any pair of random effects. Note that the model $\mathcal{M}_4$ is a sub-model of model $\mathcal{M}_3$ obtained by setting the value of ρ_1 equal to zero.

The covariance matrix Σ of model $\mathcal{M}_4$ is given by:

$$\Sigma = \sigma^2 I_4,$$

where $\sigma^2 > 0$ and I_4 denotes the identity matrix of size 4.

Model $\mathcal{M}_4$ thus assumes independence between the event times in a cluster, or stated differently: the fact that an udder quarter has an infection does not have an influence on the hazard of infection for the three other udder quarters. Model $\mathcal{M}_4$ is termed as univariate frailty model, because each subject has its own frailty term.

We will now focus on parameter estimation in these four frailty models. We insist again on the fact that the four models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ and $\mathcal{M}_4$ are nested. This allows us to apply the likelihood ratio test to compare them in the forthcoming sections.

9.3. Statistical analysis of the mastitis dataset

We choose a frequentist approach and consider the Maximum Likelihood Estimate (MLE) for the parameter θ, namely the value $\hat{\theta}_N$ of θ which maximizes the marginal likelihood $L_N^{obs}(\mathbf{Y}, \mathbf{\Delta}; \theta)$ (see Duchateau and Janssen [2008]). We apply the estimation algorithm for frailty models presented in [A6].

We focus now on the theoretical properties of the algorithm.

Some particular assumptions on the frailty model have been proposed and stated in Theorem 1 in [A6] to ensure the almost sure convergence of this stochastic estimation algorithm toward the set of stationary points of the observed likelihood. They are fulfilled by the Weibull baseline function and by the multivariate Gaussian distribution for the frailty. Moreover the model $\mathcal{M}_1$, as well as the three sub-models $\mathcal{M}_2, \mathcal{M}_3$ and $\mathcal{M}_4$, satisfy the other additional regularity assumptions (denoted by **(H1-H2) and (SAEM2)** in Theorem 1) required to apply the convergence theorem. Furthermore we choose a step size sequence $\boldsymbol{\nu} = (\nu_k)_k$ and a transition probability kernel Π_θ for the stochastic approximation procedure which satisfy assumptions **(SAEM1)** and **(SAEM3)** of Theorem 1. Thus applying the result of Theorem 1 in [A6] proves that the parameter estimated sequence $(\theta_k)_k$ generated by our algorithm converges almost surely towards the set of stationary points of the observed likelihood

Parameters and mean model based standard errors estimates are obtained applying the methods described in Section 8, for each of the four models (cf Table 8). The mean model based standard errors are obtained as the square root of the diagonal components of the inverse of the Fisher Information Matrix.

Table 8

Parameter estimates and mean model based standard errors in parenthesis for the mastitis dataset for each of the four models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ with λ_0 and γ the parameters of the Weibull baseline hazard, β_1 the effect of parity, β_2 the effect of location and σ^2, ρ_1, ρ_2 and ρ_3 the parameters of the random effect distribution. The last line gives the log-likelihood.

	Models			
Parameters	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
λ_0	0.0104	0.0097	0.0098	0.0120
	(0.0017)	(0.0018)	(0.0018)	(0.0016)
γ	2.6373	2.6568	2.6433	2.6807
	(0.0498)	(0.0486)	(0.0521)	(0.0762)
β_1	0.9321	0.9848	0.9937	0.9983
	(0.2175)	(0.2085)	(0.2001)	(0.1116)
β_2	-0.3189	-0.3237	-0.3234	-0.2678
	(0.0659)	(0.0686)	(0.0683)	(0.1040)
σ^2	9.0715	9.1781	9.6812	7.7059
	(0.2934)	(0.2821)	(0.3919)	(0.5208)
ρ_1	0.9026	0.9003	0.8883	×
	(0.0124)	(0.0133)	(0.0123)	
ρ_2	0.8913	0.8772	×	×
	(0.0214)	(0.0145)		
ρ_3	0.8700	×	×	×
	(0.0198)			
Loglik.	-3889.6	-3889.8	-3891.8	-4867.5

Let us make some comments on the numerical results obtained from the real data set of mastitis. The parameter estimates of the baseline hazard function and of the covariates effects are similar, except for those obtained for model $\mathcal{M}_4$ which has a very elementary frailty correlation structure. The residual variance parameter σ^2 is higher in the models $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$, compare to model $\mathcal{M}_4$. This can be explain by the fact that this term grows to allow the model to fit the data under more constrain correlation structure. The correlation parameters ρ_1 and ρ_2 are stable in these three models and very close to 1. The estimated log-likelihoods are also similar except for the model $\mathcal{M}_4$ which has a much lower one. and lead us to conclude via a likelihood ratio test that the model $\mathcal{M}_4$ is very too simple for fitting this data set. The three other models give similar estimation results and log-likelihood values. We next compare model $\mathcal{M}_3$ with model $\mathcal{M}_2$. The likelihood ratio test statistic equals 4.23, leading to the p-value $P(X > 4.23) = 0.04$, where X is a chi-squared random variable with one degree of freedom. Therefore, model $\mathcal{M}_3$ can be rejected in favour of model $\mathcal{M}_2$ at the 5% significance level. Nevertheless let us keep in mind that the values of the stochastically estimated log-likelihood are very closed from each others leading to take this result with some precaution. Finally it is clear that model $\mathcal{M}_1$ does not lead to a better fit, as the log-likelihood value is only increasing minimally.

Model $\mathcal{M}_2$ is withheld for the current data set. The correlation structure in this model demonstrates first the substantial correlation between the udder quarter infection times within a cow. Therefore, an infected udder quarter is a higher risk factor for the three other udder quarters of the same cow than for udder quarters of a different cow. Many of the observed infections are from bacteria that are widespread in the environment so it could be rather due to the status of the cow than to the nearby presence of an infected udder quarter that makes that the other udder quarters of the same cow are more at risk.

Table 9

The mean of the parameter estimates of the 500 *generated datasets for the different parameters with the empirical and mean model based standard errors in parenthesis in models* $\mathcal{M}_2$ *and* $\mathcal{M}_3$.

Para.	model $\mathcal{M}_2$	model $\mathcal{M}_3$
$\lambda_0 = 0.01$	0.0114(0.0031/0.0015)	0.0098 (0.0018/0.0016)
$\gamma = 2.75$	2.7171(0.1427/0.0467)	2.7703 (0.1131/0.0502)
$\beta_1 = 0.90$	0.8721(0.2017/0.1891)	0.9253 (0.2107/0.1902)
β_2=-0.32	-0.3136(0.0787/0.0911)	-0.3358 (0.0833/0.0773)
$\sigma^2 = 9$	8.7712(1.0588/0.3087)	9.1954 (1.0276/0.4801)
$\rho_1 = 0.8$	0.8056(0.0259/0.0186)	0.7989 (0.0193/0.0132)
$\rho_2 = 0.7$	0.7039(0.0257/0.0498)	×

On the other hand, model $\mathcal{M}_2$ also reveals that the risk factor for an udder quarter is increased more if the udder quarter is in the same region, i.e., either front or rear, as the infected udder quarter. This finding means that the nearby presence of an infected udder quarter has the effect of increasing the risk of infection.

9.4. Simulation studies

The aim of the simulation studies is the validation of the results obtained for the mastitis data set in Section 9.1. In the first subsection 9.4.1 the bias of the parameter estimates for the two best models $\mathcal{M}_2$ and $\mathcal{M}_3$ is investigated through simulations in similar settings as the real dataset. In the second subsection 9.4.2, the finite sample size property of the likelihood ratio test for models $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ is considered to ensure that the level of the test is close to the nominal one.

9.4.1. Parameter Estimation Evaluation for Models $\mathcal{M}_2$ and $\mathcal{M}_3$

Data are simulated from models $\mathcal{M}_2$ and $\mathcal{M}_3$ using a similar setting as in the mastitis dataset. In total $N = 1196$ clusters are chosen, each cluster consisting in 4 infection times. Also the effect of parity and location is included.

The following parameter values are used to generate the data: $\lambda_0 = 0.01, \gamma = 2.75, \beta_1 = 0.90, \beta_2 = -0.32, \sigma^2 = 9, \rho_1 = 0.8$ and $\rho_2 = 0.7$ (only for model $\mathcal{M}_2$).

For each of the two models $\mathcal{M}_2$ and $\mathcal{M}_3$, 500 datasets were generated and the methods of Section 8 are applied to obtain the parameter and mean model based standard errors estimates for each of the datasets. The simulation results are presented in Table 9 and summarized by the mean of the parameter estimates of the 500 datasets, the empirical standar errors and the mean of the mean model based standard errors of the 500 datasets.

The mean of the estimates is closed to the true value for all parameters in each of the two models $\mathcal{M}_2$ and $\mathcal{M}_3$. In most cases, the values obtained for the estimated variances are also coherent, the model based one being lower than the empirical one.

Table 10

Empirical percentage of rejection of the LRT of H_0: the true model is $\mathcal{M}_4$ against H_1: the true model is $\mathcal{M}_3$ at level α.

α/σ^2	4	9	16
0.01	0.005	0.015	0.020
0.05	0.045	0.065	0.080
0.10	0.110	0.125	0.135

Table 11

Empirical percentage of rejection of the LRT of H_0: the true model is $\mathcal{M}_3$ against H_1: the true model is $\mathcal{M}_2$ at level α ($\sigma^2 = 9$).

α/ρ_1	0.3	0.5	0.8
0.01	0.030	0.015	0.010
0.05	0.085	0.060	0.055
0.10	0.135	0.115	0.095

9.4.2. Likelihood ratio tests

In this section we investigate through a simulation study the finite sample size property of the likelihood ratio test (LRT) for testing the hypothesis H_0: the true model is model $\mathcal{M}_4$ against the hypothesis H_1: the true model is model $\mathcal{M}_3$. To that purpose, we repeat 200 times the following experiment: we simulate under model $\mathcal{M}_4$ a data set using the general setting described in Subsection 9.4.1, we estimate from these simulated data the MLE of the parameters in model $\mathcal{M}_4$ and in model $\mathcal{M}_3$ and compute the corresponding log-likelihood ratio test statistic. Then we compute the empirical confidence level for different levels $\alpha \in \{0.01, 0.05, 0.10\}$. Note that the model $\mathcal{M}_4$ is a sub-model of model $\mathcal{M}_3$ with the value of parameter ρ_1 being equal to zero which is not at the boundary of the parameter space. Thus the likelihood ratio test statistic converges in law towards a chi square random variable with one degree of freedom (see Cox and Hinkley [1974]). The corresponding quantiles are then respectively equal to 6.63, 3.84 and 2.70. We repeat this experiment for different values of the parameter σ^2 chosen in $\{4, 9, 16\}$. The results are presented in Table 10. For example, for $\sigma^2 = 9$, we get an empirical percentage of rejection for the LRT of 0.065 which is close to the nominal level $\alpha = 0.05$. We observe that the empirical percentage of rejection differs more from the nominal level α when the variance parameter equals to 16.

We conduct the same experiment taking H_0: the true model is $\mathcal{M}_3$ against hypothesis H_1: the true model is $\mathcal{M}_2$. We fix the residual variance parameter σ^2 to 9 and let the correlation parameter ρ_1 vary between 0.3, 0.5 and 0.8. The corresponding results are presented in Table 11. We observe that the empirical percentage of rejection differs more from the nominal level when the correlation parameter ρ_1 equals 0.3.

The empirical results given in Tables 10 and 11 demonstrate that the LRT can be applied for a finite sample size N equal to 1196 when the parameter σ^2 is closed to 9 (and smaller) and when the parameter ρ_1 is closed to 0.5 (and larger). This heuristic observation ensures the validity of the LRT with the mastitis data in the previous section.

Conclusions and perspectives

This manuscript summarizes my contributions to the field of statistical research. They mainly deal with developments, studies and applications of new stochastic algorithmic methods derived from the Expectation Maximization algorithm for estimation in complex latent variable models. I attempted to address the theoretical aspects and the practical ones in each of my contributions. Most of my research was inspired by practical problems. Their consequences were mainly to provide more flexibility in modeling since they give solutions for parameter estimation in many complex latent variable models used in different application fields. This makes it possible, in particular, to more efficiently assess numerous complex phenomena.

All of the research topics mentioned in this document were not treated in the same way. Some were explored in depth from several different angles, leading to other developments. Other ones, generally more recent, are still being investigated and will most certainly lead to new research directions. I mentioned in the manuscript below some contributions some open questions which could lead to further developments. Besides my major actual perspectives motivated by practical questions are detailed in the next paragrafs.

Estimation in partially observed multi-type branching processes

In collaboration with Catherine Larédo (INRA, MIA), we will address the estimation in partially observed multi-type branching processes with immigration. This work is motivated by the analysis of the dynamics of rape populations. This plant has two specific properties: first, it is feral, meaning that it is able to grow by itself without particular care; second, the seeds can remain in a seed bank in the soil for several years before eventually giving rise to a new plant. A multi-type branching process is well adapted to modeling this dynamics where the different types correspond to the flowers, the seeds on the soil, the seeds in the seed bank in the soil, etc. Nevertheless, only some types were observed, namely the flowers. The model parameters such as the probability of a seed in the seek bank in the soil eventually giving rise to a new plant, characterize the dynamics and are of great interest for controlling the behavior of such a population. The estimation of these parameters is usually done in branching processes through optimization of contrast processes such as conditional least squares. However, in the presence of unobserved types in multi-type branching processes, this is no longer possible. One alternative to these approaches could be to consider the maximum likelihood approach. Note that the unobserved types are latent variable of the multi-type branching processes. However it is not possible to apply directly an EM like algorithm, the model being non exponential when considering usual reproduction and immigration distributions. Therefore we plan to consider an approximated Gaussian model for the multi-type branching process in which the estimation can be carried out through maximum likelihood by using a stochastic estimation algorithm. Our purpose is to adapt the work presented in Section 2.2 to this context to compute the MLE in the approximated Gaussian model. The proposed estimation algorithm will be apply to a rape dataset collected in Selomnes. We also plan to study the asymptotic property of the maximum likelihood estimate in the approximated model under misspecified model conditions. Our approach may be generalized to other multivariate partially observed models.

Modeling and estimation of epidemic dynamics

In collaboration with Elisabeta Vergu, (INRA, MIAJ), we are studying the epidemic dynamics modeled by compartment models and estimation in such models. The propagation of epidemic is often described by compartments, where one compartment corresponds to one possible status of an individual, e.g., sensitive, infected or recovered, in the simple case. The estimation task is difficult since observations are very often partial. Models in which the estimation can be carried out often require stringent structural assumptions of the population dynamics, e.g. Markovian. Such an assumption is of course not realistic since it implies that the recovering rate is constant, i.e., the sojourn times in the different compartments are distributed from an exponential distribution that is not realistic. We propose to consider a new non Markovian dynamics compartment model, that also takes the errors raised by the observation process into account. We plan to consider the maximum likelihood estimate and to evaluate it in practice through a stochastic estimation algorithm. We project to study the theoretical properties of the maximum likelihood estimate and of the stochastic estimation algorithm. Simulations studies and application to real epidemic data will also be carried out.

Modeling and estimation in survival data analysis

The two contributions to estimation in the frailty models presented in the third part of the manuscript have given rise to several interesting questions.

In collaboration with Charles El-Nouty (University Paris 13 Sorbonne Paris Cité, LAGA), we plan to address the case of maximum likelihood estimation in frailty models with a piecewise constant baseline function or a non parametric one. From a modeling point of view, this will make it possible to be free from the choice of a specific parametric form for the baseline, which is not obvious in practice. Another similar theoretical development will be to consider the partial likelihood, as was done for the estimation in the Cox model, rather than the marginal likelihood. This also allows us to bypass the effect of the parametric choice for the baseline. Finally, motivated by practical applications in epidemiology, we are interested in the estimation in competing risk models that integrate frailty terms. In competing risk approaches, the hazard rate is obtained as the sum of several independent hazards, each corresponding to different risks. Such modeling is, for example, well adapted to considering the risk of death, on the one hand, from a cancer, and, on the other, from cardiovascular disease.

In collaboration with Luc Duchateau and Klaartje Goethals (Faculty of Veterinary Medicine, Ghent University), we plan to assess the correlation structure in a malaria epidemic dataset in Ethiopia by introducing spatial modeling in the correlation matrix of the frailty vector. Moreover, in this dataset, the individuals are clustered through two stages, namely the villages and the houses. This will be modeled by using hierarchical models.

Another open question in the frailty model setting is the one of model validation. It would be interesting to propose goodness-of-fit tests, in particular for testing hypotheses on the covariates used in the design matrices and on the correlation structure of the frailty terms.

Analysis of plant growth

In collaboration with Hervé Monod (INRA, MIAJ), Alain Charcosset (INRA, URGV Le Moulon) and François Tardieu (INRA, LEPSE), we are interested in genotype by environment interaction. In particular, we plan to address parameter estimation in dynamic plant growth models based on several types of data including different genotypes and different environmental conditions. Indeed multiple information sources such as bibliographies, datasets obtained on platforms and datasets obtained in field conditions exist. All of this information is relevant when estimating model parameters. However, since the objective is to use the dynamic model to predict phenotypic characteristics in field conditions for existing or possibly new genotypes in different climatic scenarios. Thus parameter estimation requires close attention, particularly concerning the use of platform datasets which do not report all the phenomena existing in fields conditions. Therefore we have to integrate in the global estimation process, possibly also in the modeling, all these different types of data taking into account their specificity. Besides modeling task can also be addressed since most of the dynamic models are deterministic, although they model stochastic plants behavior. Thus it would be interesting to enrich these models by adding stochastic terms taking the existing variablity into account.

This work takes place within the project ”investissement d'avenir” AMAIZING coordinated by Alain Charcosset.

In collaboration with Paul-Henry Cournède and Charlotte Baey (ECP, Digiplante), we are provided with a given deterministic dynamic model for plant growth, having parameters which may depend on the plants genotype. First, we are interested in identifying which of the model parameters could be considered fixed for several genotypes to reduce the model dimension. We plan to use non linear mixed effect models and to determine which parameters can be modeled as population parameters to analyse some arabidopsis data. In a second time, we plan to cluster the different genotypes of our dataset by using a mixture model in the spirit of [A5].

This work takes place within the project ”plante entière” of the Institut de Modélisation des Sciences du Vivant (IMSV), directed by Vincent Fromion (INRA, MIG).

In collaboration with Adrienne Ressayre (INRA, URGV Le Moulon), we plan to analyze maize leave dataset. Geneticians are interested in leaf area, since it is directly link with photosynthesis and therefore with plant development, resulting in early or late flowering time. Moreover, leaf area seems to be a genotypic characteristic. To validate this hypothesis, we plan to use the BME template model presented in [A3] to estimate leaf patterns for several genotypes possibly clustered.

This work takes place within to a global project ITEMAIZE dedicated to the study of the early development of maize, in particular the floral transition, coordinated by Christine Dillmann (UPS, URGV Le Moulon) actually submitted.

Motivated by the collection processes for agronomic data in plant growth studies, I am interested in considering more precisely the particular form of partially observed data called current status data: observations are composed of the observation date and of the status of the individual at this date. This can be viewed as a form of complex censoring. Let us consider the following example: the interest is in the appearance of the third leaf of the maize plant. In practice, may be the technician crosses the field every two days only, leading to the observation, for example, that on day d, the maize plant has two leaves, and that on day $d+2$, it has three leaves, generating the censored information that the third leaf appears between days d and $d+2$. Such data frequently arise in plant growth studies in agronomy. Note that such data are also common in the field of epidemiology, leading to many possible applications. Modeling such partially observations leads to specific complex latent variable models. Parameter estimation is therefore particularly intricated. Existing methods are based on several types of approximations and suffer from theoretical background and numerical limitation. New subtantial developments from both the theoretical and the practical point of view are therefore required in this field.

Bibliography

References

O. Aalen and S. Tretli. Analysing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, 10:285–292, 1999.

J. C. Abrahantes, C. Legrand, T. Burzykowski, P. Janssen, V. Ducrocq, and L. Duchateau. Comparison of different estimation procedures for proportional hazards model with random effects. *Comput. Statist. Data Anal.*, 51(8):3913–3930, 2007. ISSN 0167-9473.

S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(1):3–29, 2007. ISSN 1369-7412.

S. Allassonnière, A. Trouvé, and L. Younes. Geodesic shotting and diffeomorphic matching via textured meshes. In A. Y. Anand Rangarajan, Baba Vemuri, editor, *Proc. of the Energy Minimization Methods for Computer Vision and Pattern Recognition (EMMCVPR 05)*, pages 365–381, November 9-11 2005.

Y. Amit. Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.*, 24 (1):122–140, 1996. ISSN 0090-5364.

Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86:376–387, 1989.

C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006. ISSN 1050-5164.

C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312 (electronic), 2005. ISSN 0363-0129.

Y. Atchadé. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, 8:235–254, 2006.

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251, 2003. ISSN 0090-5364. .

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J. Comp. Vis.*, 61(2):139–157, 2005.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations.* Springer-Verlag, Berlin, 1990. ISBN 3-540-52894-6. Translated from the French by Stephen S. Wilson.

J. Bigot and B. Charlier. On the consistency of Fréchet means in deformable models for curve and image analysis. *Electron. J. Stat.*, 5:1054–1089, 2011. ISSN 1935-7524.

G. Celeux and J. Diebolt. L'algorithme SEM: Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. (The SEM algorithm: An algorithm of probabilistic learning for the determination of mixtures of densities). *Rev. Stat. Appl.*, 34(2):35–52, 1986.

H. Chen, L. Guo, and A. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Proc. Appl.*, 1988.

J. C. Chen. Testing for no effect in nonparametric regression via spline smoothing techniques. *Ann. Inst. Statist. Math.*, 46(2):251–265, 1994. ISSN 0020-3157.

N. Chopin, T. Lelièvre, and G. Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Stat. Comput.*, 22(4):897–916, 2012. ISSN 0960-3174.

E. Christensen, G, D. Rabbitt, R, and I. Miller, M. Deformable templates using large deformation kinematics. *IEEE trans. Image Proc.*, 1996.

G. Claeskens, H. Ding, and M. Jansen. Lack-of-fit tests in linear mixed models with application to wavelet tests. *J. Nonparametr. Stat.*, 23(4):853–865, 2011. ISSN 1048-5252. .

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. Ser. A*, 148(2):82–117, 1985. ISSN 0035-9238.

D. Concordet and O. G. Nunez. A simulated pseudo-maximum likelihood estimator for nonlinear mixed models. *Comput. Statist. Data Anal.*, 39(2):187–201, 2002. ISSN 0167-9473.

J. Cortiñas Abrahantes and T. Burzykowski. A version of the EM algorithm for proportional hazard model with random effects. *Biom. J.*, 47(6):847–862, 2005. ISSN 0323-3847.

D. R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. ISSN 0035-9246. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

D. R. Cox and D. V. Hinkley. *Theoretical statistics.* Chapman and Hall, London, 1974.

C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):165–185, 2004. ISSN 1369-7412.

M. Davidian and D. M. Giltinan. *Nonlinear models for Repeated Measurement Data.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1995. ISBN 0-412-98341-9.

B. Delyon. Stochastic approximation with decreasing gain: convergence and asymptotic theory. *Technical Report: Publication interne 952, IRISA*, 2000.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 1999. ISSN 0090-5364.

E. Demidenko. *Mixed models.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2004. ISBN 0-471-60161-6. . Theory and applications.

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

H. Dette and A. Munk. Validation of linear regression models. *Ann. Statist.*, 26(2):778–800, 1998. ISSN 0090-5364.

J. Diebolt and J. Zuber. On testing the goodness-of-fit of nonlinear heteroscedastic regression models. *Comm. Statist. Simulation Comput.*, 30(1):195–216, 2001. ISSN 0361-0918.

P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of longitudinal data*, volume 25 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition, 2002. ISBN 0-19-852484-6.

L. Duchateau and P. Janssen. *The Frailty Model*. Statistics for Biology and Health. Springer-Verlag, New York, 2008. ISBN 978-0387728346.

V. Ducrocq and G. Casella. A bayesian analysis of mixed survival models. *Genet Sel Evol*, 28:509–529, 1996.

P. Dupuis, U. Grenander, and M. Miller. Variational problems on flows of diffeomorphisms for image matching. *Quaterly of Applied Math.*, 1998.

S. Durrleman. *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution*. PhD thesis, Ecole Normale Supérieure de Cachan, France, 2010.

S. Durrleman, M. Prastawa, G. Gerig, and S. Joshi. Optimal data-driven sparse parameterization of diffeomorphisms for population analysis. In G. Székely and H. Hahn, editors, *Information Processing in Medical Imaging (IPMI)*, volume 6801 of *LNCS*, pages 123–134, 2011.

S. Durrleman, S. Allassonnière, and S. Joshi. Sparse adaptive parameterization of variability in image ensembles. *IJCV*, 2012.

R. L. Eubank and V. N. LaRiccia. Testing for no effect in nonparametric regression. *J. Statist. Plann. Inference*, 36(1):1–14, 1993. ISSN 0378-3758. .

G. Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *Accepted for publication in EsaimPS*, 2014.

G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Stat.*, 31(4):1220–1259, 2003.

G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 2012.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Interdisciplinary Statistics. Chapman & Hall, London, 1996. ISBN 0412055511.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(2):123–214, 2011. ISSN 1369-7412. With discussion and a reply by the authors.

C. A. Glasbey and K. V. Mardia. A penalised likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B*, 63:465–492, 2001.

J. Glaunès, M. Vaillant, and M. I. Miller. Landmark matching via large deformation diffeomorphisms on the sphere. *Journal of Mathematical Imaging and Vision, MIA 2002 special issue*, (to appear) 2003.

K. Goethals, B. Ampe, H. Berkvens, H. Laevens, P. Janssen, and L. Duchateau. Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model. *JABES*, 14(1):1–14, 2009.

W. Gonzalez-Manteiga and R. M. Crujeiras. An updated review of Goodness-of-Fit tests for regression models. *TEST*, 22(3):361–411, SEP 2013.

R. Gray. Spline-based tests in survival analysis. *Biometrics*, 50(3):640–652, 1994. ISSN 0006-341X.

R. Gray. Tests for variation over groups in survival data. *J. Amer. Statist. Assoc.*, 90(429):198–203, 1995.

ISSN 0162-1459.

S. Greven and C. M. Crainiceanu. On likelihood ratio testing for penalized splines. *ASTA-Advances in Statistical Analysis*, 97(4):387–402, Oct 2013. ISSN 1863-8171. .

S. Greven, C. M. Crainiceanu, H. Küchenhoff, and A. Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *J. Comput. Graph. Statist.*, 17(4):870–891, 2008. ISSN 1061-8600. .

E. Guerre and P. Lavergne. Data-driven rate-optimal specification testing in regression models. *Ann. Statist.*, 33(2):840–870, 2005. ISSN 0090-5364. .

W. Härdle and E. Mammen. Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947, 1993. ISSN 0090-5364. .

W. Härdle, E. Mammen, and M. Müller. Testing parametric versus semiparametric modeling in generalized linear models. *J. Amer. Statist. Assoc.*, 93(444):1461–1474, 1998. ISSN 0162-1459.

J. D. Hart. *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. Springer-Verlag, New York, 1997. ISBN 0-387-94980-1.

R. Holm, D, T. Ratnanather, J, A. Trouvé, and L. Younes. Soliton dynamics in computational anatomy. *Neuroimage*, 23:S170–S178, 2004.

J. L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-92869-2.

J. L. Horowitz and V. G. Spokoiny. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631, 2001. ISSN 0012-9682. .

P. Hougaard. *Analysis of multivariate survival data*. Statistics for Biology and Health. Springer-Verlag, New York, 2000. ISBN 0-387-98873-4.

M. Huang and D. Zhang. Testing polynomial covariate effects in linear and generalized linear mixed models. *Stat. Surv.*, 2:154–169, 2008. ISSN 1935-7516.

S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000. ISSN 0304-4149.

R. H. Jones. *Longitudinal data with serial correlation: a state-space approach*, volume 47 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1993. ISBN 0-412-40650-0.

S. Joshi and M. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions in image processing*, 9(8):1357–1370, 2000.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958. ISSN 0162-1459.

H. J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1997. ISBN 0-387-94916-X.

K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 57 (2):425–437, 1995. ISSN 0035-9246.

C. Legrand, V. Ducrocq, P. Janssen, R. Sylvester, and L. Duchateau. A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. *Stat. Med.*, 24 (24):3789–3804, 2005. ISSN 0277-6715.

C. Legrand, L. Duchateau, P. Janssen, V. Ducrocq, and R. Sylvester. Validation of prognostic indices using the frailty model. *Lifetime Data Anal.*, 15(1):59–78, 2009. ISSN 1380-7870.

T. Lelièvre and K. Minoukadeh. Long-time convergence of an adaptive biasing force method: the bi-channel case. *Arch. Ration. Mech. Anal.*, 202(1):1–34, 2011. ISSN 0003-9527.

T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with adaptive parallel dynamics. *J. Chem. Phys.*, 2007.

T. Lelièvre, M. Rousset, and G. Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155–1181, 2008. ISSN 0951-7715.

M. J. Lombardía and S. Sperlich. Semiparametric inference in generalized mixed effects models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):913–930, 2008. ISSN 1369-7412. .

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 44(2):226–233, 1982. ISSN 0035-9246.

T. Marshall and G. Roberts. An adaptive approach to langevin MCMC. *Statistics and Computing*, 22 (5):1041–1057, 2012.

S. Marsland, C. Twining, and C. Taylor. A minimum description length objective function for groupwise non rigid image registration. *Image and Vision Computing*, 2007.

C. McGilchrist and C. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47:461–466, 1991.

X.-L. Meng. On the rate of convergence of the ECM algorithm. *Ann. Stat.*, 22(1):326–339, 1994.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 2009. ISBN 3-540-19832-6.

M. Miller and L. Younes. Group action, diffeomorphism and matching: a general framework. *Int. J. Comp. Vis*, 41:61–84, 2001. (*Originally published in electronic form in: Proceeding of SCTV 99, http://www.cis.ohio-state.edu/ szhu/SCTV99.html*).

M. I. Miller, A. Trouvé, and L. Younes. On the metrics and Euler-Lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4:375–405, 2002.

M. I. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, 2006. ISSN 0924-9907. .

K. Minoukadeh, C. Chipot, and T. Lelièvre. Potential of mean force calculations: a multiple-walker adaptive biasing force approach. *J. Chem. Th. Comput.*, 2010.

H.-G. Müller. Goodness-of-fit diagnostics for regression models. *Scand. J. Statist.*, 19(2):157–172, 1992. ISSN 0303-6898.

S. A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, 22(2):712–731, 1994. ISSN 0090-5364.

S. A. Murphy. Asymptotic theory for the frailty model. *Ann. Statist.*, 23(1):182–198, 1995. ISSN 0090-5364.

L. Nie and M. Yang. Strong consistency of MLE in nonlinear mixed-effects models with large cluster size. *Sankhyā*, 67(4):736–763, 2005. ISSN 0972-7671.

G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sørensen. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, 19(1):25–43, 1992. ISSN 0303-6898.

Z. Pan and D. Y. Lin. Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4): 1000–1009, 2005. ISSN 0006-341X. .

E. Parner. Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, 26(1):183–214, 1998. ISSN 0090-5364.

X. Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, July 2006.

J. Pinheiro and D. Bates. *Mixed-effects models in S and S-Plus*. Springer, New York, 2000.

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

S. Ripatti, K. Larsen, and J. Palmgren. Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Anal.*, 8(4):349–360, 2002. ISSN 1380-7870.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 1951.

C. Robert. *Méthodes de Monte Carlo par chaînes de Markov*. Statistique Mathématique et Probabilité. [Mathematical Statistics and Probability]. Éditions Économica, Paris, 1996. ISBN 2-7178-3154-1.

V. Rondeau, D. Commenges, and P. Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Anal.*, 9(2):139–153, 2003. ISSN 1380-7870.

F. Scheipl, S. Greven, and H. Küchenhoff. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Statist. Data Anal.*, 52(7): 3283–3299, 2008. ISSN 0167-9473.

G. A. F. Seber and C. J. Wild. *Nonlinear regression*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1989. ISBN 0-471-61760-1. .

L. Sheiner, Y. Hashimoto, and S. Beal. A simulation study comparing designs for dose ranging. *Statistics in Medicine*, 10:303–321, 1991.

J. G. Staniswalis and T. A. Severini. Diagnostics for assessing regression models. *J. Amer. Statist. Assoc.*, 86(415):684–692, 1991. ISSN 0162-1459.

W. Stute. Nonparametric model checks for regression. *Ann. Statist.*, 25(2):613–641, 1997. ISSN 0090-5364.

J. Q. Su and L. J. Wei. A lack-of-fit test for the mean function in a generalized linear model. *J. Amer. Statist. Assoc.*, 86(414):420–426, 1991. ISSN 0162-1459.

R. Sylvester, A. van der Meijden, W. Oosterlinck, J. Witjes, C. Bouffioux, L. Denis, D. Newling, and K. Kurth. Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using eortc risk tables: A combined analysis of 2596 patients from seven eortc trials. *European Urology*, 49:466–477, 2006.

T. Therneau and P. Grambsch. *Modeling Survival Data : Extending the Cox Model*. Springer-Verlag, New York, 2000.

A. Trouvé. Diffeomorphism groups and pattern matching in image analysis. *International Journal of Computer Vision*, 28(3):213–221, 1998.

F. Vaida. Parameter convergence for EM and MM algorithms. *Statist. Sinica*, 15(3):831–840, 2005. ISSN

1017-0405.

F. Vaida and R. Xu. Proportional hazards model with random effects. *Statis. Med.*, 19:3309–3324, 2000.

M. Vaillant, I. Miller, M, A. Trouvé, and L. Younes. Statistics on diffeomorphisms via tangent space representations. *Neuroimage*, 23(S1):S161–S169, 2004.

J. W. Vaupel, K. G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

E. F. Vonesh. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83(2):447–452, 1996. ISSN 0006-3444.

E. F. Vonesh and V. M. Chinchilli. *Linear and nonlinear models for the analysis of repeated measurements*, volume 154 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York, 1997. ISBN 0-8247-8248-8. With 1 IBM-PC floppy disk (3.5 inch; HD).

J. Wakefield. The Bayesian analysis of population pharmacokinetic models. *J. Am. Stat. Assoc.*, 91(433): 62–75, 1996.

J. Wakefield, A. Smith, A. Racine-Poon, and A. Gelfand. Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. *J. R. Stat. Soc., Ser. C*, 43(1):201–221, 1994.

S. Walker. An EM algorithm for nonlinear random effects models. *Biometrics*, 52(3):934–944, 1996. ISSN 0006-341X.

F. Wang and D. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 2001.

G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithms. *J. Amer. Statist. Assoc.*, 85(411):699–704, 1990.

C.-F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95–103, 1983. ISSN 0090-5364.

M. Zhang, N. Singh, and P. T. Fletcher. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. *IPMI*, 2013.

List of works

List of publications

[A1] Kuhn, Estelle and Lavielle, Marc (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM Probab. & Stat.*, **8**, 115–131.

[A2] Kuhn, Estelle and Lavielle, Marc (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, **49**, 4, 1020–1038.

[A3] Allassonnière, Stéphanie and Kuhn, Estelle and Trouvé, Alain (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli*, **16**, 3, 641–678.

[A4] Allassonnière, Stéphanie and Kuhn, Estelle and Trouvé, Alain (2010). Bayesian Consistent Estimation in Deformable Models using Stochastic Algorithms: Applications to Medical Images. *Journal de la Société Française de Statistique*, **151**, 1, 1–16.

[A5] Allassonnière, Stéphanie and Kuhn, Estelle (2010). Stochastic algorithm for Bayesian mixture effect template estimation. *ESAIM Probab. & Stat.*, **14**, 382–408.

[A6] Kuhn, Estelle and El-Nouty, Charles (2013). On one convergent stochastic estimation algorithm for frailty models. *Statistics and Computing*, **23**, 3, 413–423.

[A7] R. Pilon, C. Picon-Cochard, J.M.G. Bloor, S. Revaillot, E. Kuhn, R. Falcimagne, P. Balandier, J.-F. Soussana (2013). Grassland root demography responses to multiple climate change drivers depend on root morphology. *Plant and Soil*, **364**, 1-2, 395–408.

[A8] Fort, Gersende and Jourdain, Benjamin and Kuhn, Estelle and Lelièvre, Tony and Stoltz, Gabriel (2014). Efficiency of the Wang-Landau algorithm: a simple test case. *Applied Mathematics Research Express*, **2014**, 2, 275–311.

[A9] R. Rincent, L. Moreau, H. Monod, E. Kuhn, A.E. Melchinger, R.A. Malvar, J. Moreno-Gonzalez, S. Nicolas, D. Madur, V. Combes, F. Dumas, T. Altmann, D. Brunel, M. Ouzunova, P. Flament, P. Dubreuil, A. Charcosset, T. Mary-Huard (2014). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics*, **197**, 1, 375–387.

[A10] Huet, Sylvie and Kuhn, Estelle (2015). Goodness-of-fit test for Gaussian regression with block correlated errors. *Statistics*, **49**, 2, 239–266.

[A11] Fort, Gersende and Jourdain, Benjamin and Kuhn, Estelle and Lelièvre, Tony and Stoltz, Gabriel (2015). Convergence and efficiency of the Wang-Landau algorithm. *Mathematics of Computation*, **84**, 2297–2327.

[A12] Allassonnière, Stéphanie and Kuhn, Estelle (2015). Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation. *Computational Statistics and Data Analysis*, **91**, 4–19.

[A13] Allassonnière, Stéphanie and Durrleman, Stanley and Kuhn, Estelle (2015). Bayesian Mixed Effect Atlas Estimation with a Diffeomorphic Deformation Model. *SIAM Journal on Imaging Sciences*, **8**, 3, 1367–1395.

List of proceedings

[P1] Kuhn, Estelle and Lavielle, Marc (2002). Convergence d'une version MCMC de l'algorithme EM. *Actes des XXXIVèmes journées de statistique*, Bruxelles.

[P2] Sahmoudi, M. and Abed-Meraim, K. and Lavielle, M. and Kuhn, E. and , Ciblat, Ph. (2005). Blind Source Separation of Noisy Mixtures Using a Semi-Parametric Approach With Application to Heavy-Tailed Signals. EUSIPCO Conference, Antalya, Turkey.

[P3] Allassonniére, Stéphanie and Kuhn, Estelle and Amit, Yali and Trouvé, Alain (2006). Generative Model and consistent estimation algorithms for non-rigid deformable models. ICASSP conference, Toulouse.

[P4] Allassonnière, Stéphanie and Kuhn, Estelle and Trouvé, Alain (2008). MAP Estimation of Statistical Deformable Template Via Nonlinear Mixed Effect Models: Deterministic and Stochastic Approaches. Mathematical Foundations of Computational Anatomy (MFCA) workshop of MICCAI 2008 conference.

[P5] Allassonnière, Stéphanie and Kuhn, Estelle and Ratnanather, J. Tilak and Trouvé, Alain (2009). Consistent Atlas Estimation on BME Template Model: Applications to 3D Biomedical Images. Probabilistic Models for Medical Image Analysis (PMMIA) worshop of the MICCAI 2009 conference.

[P6] Kuhn, Estelle and El-Nouty, Charles (2010). Modèles de fragilité et algorithme EM stochastique. *Actes des XXXXIIèmes journées de statistique*, Marseille.

List of technical reports and preprints

[R1] Kuhn, Estelle and Goethals, Klaartje and El-Nouty, Charles and Duchateau, Luc (2014). Using correlated frailty model to analyze cow udder quarter infection times. Technical report INRA, in revision.

Thesis

[T] Kuhn, Estelle (2003). Estimation par maximum de vraisemblance dans des problèmes inverses non linéaires. Thèse de l'Université Paris Sud (Orsay).

Main co-authors and collaborators

Stéphanie Allassonnière, Ecole Polytechnique, CMAP, Palaiseau.
Alain Charcosset, INRA, UMR Génétique Quantitative et Evolution, Le Moulon.
Luc Duchateau, Ghent University, Faculty of Veterinary Medecine, Ghent, Belgique.
Charles El-Nouty, Université Paris Nord, LAGA, Villetaneuse.
Gersende Fort, Telecom Paris Tech, LTCI, Paris.
Klaartje Goethals, Ghent University, Faculty of Veterinary Medecine, Ghent, Belgique.
Sylvie Huet, INRA, MaIAGE, Jouy-en-Josas.
Benjamin Jourdain, Ecole Nationale des Ponts et Chaussées, CERMICS, Marne La Vallée.
Catherine Larédo, INRA, MaIAGE, Jouy-en-Josas.
Marc Lavielle, Université Paris Sud, Laboratoire de Mathématiques, Orsay.
Tony Lelièvre, Ecole Nationale des Ponts et Chaussées, CERMICS, Marne La Vallée.
Hervé Monod, INRA, MaIAGE, Jouy-en-Josas.
Renaud Rincent, INRA, UMR Génétique, Diversité et Ecophysiologie, Clermont-Ferrand.
Gabriel Stoltz, Ecole Nationale des Ponts et Chaussées, CERMICS, Marne La Vallée.
Alain Trouvé, ENS , CMLA, Cachan.
Elisabeta Vergu, INRA, MaIAGE, Jouy-en-Josas.