# Statistical modelling of medical data and theoretical analysis of estimation algorithms

Vianney Debavelaere

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAX034

Thèse de doctorat

# Modélisation statistique de données médicales et analyse théorique des algorithmes d'estimation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Ecole Polytechnique

École doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 25 juin 2021, par

**VIANNEY DEBAVELAERE**

Composition du Jury :

| | |
|---|---|
| Emmanuel Gobet<br>Professeur, École Polytechnique | Président |
| Christophe Andrieu<br>Professeur, University of Bristol | Rapporteur |
| Jean-Michel Marin<br>Professeur, Université de Montpellier | Rapporteur |
| Maria Vakalopoulou<br>Maitre de conférence, Centrale Supélec | Examinatrice |
| Stéphanie Allassonnière<br>Professeur, Université Paris-Descartes | Directrice de thèse |
| Stanley Durrleman<br>Directeur de Recherche, Inria | Co-directeur de thèse |






**Vianney Debavelaere**
*vianney.debavelaere@polytechnique.edu*

# Abstract

The emergence of large longitudinal data sets (subjects observed repeatedly at different time points) has allowed the construction of different models improving the understanding of biological or natural phenomena. Longitudinal studies have numerous applications: understanding of the differences of progression in neurodegenerative disease such as Alzheimer's, chemotherapy monitoring, facial recognition, etc.. To process those data sets, mixed effects hierarchical models can be used. They separate the population parameters (called *fixed effects*) from the subjects random variables. One possible aim of those models is to create an atlas i.e. to estimate a representative image, shape or even trajectory of the population together with the inter-individual variability. Doing so, it is for instance possible to highlight the influence of a disease on a particular organ compared to a normal ageing. The random variables then give the population variability. One is able to use them to reconstruct the trajectory of each patient, predict his future, issue a diagnosis, etc..

In practice, different data sets can be studied from the medical domain. Two or three dimensional images are obtained from scanners, MRI, radiography, etc.. Sometimes, meshes of a particular organ are extracted from those images. To manage those different structures, a unique framework has been developed. Using Riemannian geometry, one can create distances, compute deformations and means via the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework. Models have first been developed for cross sectional data sets and have then been generalized for longitudinal data sets. They, however, present several limitations that we present here and that we will overcome in this thesis.

First, in the longitudinal case, those models assume that the population representative trajectory follows a unidirectional dynamic. While correct in certain cases (atrophy of the hippocampus for instance), it is not verified in others. In chemotherapy monitoring, a tumor often first shrinks before growing again. A unidirectional trajectory cannot explain this behaviour. To overcome this problem, we will propose a hierarchical mixed effect model allowing to consider branching populations.
One of the advantages of the LDDMM framework is the fact that it uses diffeomorphic deformations. However, this advantage can also pose a problem for some data sets. Indeed, for an images data set issued from chemotherapy monitoring, each patient can have a different number of tumors. But, being a diffeomorphic deformation from a unique representative image, the reconstruction of any subject will always have the same number of tumors as the representative trajectory. In this thesis, we model the residuals as a sparse matrix, allowing to detect and recover anomalies, such as tumors for example, on an observation.

In addition to providing new models, we will also focus on the estimation of the parameters of non-linear mixed effects models as the one we have used. In practice, one uses Stochastic Approximation Expectation Maximization (SAEM) algorithms coupled with Markov Chain Monte Carlo methods. We will relax two hypotheses of these algorithms. The first one is the necessity of geometric ergodicity of the Markov Chain, preventing from considering distributions with heavy tails. We will relax this hypothesis by proposing a new set of assumptions, asking only subgeometric ergodicity. Moreover, the SAEM algorithm asks the joint distribution of the mixed effect model to belong to the curved exponential family. This hypothesis is in fact a bottleneck in lots of situations. Here, we study an idea proposed by Kuhn and Lavielle (2005) allowing to consider distributions that do not verify this assumption. We however show that this trick can introduce a bias in the estimation and propose a new algorithm reducing it.

# Contents

# Remerciements

C'est avec beaucoup de reconnaissance que je commence ce manuscrit par mes remerciements à ceux, très nombreux, qui m'ont aidé et soutenu au cours de ma thèse.

Dans un premier temps, je me tourne bien entendu vers mes directeurs de thèse. Stéphanie, merci d'avoir toujours été présente au cours de ces trois années. Merci pour ces innombrables rendez-vous et visio-conférences. Ton enthousiasme m'a permis de ressortir toujours plus motivé de nos rencontres et c'est en grande partie grâce à toi que ces trois années ont pu être si enrichissantes. Je suis très heureux de pouvoir continuer à travailler à tes côtés par la suite.
Si j'ai moins travaillé avec toi Stanley après la première année, tu as toujours été présent quand j'en avais besoin : aussi bien pour relire un article, conseiller sur la rédaction ou trouver une salle pour la soutenance. Merci pour tous ces conseils précieux !

Je remercie également Christophe Andrieu et Jean Michel Marin d'avoir immédiatement et avec beaucoup d'enthousiasme accepté de relire mon manuscrit. Merci aussi à Emmanuel Gobet et Maria Vakalopoulou d'avoir accepté de faire partie de mon jury.

Je me tourne maintenant vers toutes les personnes avec lesquelles j'ai travaillé au cours de ces trois années. Je pense notamment à Alexandre et Igor qui ont, avec beaucoup de patience, répondu à mes innombrables questions sur deformetrica. Je remercie également toute l'équipe de doctorants de Stéphanie. Rémi, Juliette et Thomas qui m'ont accueilli à mon arrivée. Je pense aussi à Clément, Solange, Fleur, le petit Clément et plus récemment Pierre et Louis. Merci pour ces journées aux Cordeliers à discuter de maths (ou autre) et toutes ces visio-conférences, en période de confinement, occupées à photoshopper des photos de Clément ! Finalement, Tom, j'ai eu beaucoup de plaisir à travailler avec toi et tes innombrables données médicales. Merci d'avoir toujours été aussi motivé, même après avoir passé une nuit blanche de garde !
J'ai également eu le plaisir et la chance de travailler avec plusieurs médecins. Un grand merci à Olivier Pellerin et Marc Sopoval pour leur confiance et enthousiasme ! Je ressortais de nos réunions toujours plus motivé et, je l'avoue, un peu stressé face à l'impact que pourrait avoir mes travaux.
Je remercie aussi pour Arthur et Anne Sophie pour tous ces vendredi passés à enseigner aux Bachelors et surtout pour tous ces moments passés à s'en plaindre !

Finalement, je remercie aussi Nicolas Villain et Olivier Nempont pour leur disponibilité et leur aide à la segmentation d'IRM de foies.

Si la thèse est une grande expérience professionnelle, il ne me faut pas oublier mes amis et ma famille qui m'ont soutenus tout au long de ces trois années.

Tout d'abord, merci à tous ceux qui ont participé à la bonne ambiance au CMAP: Kevish, Felipe, Benoit, Corentin, Claire, Apolline, Dominik, Louis, Pablo, Thomas et j'en oublie sûrement ! Je tiens également à remercier tous mes amis de l'ENS. Merci à Alexandre, Antoine, Timothée, Nicolas et Shmuel pour ces soirées jeux de société en présentiel ou à distance. Merci aussi à Julien et Alexandra pour toutes ces sorties ciné, pique-niques et bières partagées ! Je pense aussi bien sûr à Julie, William et Cheikh qui m'ont bien aidé à tenir ces mois de confinements lors de nos sorties à "environ" moins d'un kilomètre ! Merci à Julie pour toutes ces pauses à 15h qui duraient bien trop longtemps. Merci à Cheikh pour les sorties le dimanche et son café imbuvable. Par contre, je ne remercie pas Julie pour toutes les souffrances endurées lors de ses séances de sport !!
Un grand merci à tous mes amis Faidherbards: Thomas, Cecile, mon amante, Valentin et Corentin. C'est toujours un plaisir de se retrouver tous ensemble en weekend à Grenoble ou en sortie sur Paris. Merci aussi bien sûr à mes deux colocs préférées, Camille et Marie-Anne qui m'ont supporté pendant ces trois années. Merci pour toutes ces soirées apéros, séances de sport et bons moments passés ensemble !
Bien sûr, je remercie aussi ceux que je retrouvais à chacun de mes retours dans le Nord dans ma maison secondaire à Noeux (maintenant Béthune !). Geraud, ça fait maintenant 17 ans que l'on se connait. Merci pour toutes ces années d'amitiés, ces nombreux weekend sur Paris et mes innombrables passages dans le Nord ! Vincent, merci pour toutes ces discussions plus ou moins philosophiques sur des sujets plus variés les uns que les autres !

Finalement, je me tourne bien évidemment vers ma famille. Papa, maman, merci de m'avoir toujours soutenu au cours de mes (longues) études ! Merci aussi à mon frère Damien et ma soeur Marie et bien sûr mes filleuls, neveu et nièce: Henri, Ambre, Victor et Baptiste. Vous illuminez chacun de mes retours dans le Nord !

# Part I

# Introduction

# CHAPTER 1

# Résumé en Français

## Abstract

L'émergence de vastes data sets longitudinaux (sujets observés de manière répété à différents temps) a permis la construction de différents modèles améliorant la compréhension de phénomènes biologiques. Les études longitudinales ont de nombreuses applications : compréhension des différences de progression pour des maladies neuro-dégénératives comme la maladie d'Alzheimer, suivi de chimiothérapie, reconnaissance faciale,... Pour traiter ces data sets, des modèles hiérarchiques à effets mixtes ont été développés. Ces modèles séparent les paramètres de population (appelés effets fixes) des variables aléatoires propres à chaque sujet. Un but possible est de créer un atlas i.e. d'estimer une image, une forme ou même une trajectoire représentative de la population ainsi que la variabilité inter-individu. Il est alors, par exemple, possible de mettre en évidence l'influence d'une maladie sur un organe particulier par rapport à un vieillissement normal. Quant aux variables aléatoires, elles nous renseignent sur la variabilité des sujets. Elles peuvent être utilisées pour reconstruire la trajectoire de chaque patient, prédire son avenir, émettre un diagnostic, ...

En pratique, différents types de données sont étudiées dans le domaine médical. Des images en deux ou trois dimensions sont obtenues à partir de scanners, d'IRM, de radiographies, etc. Parfois, le maillage d'un organe est extrait. Pour gérer ces différentes structures, un cadre commun a été développé. En utilisant des notions de géométrie riemannienne et plus particulièrement le Large Deformation Diffeomorphic Metric Mapping (LDDMM), il est possible de calculer des distances, des déformations et des moyennes. Ces modèles ont d'abord été développés pour des ensembles de données transversales et ont ensuite été généralisés à des data sets longitudinaux. Ils possèdent cependant plusieurs limitations que nous présentons ici et que nous allons surmonter dans cette thèse.

Tout d'abord, dans le cas longitudinal, ces modèles supposent que la trajectoire de la population représentative suit une dynamique unidirectionnelle. Bien que correcte dans certains cas (atrophie de l'hippocampe par exemple), cette hypothèse n'est pas vérifiée dans d'autres. Dans

le cadre du suivi de chimiothérapie par exemple, une tumeur commence souvent par rétrécir avant de s'agrandir à nouveau. Une trajectoire unidirectionnelle ne peut donc pas expliquer ce comportement. Pour surmonter ce problème, nous proposerons un nouveau modèle hiérarchique à effets mixtes permettant de considérer des populations à branchement.

L'un des avantages du LDDMM est son utilisation de déformations difféomorphiques. Toutefois, cet avantage peut également poser un problème pour certains types de données. En effet, pour un ensemble d'images issues du suivi de chimiothérapie, chaque patient peut avoir un nombre de tumeurs différent. Mais, ces patients étant considérés comme une déformation difféomorphe d'une unique image représentative, leur reconstruction aura toujours le même nombre de tumeurs. Dans cette thèse, nous proposons de représenter les résidus de reconstruction par une matrice creuse, nous permettant de détecter et récupérer des anomalies, comme des tumeurs par exemple, sur une observation.

En plus de fournir de nouveaux modèles, nous nous concentrons également sur l'estimation des paramètres des modèles à effets mixtes utilisés. En pratique, des algorithmes de Stochastic Approximation Expectation Maximization (SAEM) couplés à des méthodes de Monte Carlo par chaîne de Markov sont utilisés. Nous allons assouplir deux hypothèses de ces algorithmes. La première est la nécessité d'une ergodicité géométrique de la chaîne de Markov, nous empêchant de considérer des distributions à queues lourdes. Nous allons assouplir cette hypothèse en proposant un nouvel ensemble d'hypothèses, ne demandant qu'une ergodicité sous-géométrique. De plus, l'algorithme SAEM requiert que la distribution jointe du modèle à effets mixtes appartienne à la famille exponentielle courbe. Cette hypothèse est en fait un obstacle dans de nombreuses situations. Ici, nous étudions une idée proposée par Kuhn and Lavielle (2005) permettant de considérer des distributions qui ne vérifient pas cette hypothèse. Nous montrons cependant que cette astuce peut introduire un biais dans l'estimation et proposons un nouvel algorithme permettant de le réduire.

## Plan de la thèse

Le manuscrit de cette thèse est séparé en deux parties. Dans la première, nous nous concentrons sur la modélisation de données médicales. Des données longitudinales et transversales seront étudiées dans deux buts différents : classifier des données longitudinales à plusieurs dynamiques et identifier des anomalies (telles que les tumeurs) dans des organes. Dans la deuxième partie, nous étudions différentes propriétés théoriques des algorithmes d'approximation stochastique et d'Expectation Maximization. Les quatre différents chapitres constituant ce manuscrit sont résumés ci-dessous.

- **Chapitre 3:** *Apprentissage de la classification de données longitudinales en un mélange de trajectoires indépendantes ou branchantes.*

  Dans ce chapitre, nous étudions des données longitudinales. L'objectif d'un atlas longitudinal est la création d'une trajectoire représentative de la population ainsi que des déformations vers chaque sujet. Les atlas longitudinaux, tels qu'ils sont présentés dans Schiratti et al. (2015), ont cependant l'inconvénient de modéliser la trajectoire représentative comme une géodésique. Toutefois, dans de nombreuses situations pratiques, il ne s'agit pas d'une hypothèse valable. Elle n'est par exemple pas vérifiée dans le cas d'une chimiothérapie lorsque la tumeur devient résistante au traitement après un certain temps.

Pour surmonter ce problème, nous modélisons la trajectoire représentative comme une géodésique par morceaux. Cette idée a été introduite pour la première fois dans Chevallier et al. (2017) où les auteurs proposent un tel modèle pour des données scalaires.
Dans ce premier chapitre, nous généralisons cette discussion en dimension plus grande en introduisant des temps de rupture auxquels la trajectoire représentative passe d'une dynamique à une autre. La modélisation de la trajectoire représentative par une géodésique par morceaux nous permet également d'envisager des ensembles de données plus complexes. En effet, nous pouvons supposer que la population est séparée en différents groupes dont les trajectoires représentatives respectives se ramifient ou se rejoignent à différents temps de rupture. C'est en partie ce qui nous motive à introduire un modèle de classification non supervisée. Ce nouveau modèle est présenté au chapitre 3 et est appliqué à différents jeux de données tels que le score RECIST dans le cas de la chimiothérapie ou le maillage de l'hippocampe dans le cas de la maladie d'Alzheimer.
Ce travail a été publié dans l'International Journal of Computer Vision (Debavelaere et al., 2020).

- **Chapitre 4:** *Détection d'anomalies dans le cadre LDDMM*

Dans ce chapitre, nous nous intéressons à la détection d'anomalies, telles que la présence de tumeurs, dans une image médicale. Plus précisément, nous nous plaçons dans le cadre transversal et supposons que nous avons à notre disposition un template de sujets témoins. Nous définissons alors une anomalie comme une structure qui ne peut pas être récupérée comme une déformation difféomorphe du template témoin.
Par exemple, dans le cas de la détection de tumeurs, le template témoin n'aura pas de tumeur et les déformations difféomorphes de ce template n'auront pas non plus de tumeur. Nous sommes donc capables de récupérer ces tumeurs dans les résidus de la déformation.
Nous montrons que ce procédé améliore la reconstruction des observations et permet effectivement de détecter les anomalies.
En particulier, notre méthode présente l'avantage de ne pas nécessiter de grands ensembles de données ou d'annotations par les médecins. De plus, elle peut être facilement appliquée à n'importe quel organe. Pour mettre en évidence ces avantages, nous appliquons cette méthode à deux jeux de données différents : un data set de foies de patients atteints de cancer colorectal métastatique et un data set de cerveaux atteints de gliomes.
Ce chapitre sera converti en un article à être soumis.

- **Chapitre 5:** *Sur la convergence des approximations stochastiques sous une dynamique de Markov sous-géométrique.*

Les théorèmes assurant la convergence des approximations stochastiques sous dynamique markovienne supposent que la chaîne de Markov est géométriquement ergodique. Dans le chapitres 3, nous utiliserons ces algorithmes avec une dynamique markovienne obtenue à partir d'algorithmes de Metropolis Hastings. Cependant, nous savons que, lorsque des distributions à queues lourdes sont ciblées, ces chaînes de Markov peuvent être sous-géométriques (Douc et al., 2004; Fort and Moulines, 2000, 2003; Jarner and Hansen, 2000). Ainsi, les garanties théoriques de convergence de ces approximations stochastiques ne sont plus respectées.

C'est pourquoi, dans le chapitre 5, nous choisissons de lever la condition d'ergodicité géométrique. Nous proposons un ensemble d'hypothèses plus générales, sous lesquelles nous prouvons la convergence des approximations stochastiques avec dynamique sous-géométrique. Ces hypothèses concernent essentiellement la vitesse de convergence de la chaîne de Markov et la régularité de son noyau. En particulier, la plupart des vitesses de convergence polynomiales satisfont ces hypothèses. Nous prouvons ensuite la convergence de deux approximations stochastiques dans ce cadre. Tout d'abord, nous étudions la convergence d'un algorithme de Metropolis Hastings dans le cas où la variance de la proposition est adaptée au long des itérations. Dans le second exemple, nous considérons un modèle d'analyse en composantes indépendantes dans le cas où des distributions à queues lourdes positives conduisent à une chaîne de Markov ergodique sous-géométrique dans un algorithme SAEM-MCMC.
Ce travail a été publié dans l'Electronic Journal of Statistics (Debavelaere et al., 2021).

- **Chapter 6:** *Famille courbe exponentielle et algorithme de Stochastic Approximation Expectation Maximization Algorithm*
  Parmi les hypothèses assurant la convergence des algorithmes SAEM et MCMC-SAEM, l'une des plus restrictives est la nécessité que la vraisemblance jointe appartienne à la famille exponentielle courbe. Cependant, cette hypothèse n'est pas toujours vérifiée (par exemple pour les modèles hétéroscédastiques). Dans ce cas, Kuhn and Lavielle (2005) propose de transformer le modèle statistique pour le rendre exponentiel. Leur solution consiste à considérer les paramètres $\theta$ du modèle initial comme des variables latentes supplémentaires suivant une distribution Gaussienne centrée sur un nouveau paramètre $\bar{\theta}$ et avec une variance fixée $\sigma^2$. Au lieu d'estimer $\theta$, nous estimons alors sa moyenne $\bar{\theta}$. Si cette méthode est souvent utilisée, il n'y a en fait aucune garantie que $\bar{\theta}$ est proche du paramètre du modèle initial.
  Dans le chapitre 6, nous montrons que l'utilisation de cette méthode peut introduire un biais dans l'estimation du maximum de vraisemblance. Nous prouvons ensuite que ce biais tend vers zéro lorsque la variance $\sigma$ tends vers 0 et donnons une borne supérieure pour $\sigma$ petit. Cependant, sur un exemple numérique, nous voyons qu'un compromis doit être fait entre l'erreur d'estimation et le temps de calcul. Pire encore, pour de très petites valeurs de $\sigma$ (et donc, théoriquement, de petites erreurs), l'algorithme ne converge pas numériquement. Pour surmonter ce problème, nous proposons dans la dernière partie du chapitre un nouvel algorithme permettant une meilleure estimation du maximum de vraisemblance en un temps de calcul raisonnable.
  Ce travail a été soumis (Debavelaere and Allassonnière, 2021).

# CHAPTER 2

# Introduction

In this chapter, we introduce the different notions needed in this thesis. After explaining the motivations behind this thesis and a quick introduction to mixed effect models, we begin by presenting, section 2.3, the Riemannian notions necessary to the understanding of models processing data living on Riemannian manifolds that we will use or generalize later on (section 2.4). To be able to present the algorithms (section 2.7) necessary to estimate the parameters of those statistical models, we also present some notions on Markov Chains (section 2.5) and Stochastic Approximations (section 2.6).

## Contents

## 2.1 Motivation

Numerous scientific fields require to study the mean behaviour of a population, whether along time (subjects observed at different time points, forming a *longitudinal* data set) or not (one observation per subject, forming a *cross-sectional* data set). For instance, the study of a data set of organs can highlight the influence of a particular disease, allow to produce a diagnosis and play a crucial role in the understanding of the disease. Similarly, the study of the temporal progression of a phenomenon helps in the development of new treatments and in the prediction of the future evolution of new patients.

One of the easiest examples of such studies is the growth of children. Development and growth studies have provided representative scenarios of weight or height evolution with time. Those scenarios give an average trajectory of progression which describes the typical evolution of height and weight of a child. They also give the variability in the population in the form of confidence intervals.

The data from such phenomena come in lots of different formats. The features may be real numbers such as heights, weights or cognitive scores. Increasingly often, in medicine, the data come from images. For instance, those images are used in neurological monitoring, oncology, traumatology, etc.. From those images, different features may be extracted. One can, for instance, measure the size of the tumors on an organ. Another possibility is to extract structures from those images: meshes can represent the surface of an anatomical shape like the hippocampus or the cortex ribbon. Finally, sometimes, the whole image is used as a feature: images from scanner, Magnetic Resonance Imaging (MRI), X-rays, etc..

The space of measurements to which the data belongs to is typically defined by smooth constraints and may not behave as a Euclidian space. For instance, operations like addition or scaling do not make sense for images or meshes. Hence, a new structure is needed to transpose these intuitive operations to more complex objects. This is the goal of Riemannian manifolds (Lafontaine et al., 2004; Younes, 2010). In such manifolds, one is able to define distances, means or even do statistics. The distance between two objects is then defined as the minimum length of the curves going from one point to the other in the Riemannian space. This curve of minimal length defines the deformation between the two objects.

For medical data, it is in particular crucial to conserve the structure of the objects studied. For instance, if we want to compute the distance between two livers by deforming one onto the other, we do not want this deformation to make holes appear. We do not want either to fold the liver over itself. The Large Diffeomorphic Deformation Metric Mapping (LDDMM) (Trouvé, 1995; Dupuis et al., 1998) has been developed in this perspective. In this framework, the distance between two objects is computed as the difficulty to obtain one as a diffeomorphic deformation of the other. The use of diffeomorphisms prevents the problems evoked above.

This framework will allow us to create atlases. An atlas of a data set is composed of an object that is representative of the population, and of the variability within this population. To this end, we introduce hierarchical mixed effect models (Lavielle and Mentré, 2007). Those statistical models analyze a population as being driven by two levels of effects. First, the model is parametric and the population parameters (called *fixed effects*) describe the population globally. Then, a second level of description is given by random variables (called *random effects*) explaining the subjects variability inside this population. The individual variables are random

variables on which we specify distributions while the fixed effects can be estimated using maximum likelihood estimates. By fitting a hierarchical mixed effect model, one can thus learn an average model and derive from it an individual-specific model. They are generative statistical models whose parameters can, most of the time, be easily interpreted.

Combining those two mathematical frameworks, our models aim at explaining the population by a representative shape (in the case of cross-sectional data sets) or trajectory (in the case of longitudinal data sets). The generative mixed effect models also allows us to reconstruct each subject from this representative object using the LDDMM framework. This in particular, allows us to obtain the population variability. Given this representative object and population variability, it is then possible to fit new subjects and observe their position in the population variability. For example, one is able to observe if a new subject is close to the representative object or in the tail of the distribution. In that last case, it could be the sign of a pathology. In the longitudinal case, it is also possible to predict the future of a patient. This can allow to issue an early diagnosis or to adapt a treatment.

The use of the LDDMM framework in mixed effect models has already been studied in various cases: cross-sectional atlases for images (Miller et al., 2002; Durrleman et al., 2011a), shapes (Durrleman et al., 2014) but also for longitudinal atlases (Muralidharan and Fletcher, 2012; Singh et al., 2016; Bône et al., 2018a). In this last case, the representative trajectory is modeled by a geodesic and a temporal variability is added to the spatial variability: each subject can indeed have its own pace of progression and offset in addition to its own anatomical evolution.

To estimate the parameters of mixed effect models, Expectation Maximization algorithms are used. Those algorithms estimate a maximum of likelihood when some variables are not observed (in our case, the individual random effects). They consist in the iteration of the computation of an expectation and a maximization. In most cases, the expectation step is in fact intractable. Stochastic Approximations (SA) coupled with Monte Carlo Markov Chains are then used to compute it.

Stochastic approximations are used to find roots of a function $h(\theta) = \mathbb{E}_\theta(H_\theta(X))$ from which we only have noisy observations. Using Markovian dynamics, one is able to find the solutions of such an equation (Andrieu et al., 2005). This framework can hence be used to compute the expectation step of EM algorithms (Delyon et al., 1999; Allassonnière et al., 2010).

If those different subjects have already been extensively studied, several points still necessitate further investigations.

First, concerning longitudinal models using the LDDMM framework, not all data sets can be correctly studied by the models mentioned above. Indeed, because they model the representative trajectory by a geodesic, they have the drawback to only allow a unique global dynamic. If it is a valid assumption for lots of real cases (atrophy of the hippocampus, evolution of cognitive scores for instance) it does not hold for others. For instance, in the case of chemotherapy monitoring, a tumor often begins by shrinking before becoming resistant to the treatment and growing again. Such a global behaviour cannot be modeled by a geodesic as two different dynamics follow one another. Here, we generalize a model modeling the representative trajectory by a piecewise geodesic. This model has first been introduced in Chevallier et al. (2017) for one dimensional data sets. We apply it here in the multi-dimensional case. In particular, we process

more complex data sets with mixture of populations following different dynamics along time and branching or joining in sub-populations.

In a second time, we apply the cross-sectional models already developed in a new setting. By creating a representative image of normal subjects, one can look for anomalies in the diffeomorphic reconstruction of new subjects. This is particularly useful for patients with cancer. Indeed, in that case, the tumors can be visible as new structures with different grey level textures on the image. But, the reconstruction of a particular subject being a diffeomorphic deformation of a representative shape without tumors, they will not appear on it. By splitting up the residuals between noise and a sparse matrix, one can retrieve those tumors. Such methods have already been studied in the framework of deep learning (Baur et al., 2018; You et al., 2019; Pawlowski et al., 2018; Yu et al., 2019). However, those methods require a big amount of data and, often, annotations from doctors. Our method has the particular advantage to be easily generalized to any organ, even if one only has few observations and no annotation.

As explained above, the Expectation Maximization algorithm used to estimate the parameters of our mixed effect models uses the framework of Stochastic Approximations coupled with Monte Carlo Markov Chain methods (SAEM-MCMC algorithm). The usual assumptions ensuring the convergence of such SA require that the subjacent Markov Chain is geometrically ergodic (i.e. converges towards its invariant distribution at a geometric rate). This can be a bottleneck in certain applications. The Metropolis Hastings algorithm is a commonly used MCMC sampler. However, when targeting distributions with heavy tails, it can produce sub-geometric ergodic chains. In that case, we no longer have any guarantee of convergence of the SAEM-MCMC algorithm. We thus choose to study the case where the Markov Chain is not geometric ergodic. We propose a new theorem of convergence of Stochastic Approximations with only subgeometric ergodicity allowing, in particular, to apply the SAEM-MCMC in a broader range of cases.

Finally, the SAEM-MCMC algorithm has the disadvantage to require the joint distribution to belong to the curved exponential family. In practice, this is rarely the case, neither for heteroscedastic models nor for most non linear models. A usual trick is to transform the model to make it curved exponential (Kuhn and Lavielle, 2005). It is this particular trick that we will use further on in our models. However, the model being changed, there is no guarantee that the estimated maximum likelihood of this new model is close to the initial one. In fact, we will show that a bias is introduced by this method and propose a new algorithm allowing to reduce it.

In the next sections, we will introduce the different notions evoked here and needed to the understanding of this thesis. Those introductions are succinct but references to more complete books or articles are given.

## 2.2 Mixed effects models

Mixed effects models explain observations through two types of effects: the *fixed effects* are shared by the whole population while the *random effects* are specific to each subject. This type of statistical model is particularly helpful for hierarchical generative models. Although very generic, we will focus here on longitudinal data analysis through these models. Given observations $(y_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$ at times $(t_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$, the model writes:

$$y_{i,j} = f(t_{i,j}, \beta, z_i) + \varepsilon_{i,j} \tag{2.1}$$

where $\beta$ is the vector of parameters, called fixed effects, $(z_i)_{1\leq i\leq n}$ are the random effects and $\varepsilon_{i,j}$ is a random variable representing the noise.

One of the easiest examples of those models is the random slope and intercept model (Cohen et al., 2013) which takes in scalar longitudinal data . Given an initial time $t_0$, this model is linear and represents the observations as:

$$y_{i,j} = (a + a_i)(t_{i,j} - t_0) + (b + b_i) + \varepsilon_{i,j} \, .$$

The vector of fixed effects is $\beta = (a, b)$ while, for each subject $i$, the random effect are $z_i = (a_i, b_i)$ and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ is the noise.
This model reflects the evolution dynamic of the population as a line: $t \mapsto a(t-t_0)+b$ and allows variability in the population by adjusting the slope and intersect for each subject.
By supposing $a_i \sim \mathcal{N}(0, \sigma_a^2)$ and $b_i \sim \mathcal{N}(0, \sigma_b^2)$, one can estimate the fixed effects using, for instance, Expectation Maximization algorithms (see section 2.7).

More generally, non linear mixed effects are often used and one can refer, among many other works, to Sheiner and Beal (1980); Bates and Watts (1988). As in the previous example, those models estimate the fixed effects explaining the population dynamic and the random effects modeling the variability in the population. Such models will be studied section 2.4 and in the following chapters. One can find a complete review of mixed effect models in Lavielle and Mentré (2007).

## 2.3 Riemannian notions

We now want to present the mixed effect models we will use later on. However, to do so, we first need some notions of Riemannian geometry. We will present them only succinctly here and we refer to Lafontaine et al. (2004); Younes (2010) for more details.

### 2.3.1 Riemannian metric and Exponential map

The structure of a Riemannian manifold is given by its metric. It is defined by a continuous collection of dot products $< \cdot \,|\, \cdot >_x$ on the tangent space $T_xM$ at each point $x \in M$.
Let $\gamma : [0, 1] \to M$ be a curve on the manifold. One can then compute the length of this curve between $t = a$ and $t = b$ as:

$$\mathcal{L}_a^b(\gamma) = \int_a^b (< \dot{\gamma}(t) \,|\, \dot{\gamma}(t) >_{\gamma(t)})^{1/2} \, dt \, .$$

We then define the distance between two points $x$ and $y$ of $M$ as the minimum length among the smooth curves joining $x$ and $y$:

$$d(x, y) = min \left\{ \mathcal{L}_0^1(\gamma) \,|\, \gamma : [0, 1) \to M, \gamma(0) = x \text{ and } \gamma(1) = y \right\} \, .$$

The curves realizing the minimum of length are called geodesics. It is possible to compute them by solving a second order differential system.
We say that the manifold is geodesically complete if the definition domain of all geodesics can be extended to $\mathbb{R}$. This means that the manifold has no boundary nor any singular point that we can reach in a finite time. In particular, such a manifold is complete for the induced distance and there always exists at least one minimizing geodesic between any two points of the manifold. In

the following, we suppose that $M$ is geodesically complete.

Since a geodesic is the solution of a second order differential equation, there exists one and only one geodesic $\gamma_{x,v}$ going through the point $x \in M$ at time $t = t_0$ with tangent vector $v \in T_xM$. We call Riemannian exponential the application mapping each vector $v$ to the value of the associated geodesic at time $t$:

$$\begin{aligned} \mathcal{E}xp_{x,t_0,t} : T_xM &\to M \\ v &\mapsto \gamma_{x,v}(t) \end{aligned}$$

This exponential map will be of the upmost importance in the next sections as we will use it to compute deformations of a shape $x$ towards our different subjects.

### 2.3.2 Exp-parallelization

We now want to generalize the notion of parallels to Riemannian manifolds. In the following, this notion will help us define spatial deformations. Let $M$ be a geodesically complete Riemannian manifold, $\gamma : I \subset \mathbb{R} \to M$ a differentiable curve on $M$ and $\omega \in T_{\gamma(t_0)}M$ a tangent vector. Given $t_0, t \in [0, 1]$, we first define the parallel transport of $w$ from $\gamma(t_0)$ to $\gamma(t)$ along $\gamma$: $P_{t_0,t}(w) \in T_{\gamma(t)}M$. We recall that this mapping is uniquely defined by the integration from $u = t_0$ to $t$ of the differential equation $\nabla_{\dot{\gamma}(u)} P_{t_0,u}(w) = 0$ with the initial condition $P_{t_0,t_0}(w) = w$ where $\nabla$ is the Levi-Civita covariant derivative.

We then define the exp-parallel variation of $\gamma$ along $\omega$ as the curve $\eta^w(\gamma; .) : I \to M$:

$$\eta^w(\gamma, .) : t \mapsto \mathcal{E}xp_{\gamma(t),0,1}(P_{\gamma,t_0,t}(\omega)) \tag{2.2}$$

Hence, to obtain the parallel of $\gamma(t)$ at the instant $t$, we first compute the parallel transport of $\omega$ on $\gamma$ and use the resulted vector in the Riemannian Exponential. This process is summarized on a sphere Figure 2.1 and a numerical process to compute it has been introduced in Louis et al. (2017).

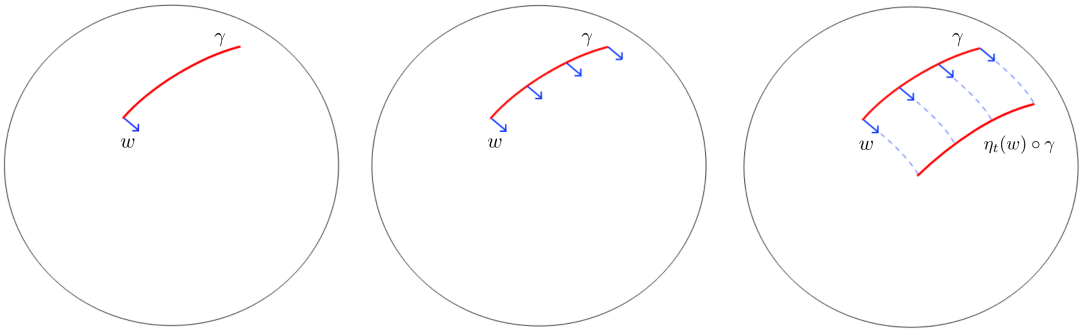



**Figure 2.1** Example of parallel transport on a sphere. On the left, we draw a trajectory $\gamma$ and the momenta to transport $w$. On the center, we transport $w$ along $\gamma$. On the right, we compute the exp-parallelization of $\gamma$ by $w$.

### 2.3.3 Fréchet mean

Given a data set, a usual question is to define its mean. However, on a Riemannian manifold, we cannot just take the arithmetic mean of the observations as it would not always belong to this manifold (this problem is particularly noticeable for data living on a sphere). To generalize the notion of mean on a Riemannian manifold, we introduce the notion of Fréchet mean. Given observations $(x_i)_{1\leq i\leq n}$ in $M$ and a distance $d$, we define:

$$\bar{x} = \text{argmin}_{y\in M} \frac{1}{n}\sum_{i=1}^{n} d(x_i, y)\,. \tag{2.3}$$

This new notion generalizes the usual mean on Riemannian manifolds. In particular, on a Euclidian space, the usual mean is also solution of 2.3.

The existence and uniqueness of the Fréchet mean have notably been studied by Karcher (1977), Kendall (1990) or Bhattacharya et al. (2003) and we send back to those papers to find conditions on the manifold ensuring the existence and uniqueness of the Fréchet mean.

### 2.3.4 Statistics on a Riemannian manifold

Different statistical models have been studied on Riemannian manifolds using the notions presented above. Theoretical results on the Fréchet mean and covariance matrix of random elements have been studied in Pennec (2006). Those properties have for instance been applied by Boisvert et al. (2006) for the variability analysis of the scoliotic spine shape where deformations belong to the Riemannian space of rigid transformations. Similarly, the Fréchet mean for different distances on the set of positive definite matrices has been studied in Dryden et al. (2009) to do statistics on covariance matrices. The Fréchet mean is also, for instance, used in Vercauteren et al. (2005) to find a globally consistent mapping of input frames to a common coordinate system.

Another point of interest from the statistical point of view has been the generalization of Principal Components Analysis (PCA) on Riemannian manifolds. The simplest generalization is the tangent PCA and consists in unfolding the whole distribution in the tangent space at the mean, and computing the principal components of the covariance matrix in the tangent space. Other possibilities consist in considering geodesic spaces (Fletcher et al., 2004) or barycentric subspaces (Pennec et al., 2018).

In the next section, we will present the Large Deformation Diffeomorphic Metric Mapping framework which will use most of the tools presented above and be used in part II of this manuscript. This framework uses the Fréchet mean and the Exponential map on the manifold space of shapes to construct a hierarchical model.

## 2.4 Large Deformation Diffeomorphic Metric Mapping

Now that we have presented the Riemannian geometry notions needed, we can introduce the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework we will use in our thesis.

### 2.4.1 First notions of shape spaces

In 1942, d'Arcy et al. (1917) introduced what would later on be generalized as shape spaces. Instead of considering shapes as distinct objects, his idea was to study the deformations allowing to go from one anatomical shape to another (see figure 2.2).

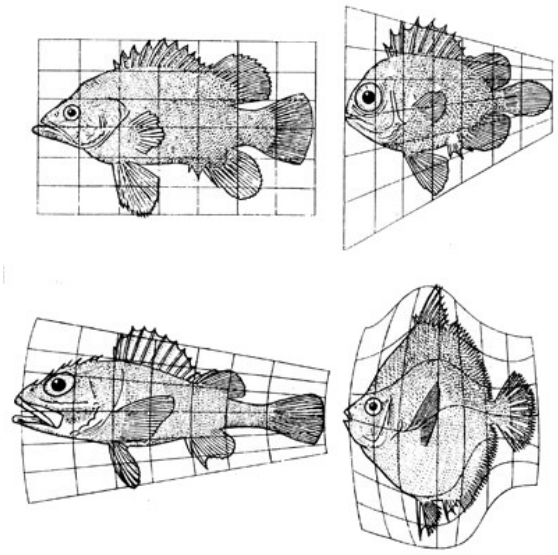


**Figure 2.2** Illustration taken from the book On Growth and Form d'Arcy et al. (1917).

Grenander (1993) has then mathematically formalized this idea. The idea is the following. We consider a shape space $M$. Here the term "shape" designs any structured data such as meshes, images, etc.. We suppose that there exists $\mathcal{G}$ a group acting transitively on $M$. We then define a shape space by considering the unique orbit of the action $\mathcal{G}.x_0$ for $x_0$ in $M$. This shape space is constituted of all the deformations of the shape $x_0$ by the group $\mathcal{G}$. In this thesis, we will mainly consider the case where $\mathcal{G}$ is a subset of the set of diffeomorphisms on $\mathbb{R}^n$.

We can already give an example of such shape spaces, composed of landmarks. We call landmark a labelled set of points. The set of all landmarks of $\mathbb{R}^n$ of size $p \in \mathbb{N}$ is:

$$M = \{x = (x_1, ..., x_p) \in (\mathbb{R}^n)^p \,|\, \forall i \neq j, x_i \neq x_j\} \ .$$

Since the points are labelled, the landmarks are easily handled. In particular, we can easily make $\mathcal{C}^1(\mathbb{R}^n)$ act on $M$ by setting, for $g \in \mathcal{C}^1(\mathbb{R}^n)$, $x \in M$,

$$g.x = (g(x_1), ..., g(x_p)) \, .$$

The landmarks shape space has in particular been studied in Kendall (1984).

### 2.4.2 Measuring the distance between two shapes

Now that we have defined our shape space as $\mathcal{G}.x_0$ with $x_0 \in M$, we can interest ourselves in the creation of a distance between two shapes. Following the idea of d'Arcy et al. (1917), we

will consider this distance as the difficulty to deform one shape onto another. More precisely, we suppose that $\mathcal{G}$ is endowed with a right invariant metric $d_{\mathcal{G}}$. We can then endow $M$ with the following pseudo metric: for any $x, y \in M$, we set:

$$d(x,y) = \inf \{ d_{\mathcal{G}}(Id, g) \,|\, g.x = y \} \ . \tag{2.4}$$

If there is no deformation from $x$ to $y$, $g = Id$ and $d(x,y) = 0$. In general, $d(x,y)$ is computed using the minimal deformation transforming $x$ into $y$.

The question is now to define the group $\mathcal{G}$ and the right invariant distance $d_{\mathcal{G}}$. The first attempt to construct such structures in the medical image setting has been done using displacement vector fields by Broit (1981). With $\Omega \subset \mathbb{R}^n$, one can consider a displacement vector field $u : \Omega \to \mathbb{R}^n$ and set $\phi_u : x \mapsto x + u(x)$ and $\phi_u^{-1} : x \mapsto x - u(x)$. $\mathcal{G}$ is then the group of all such maps and acts on an image $I_0$ by: $\phi.I_0(x) = I_0(\phi^{-1}(x))$. We then define a distance on $\mathcal{G}$ measuring the smoothness of the displacement field by setting:

$$d_{\mathcal{G}}(Id, \phi_u) = || - \alpha \Delta u + \gamma u||_{L^2} \ .$$

This distance on $\mathcal{G}$ then allows us to compute a distance between images using equation 2.4.

However, this small deformation approach presents some limitations. First, it does not ensure a one to one transformation. It has also been showed in Christensen (1994) that it can sometimes generate transformations folding the grid over itself and so destroying the neighbourhood structure. Moreover, it does not always generate invertible transformations. To overcome those drawbacks, Trouvé (1995), Dupuis et al. (1998) and Beg et al. (2005) chose to proceed infinitesimally by introducing the Large Deformation Diffeomorphic Metric Mapping (LDDMM). They chose to only consider diffeomorphic transformations. Diffeomorphisms are particularly well adapted here as they are smooth transformations with smooth inverses preserving connected sets and smoothness of anatomical features. We present this framework below.

Let $V$ be a set of vector fields over $\mathbb{R}^n$ endowed with a Hilbert structure and continuously embedded into the space of diffeomorphisms vanishing at infinity and whose differential also vanishes at infinity: $\mathcal{C}_0^1(\mathbb{R}^n)$. We also set $\mathbb{L}^2([0,1], V)$ the set of time dependent vector fields, $\mathbb{L}^2$-integrable with respect to $t$:

$$\mathbb{L}^2([0,1], V) = \left\{ (v_t)_{t\in[0,1]} \,\middle|\, \forall t \in [0,1], v_t \in V \text{ and } \int_0^1 ||v_t||_V^2 \, dt < \infty \right\}$$

Instead of considering a deformation as $Id + v$ as done in the small deformation framework, we suppose that a deformation $\phi$ is obtained as the flow of a vector field $v_t$. More precisely, for $v \in L^2([0,1], V)$, we set $\phi_1^v$ the diffeomorphism obtained as the flow at time $1$ of the vector field $v$:

$$\begin{cases} \partial_t \phi_t^v = v_t \circ \phi_t^v \\ \phi_0^v = Id \, . \end{cases} \tag{2.5}$$

We then set $\mathcal{G} = \{\phi_1^v | v \in \mathbb{L}^2([0,1], V)\}$ the group of such diffeomorphisms. It is now easy to define a distance on $\mathcal{G}$. For $\phi, \phi' \in \mathcal{G}$, we set:

$$d_{\mathcal{G}}(Id, \phi) = \inf \left\{ \left( \int_0^1 ||v_t||_V^2 \, dt \right)^{1/2} \,\middle|\, v \in \mathbb{L}^2([0,1], V) \text{ and } \phi_1^v = \phi \right\}$$

and

$$d_{\mathcal{G}}(\phi, \phi') = d_{\mathcal{G}}(Id, \phi' \circ \phi^{-1}) .$$

It can be remarked that this distance depends on the norm $||.||_V$ we put on $V$. The choice of this norm will be discussed subsection 2.4.4.

This exactly states that $\mathcal{G}$ is given the structure of a manifold on which distances are computed as the length of minimal geodesic paths $(\phi_t^v)_{t\in[0,1]}$ connecting two elements. In particular, it means that we can use the different tools introduced section 2.3.

It has been showed that this infimum is in fact a minimum and that the distance is right invariant (Trouvé, 1995; Younes, 2010). Moreover, a geodesic in $\mathcal{G}$ passing through $Id$ at an initial time $t_0$ is then uniquely defined by an initial velocity $v_0$. In the following, we will write $\mathcal{E}xp_{t_0,t}(v_0)$ the value of this geodesic at the time $t$, as introduced section 2.3 (contrary to the notation introduced in that section, we do not specify the initial point $Id$).

Hence, for two shapes $x$ and $y \in M$, we set:

$$d(x,y) = \inf \left\{ \left( \int_0^1 ||v_t||^2 \, dt \right)^{1/2} \middle| \; v \in \mathbb{L}^2([0,1], V) \text{ and } \phi_1^v.x = y \right\} .$$

This distance measures the shortest length of the path relying $x$ to $y$ using the diffeormophisms $\phi^v$. It also allows to define a Riemannian structure on $M$. A geodesic on $M$ will then be defined using an initial shape $p_0$ and initial velocity $v_0$ by $t \mapsto \mathcal{E}xp_{t_0,t}(v_0)(p_0)$.

This construction answers our previous concerns: we measure the distance between two shapes as the difficulty to deform one onto another. Moreover, as $\phi_1^v$ is a diffeomorphism, it is invertible and preserves the smoothness and structure of the shapes.

### 2.4.3 Matching of two shapes

In the previous subsection, we have defined a distance between two shapes as a cost of deformation. We now ask ourselves how to compute in practice this deformation. To do so, we will use inexact matching by minimizing a function expressing a balance between length of the path and target correspondence. We set, for $\lambda > 0$, $x, y \in M$ and for $v \in \mathbb{L}^2([0,1], V)$,

$$J(v) = \int_0^1 ||v_t||_V^2 \, dt + \lambda A(\phi_1^v.x, y) . \tag{2.6}$$

$A$ is the data attachment term. It measures the distance between the target $y$ and the deformed initial shape: $\phi_1^v.x$. In particular, as $A$ only depends of $\phi_1^v$, we can see that any $v$ minimizing $J$ will also verify

$$v = \text{argmin} \left\{ \left( \int_0^1 ||v_t||^2 \, dt \right)^{1/2} \middle| \; v \in \mathbb{L}^2([0,1], V) \text{ and } \phi_1^v = \phi \right\} .$$

The path $(\phi_t^v)_{t\in[0,1]}$ will thus be a geodesic in the manifold $\mathcal{G}$.

Different data attachment terms exist depending on the nature of the shape observed. For images, the $\mathbb{L}^2$ distance between the two images is used. Similarly, for landmarks, as the points

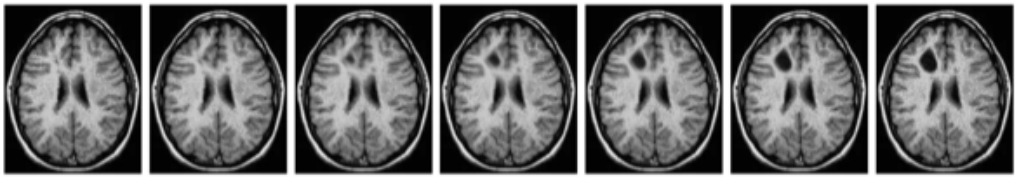
are labelled, we just use a square distance:

$$\forall x = (x_1, ..., x_p) \in (\mathbb{R}^n)^p, y = (y_1, ..., y_p) \in (\mathbb{R}^n)^p, v \in V,\ A(\phi_1^v.x, y) = \sum_{i=1}^{p} ||\phi_1^v(x_i) - y_i||^2 .$$

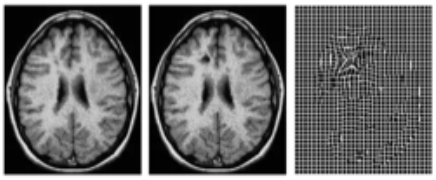
The problem is more complex when the points are not labelled and when two shapes can have a different number of points (meshes for instance). In that case, different attachment terms measuring the distance between two shapes have been created. Using discrete measures, Glaunes et al. (2004) proposes an attachment term for non-labelled points. Vaillant and Glaunès (2005) introduces the notion of currents to match oriented shapes. As for non oriented shapes, Charon and Trouvé (2013) introduces the concept of varifolds.

In the following parts, we will either use varifolds or currents for meshes and the $\mathbb{L}^2$ norm for images. Hence, once we have defined this attachment term, we are able to perform the inexact matching by minimizing the function $J$. To do so, the computation of its gradient has been the object of several articles (see Miller and Younes (2001); Miller et al. (2002); Beg et al. (2005) among others). More recently, the use of automatic differentiation is often privileged. An illustration of a matching and of the deformation field is given Figure 2.3.

6.2.5. *Tumor.* In this next experiment, a brain without a tumor is matched to the same one with a tumor. The tumor appears progressively during the morphing process. One notices large deformations around the tumor, and almost no deformation in other places.



A tumor progressively appearing on a brain



Tumor: Reference image, registered target and deformation

**Figure 2.3** Illustration taken from Miller et al. (2002).

### 2.4.4 Finite parametrization of the vector fields

To finalize our explanation of matching, we still have to define the norm we use on our vector fields $||.||_V$. To do so, Joshi and Miller (2000) and Durrleman et al. (2011a) propose to characterize the vector fields by a finite number of parameters: control points and momenta. They endow $V$ with a Reproducing Kernel Hilbert Space (RKHS) structure. A RKHS is associated with a kernel such that for any function $f$, the operation $f(x)$ can be performed by taking an inner product with a function determined by the kernel $K_V$.

In this RKHS, for $x \in \mathbb{R}^n$, a vector field $v$ is represented as:

$$v(x) = \sum_{i=1}^{n_{cp}} K_V(c_i, x)\alpha_i \tag{2.7}$$

where $(c_i)_{1\leq i\leq n_{cp}}$ are called control points and $(\alpha_i)_{1\leq i\leq n_{cp}}$ are called momenta. $v$ is thus represented as the interpolation of the momenta at the control points using the kernel $K_V$. In practice, we choose $K_V$ to be a Gaussian kernel with variance $\sigma_V^2$: for $x, y \in \mathbb{R}^n$,

$$K_V(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma_V^2}\right).$$

This construction can be seen Figure 2.4 where the control points are represented by red points, the momenta by red arrows and the vector field by blue arrows.

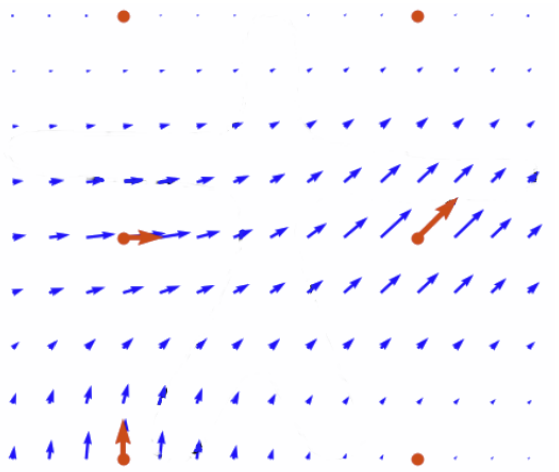


**Figure 2.4** Construction of the vector field $v$ (blue arrows) using the control points (red points) and momenta (red arrows).

With such notations, the length of the path parameterized by $(v_t)_{t\in[0,1]}$ is:

$$\int_0^1 ||v_t||_V^2 \, dt = \int_0^1 \sum_{i,j=1}^{n_{cp}} \alpha_i(t)^t K_V(c_i(t), c_j(t))\alpha_j(t) \, dt \, .$$

We come back to our matching problem. Given two shapes, we are looking for a vector field $(v_t)_{t\in[0,1]}$ minimizing equation (2.6). We will now see that it is enough to define initial control points and momenta at $t = 0$ to be able to compute at any time point the velocity field $(v_t)_{t\in[0,1]}$ minimizing the function $J$.

As explained in the previous subsection, a vector field minimizing $J$ generates a geodesic path in $\mathcal{G}$ beginning at $Id$. Hence, it is enough to look for geodesic paths and so for the initial velocity.

It has been showed in Miller et al. (2006) that, in a RKHS, a geodesic path $(\phi_t^v)_{t\in[0,1]}$ defined by a velocity field $(v_t)_{t\in[0,1]}$ verifies:

$$v_t(x) = \sum_{i=1}^{n_{cp}} K_g(c_i(t), x)\alpha_i(t)\,. \tag{2.8}$$

where the time dependent control points and momenta are solutions of the Hamiltonian equations:

$$\begin{cases} \dot{c}(t) = K_V(t)m(t) \\ \dot{m}(t) = \nabla_{c(t)}\left(m(t)^T K_V(t)m(t)\right) \end{cases} \tag{2.9}$$

with initial conditions $m(0) = (m_k(0))_{1\leq k\leq n_{cp}}$, $c(0) = (c_k(0))_{1\leq k\leq n_{cp}}$ and where $K_V(t)$ is the $n_{cp} \times n_{cp}$ kernel matrix $(K_V(c_i(t), c_j(t)))_{1\leq i,j\leq n_{cp}}$.
Hence, the geodesic path $t \mapsto \mathcal{E}xp_{t_0,t}(v_0)$ is uniquely defined by its initial momenta and control points.

To summarize, given initial control points and momenta, we are able to generate a geodesic path. Hence, to solve our matching problem, we now just have to find those initial vectors minimizing the function $J$. This is to be compared with the previous situations where one had to estimate a vector field $v_0$ on the whole space. We now just have to estimate a finite number of initial vectors to solve the same problem.

An example of a matching using this method is presented Figure 2.5. In part II, it is this decomposition of vector fields using momenta and control points that we will use in our experiments.

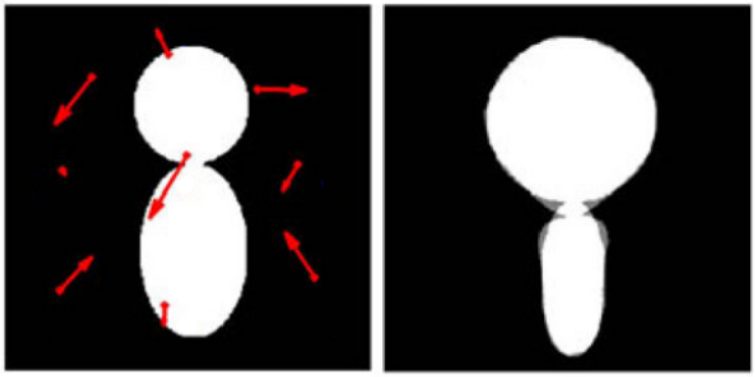


**Figure 2.5** Illustration taken from Durrleman et al. (2011a). The red points are the initial control points, the red arrows, the initial momenta. On the left is the initial shape. On the right is the deformed shape (white) superposed with the target (grey).

### 2.4.5 Cross sectional atlas

Now that we know how to compute a distance between two shapes, we can interest ourselves in the creation of an atlas. An atlas is constituted of a *representative shape* (also called *template*) of the population as well as the deformations from this representative shape towards each

subject. To create this representative shape, we use the concept of the Fréchet mean introduced subsection 2.3.3. This notion has been used on different data sets in numerous papers, see Guimond et al. (1998); Le and Kume (2000); Pennec (2006); Mio et al. (2007); Ma et al. (2008); Allassonnière et al. (2007) among others. We present this approach here.
Given a data set of observations $(y_i)_{1\leq i\leq m}$, we will minimize the function:

$$J(p, v_1, ..., v_m) = \frac{1}{2\sigma^2}\sum_{i=1}^{m} A(\phi_1^{v_i}.p, y_i) + \text{Reg}(p, v_1, ..., v_m)$$

Hence, we are looking for a shape $p$ and for the deformation fields $v_i$ transforming $p$ onto an approximation of each observation. $A$ is the data attachment term and we also add a regularity term allowing, for instance, to add smoothness conditions. The parameter $\sigma$ allows us to balance between the regularity and the attachment to the data desired.

This minimization problem can be solved using gradient descents algorithms. It is also possible to transform it into a Bayesian problem by associating it to a statistical model. The observations $y_i$ are then supposed to follow a normal law centered in $\phi_1^{v_i}.p$ and of variance $\sigma^2$. The regularization terms are composed of the sum of the log-likelihood of $v_i$ (often, $v_i \sim \mathcal{N}(0, \Sigma)$) and of the log-likelihood of priors. In that case, Expectation-Maximization algorithms are used to estimate the parameters of the model (see section 2.7). We no longer estimate $p$ as a Fréchet mean but as a maximum likelihood estimate. It has been showed in Devilliers et al. (2017) that a bias is introduced between those two notions when the data is noised.

An example of a dataset and representative shapes obtained by this method is presented Figure 2.6.

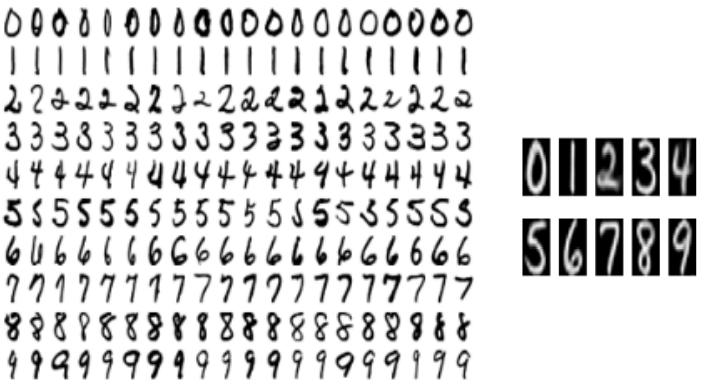


**Figure 2.6** Figure obtained from Allassonnière et al. (2010). On the left, the dataset constituted of numbers from the USPS data set. On the right, the templates obtained for each number using a SAEM algorithm.

### 2.4.6 First longitudinal models

We now interest ourselves in longitudinal data sets. In that case, we are given a data set $(y_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$ of observations of $n$ subjects, each being observed at $k_i$ different times $t_{i,j}$.

We present here different models developed in the literature to tackle that case.

**Geodesic shooting:**
To explain the longitudinal trajectory of a particular subject, Fletcher (2013) considers it as a geodesic. It is the equivalent to a linear regression model but for Riemannian manifolds. Instead of estimating an intercept and a slope, we will estimate a point $p$ on the manifold $M$ and a vector $v \in TM$ the tangent bundle of $M$. More precisely, given a trajectory $(y_j)_{1\leq j\leq m}$ of a particular subject associated to times $(t_j)_{1\leq j\leq m}$, we write:

$$y_j = \mathcal{E}xp_{0,1}(\varepsilon)\left(\mathcal{E}xp_{0,1}(t_j v)(p)\right) ,$$

where $\varepsilon$ is a random variable taking its value in the tangent space of $M$ at the point $\mathcal{E}xp_{0,1}(t_j v)(p)$ and where $\mathcal{E}xp_{0,1}(v)(p)$ designs the value at time $1$ of the unique geodesic in $M$ passing through $p$ at time $0$ with initial speed $v$. This notation means that $y_j$ is written as the deformation of an initial point to which we add a noise.
Fletcher (2013) then uses a least square method to estimate the parameters of the model and applies it to the Corpus Callosum aging. This model reconstructs the trajectory of each subject independently of the others and hence does not give a representative trajectory of the population. To obtain one, the next model presented chooses to obtain the subject trajectories as the deformation of a mean one.

**Shapes-based approach:**
To obtain an atlas from a population $(y_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$, we need to define a hierarchical model. Muralidharan and Fletcher (2012) endow the tangent bundle of $M$ with a Riemannian metric: the Sasaki metric. It allows them to consider the random effects $(p_i, v_i)$ as a geodesic perturbation on the tangent bundle of $M$. More precisely, let $(\alpha, \beta) \in TM$ be the fixed effects and, for $1 \leq i \leq n$, let $(q_i, w_i)$ be a vector in the tangent space of $TM$ at $(\alpha, \beta)$. We then write,

$$\begin{cases} (p_i, v_i) = \mathcal{E}xp_{S_{0,1}}((q_i, w_i))(\alpha, \beta) \\ y_{i,j} = \mathcal{E}xp_{0,1}(\varepsilon_{i,j})\left(\mathcal{E}xp(t_{i,j} v_i)(p_i)\right) \end{cases} \tag{2.10}$$

Here $\mathcal{E}xp_S$ designs the Riemannian Exponential associated to the Sasaki metric on the tangent bundle $TM$ of $M$ and $\varepsilon_{i,j}$ is a random variable taking its value in the tangent space of $M$ at the point $\mathcal{E}xp_{0,1}(t_{i,j} v_i)(p_i)$. This allows them to define a representative trajectory as $\mathcal{E}xp_{0,1}(t_{i,j}\beta)(\alpha)$. This model is once again applied to the Corpus Callosum aging.

**Diffeomorphism-based approach:**
Another way to construct an atlas of a longitudinal data set is to directly look for a geodesic trajectory of diffeomorphisms representing the population dynamic as done in Singh et al. (2016). In this model, each individual trajectory is modeled by a geodesic parameterized by its value and direction at the initial time of observation. Those two random effects are obtained as a perturbation of the fixed effects defining the population geodesic.

This model however has the drawback to heavily depend on the initial times of observation of each subject. A change of this initial time then modifies all the parameters. However, in most practical cases, this initial time has no intrinsic meaning. We will thus look for a model that is robust to change of time origin.

The different models evoked here all have the disadvantage not to take into account the difference of temporality between subjects. Different problems arise from this fact. When we

study the progression of a disease, it is not always possible to know when this disease starts. Moreover, it can develop slower or faster from one patient to another. Thus, constructing a population trajectory by comparing all patients at the same age can result to a bad representation of the mean progression of the disease.

To overcome this problem, Schiratti et al. (2015) proposed to learn a temporal reparameterization together with the spatial deformations. This is the model we present in the next section.

### 2.4.7 Hierarchical spatio-temporal model

Exactly as we have explained how to create a cross-sectional atlas, we now want to create a longitudinal atlas. This time we observe $(y_{i,j})_{1\leq i\leq m, 1\leq j\leq k_i}$ the trajectory of $m$ subjects, each of them being observed at times $(t_{i,j})_{1\leq j\leq k_i}$. In particular, each subject can be observed a different number of times and at different ages.

The goal is to estimate a representative trajectory as well as the distributions of the spatio-temporal deformations from that representative trajectory towards the population of subjects. In Schiratti et al. (2015), the authors propose a hierarchical model where each subject is obtained as a spatio-temporal deformation of a representative curve.

More precisely, the representative curve $\gamma_0$ is now a trajectory of shapes along time. We suppose that it is obtained using a geodesic flow applied on an initial shape. The trajectory of a subject $i$ is then obtained using a spatial deformation $\phi_i$ and a temporal one $\psi_i$. This writes:

$$y_{i,j} = \phi_i \circ \gamma_0(\psi_i(t_{i,j})) + \varepsilon_{i,j}\,, \tag{2.11}$$

where $\varepsilon_{i,j}$ is a Gaussian noise.

As we suppose the representative trajectory to be obtained from a geodesic, it means that we can parameterize it using an initial shape $p_0$ at an initial time $t_0$ with an initial velocity $v_0$. This writes:

$$\gamma_0(t) = \mathcal{E}xp_{t_0,t}(v_0)(p_0)\,, \tag{2.12}$$

where $\mathcal{E}xp$ has been defined subsection 2.4.2 and is the geodesic path in $\mathcal{G}$ with initial velocity $v_0$ at time $t_0$. As explained subsection 2.4.4, the velocity $v_0$ will be obtained as the interpolation of momenta $m_0$ at control points $c_0$. $(t_0, m_0, c_0, p_0)$ are then the fixed effects to estimate.

Concerning the temporal deformation, it must be noticed that each subject can have its own acceleration and time shift. For instance, if we study a disease, it can be declared at a younger or older age and then develop itself more or less quickly. As we want to construct the representative trajectory of the disease, we need to compare the subjects at the same stage. Hence the necessity of a temporal reparameterization. We thus consider for each subject an acceleration parameter $\alpha_i$ and a time shift parameter $\tau_i$ and write:

$$\phi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0\,. \tag{2.13}$$

As for the spatial deformation, we use the notion of exp-parallelization defined section 2.3.2: the trajectory of a subject $i$ is obtained as the exp-parallelization of the representative trajectory using a certain vector $w_i$: $\eta^{w_i}(\gamma_0, \cdot)$.

Thus, using those temporal and spatial deformation, we define the trajectory of the subject $i$ as:

$$\gamma_i : t \mapsto \eta^{\omega_i}(\gamma_0, \psi_i(t)) \tag{2.14}$$

Hence, knowing the representative trajectory, the deformation of each subject is obtained using only 3 variables: the acceleration $\alpha_i$, the time shift $\tau_i$ and the space shift $\omega_i$.

For the model to be identifiable, the vectors $\omega_i$ are supposed to be orthogonal to the trajectory $\gamma_0$. It prevents the model from considering an acceleration with respect to the representative trajectory as a space shift (Schiratti et al., 2017). Moreover, to reduce the dimension of the problem, one can assume that the space shifts $w_i$ are obtained as linear combinations of independent sources: $w_i = As_i$ with $s_i$ a vector whose dimension is small compared to the dimension of $\omega_i$.

Finally, writing $z_i = (\alpha_i, \tau_i, s_i)$, the statistical model can be written as:

$$\begin{cases} y_{i,j} \,|\, z_i, \theta \sim \mathcal{N}(\gamma_i(t_{i,j}), \sigma^2) \\ z_i \,|\, \theta \sim \mathcal{N}(0, \Sigma) \end{cases} \tag{2.15}$$

where the parameters to estimate are $\theta = (t_0, m_0, c_0, p_0, A, \sigma, \Sigma)$. Often priors are added to theoretically ensure the existence of the maximum a posteriori estimate of the parameters.

To estimate $\theta$, we can use Expectation Maximization algorithms that we will present section 2.7. This spatio-temporal model has in particular be implemented in Deformetrica and can be accessed in open access (Bône et al., 2018b).

The principal limitation of the spatio-temporal model presented above is the necessity for the representative trajectory to be a geodesic. In particular, it means that the representative trajectory cannot take the same value twice. In certain cases, it is not a problem. For instance, for the Alzheimer's disease, this model has been successfully applied for an early detection (Koval et al., 2018; Bône et al., 2018a). However, it can be a severe limitation in other cases. In the case of chemotherapy, the treatment is often efficient at first, but, after a certain time, the tumor becomes resistant and its size increases again. Such a dynamic cannot be reproduced by the model previously presented. However, this dynamic, and particularly the time at which the tumor becomes resistant, would be of the upmost interest to doctors. One of the goals of this thesis will be to remove this limitation while considering heterogeneous populations.

## 2.5 Markov Chains and Metropolis Hastings algorithms

We now want to present the estimation algorithms we will use to estimate the parameters of the models presented previous section. To do so, we first need notions about Markov Chains, Metropolis Hastings algorithms and Stochastic Approximations. One can find a review of the Markov Chains notions presented here in Meyn and Tweedie (2012) or Douc et al. (2018). We only focus in this section on the required notions necessary to understand the rest of this thesis.

### 2.5.1 Markov Chains

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a separable space and $P$ a transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For any probability measure $\mu$ on $\mathcal{B}(\mathcal{X})$, one can define a probability measure $P_\mu$ on $\mathcal{B}(\mathcal{X})^{\otimes\mathbb{N}}$ such that, for all $n \geq 0$ and $A_0 \times ... \times A_n \in B(\mathcal{X})^{n+1}$,

$$P_\mu(X_0 \in A_0, ..., X_n \in A_n) = \int_{A_0 \times ... \times A_n} \mu(dx_0)P(x_0, dx_1)...P(x_{n-1}, dx_n)\,.$$

$(X_i)_{i\in\mathbb{N}}$ is then called the canonical Markov Chain.

We will quickly define four notions about Markov Chains that we will need further on: irreducibility, aperiodicity, small sets and ergodicity.

We say that a kernel $P$ is $\phi$-*irreducible* if there exists a non trivial measure $\phi$ on $\mathcal{B}(\mathcal{X})$ such that, for all $x \in \mathcal{X}$ and for all $A \in \mathcal{B}(\mathcal{X})$ verifying $\phi(A) \neq 0$, there exists $n \geq 1$ such that $P^n(x, A) > 0$. It means that all sets of $\phi$ measure positive are accessible. We call $\phi$ maximal if it dominates all other irreducibility measures.

$P$ is said to be *aperiodic* if there are no $d$ measurable sets $(A_i)_{1\leq i\leq d}$ with $d \geq 2$ and $\psi(\bigcup_{i=1}^d A_i) = 1$ for a certain maximal irreducibility measure $\psi$ such that $P(x, A_{i+1}) = 1$ for all $x \in A_i$, $1 \leq i \leq d$.

We then define the concept of *small sets*. If $P$ is irreducible and aperiodic, we say that a set $C \in \mathcal{B}(\mathcal{X})$ is $\nu$-small if there exist constants $\varepsilon > 0$, $m \geq 1$ and a probability measure $\nu$ such that

$$\forall x \in C,\ P^m(x, \cdot) \geq \varepsilon\nu\,.$$

The existence of small sets will be one of the conditions ensuring convergence of stochastic approximations with Markovian dynamic. In practice, if $P$ is $\psi$-irreducible aperiodic and $\mathcal{X}$ is separable, any set of positive $\psi$-measure contains a small set (Douc et al., 2018). Hence, it is not a restrictive condition.

We also define the notion of *petite sets*. $C \in \mathcal{B}(\mathcal{X})$ is petite if there exist $\varepsilon > 0$, probability measures $a$ on $\mathbb{N}$ and $\nu_a$ on $\mathcal{B}(\mathcal{X})$ such that, for all $x \in C$:

$$\sum_{n\in\mathbb{N}} a(n)P^n(x, \cdot) \geq \varepsilon\nu_a\,.$$

When $P$ is $\psi$-irreducible and aperiodic, petite sets are exactly the $\nu$-small sets.

We can now define the notion of *ergodicity*. A probability measure $\pi$ is invariant for $P$ if, for all $A \in \mathcal{B}(\mathcal{X})$,

$$\pi(A) = \int \pi(dy)P(y, A)\,.$$

$P$ is then said to be ergodic if it is $\psi$-irreducible, aperiodic, has an invariant probability measure $\pi$ and, for $\psi$-almost every $x$,

$$\lim_n ||P^n(x, \cdot) - \pi||_{TV} = 0\,,$$

where the total variation norm for a signed measure $\mu$ is defined by:

$$||\mu||_{TV} = \sup_{A\in\mathcal{B}(\mathcal{X})} \mu(A) - \inf_{A\in\mathcal{B}(\mathcal{X})} \mu(A)\,.$$

In particular, it means that, for any bounded function $g : \mathcal{X} \to \mathbb{R}$ and, for $\psi$-almost every $x$,

$$\lim_n |P^n g(x) - \pi(g)| = 0\,,$$

where $P^n g(x) := \int P^{n-1} g(y) P(x, dy)$, $P^0(x, dy) := \delta_x(dy)$ and $\pi(g) := \int g(y)\pi(dy)$.

Instead of interesting ourselves only to bounded functions $g$, we can choose to study functions $g$ increasing at infinity at most like another function $f$. To do so, we define, for $f \geq 1$, the $f$-norm of a function $g : \mathcal{X} \to \mathbb{R}$ by:

$$||g||_f = \sup_{x \in \mathcal{X}} \frac{|g(x)|}{f(x)}\,.$$

We write $\mathcal{L}_f = \{g : \mathcal{X} \to \mathbb{R}\,|\,||g||_f < \infty\}$. We can then define the $f$-norm of a signed measure $\mu$ by:

$$||\mu||_f := \sup_{||g||_f = 1} |\mu(g)|\,.$$

As before, we can now define the concept of $f$-ergodicity. $P$ is said to be $f$-ergodic if it is $\psi$-irreducible, aperiodic, has an invariant probability measure $\pi$ and for $\psi$-almost every $x$,

$$\lim_n ||P^n(x, \cdot) - \pi||_f = 0$$

It is rarely enough to know that a function is ergodic: we also need to know the speed of convergence. This means that we need to find a sequence $(r_n)_{n \in \mathbb{N}}$ positive and increasing such that;

$$\lim_n r_n ||P^n(x, \cdot) - \pi||_f = 0$$

$\psi$-almost everywhere.
When $r_n = \lambda^{-n}$ with $\lambda < 1$, we say that the sequence is $f$-geometrically ergodic. Otherwise, we say that it is $f$-subgeometric. In practice, it is often difficult to verify the ergodicity of a Markov Chain. To do so, we will use drift conditions presented below.

The first drift condition gives us the geometrical ergodicity of the Markov Chain. Suppose that there exist $V : \mathcal{X} \to [1, \infty)$ a measurable function, $\lambda \in [0, 1)$, $b \in [0, \infty)$, an integer $m$ and $C \in \mathcal{B}(\mathcal{X})$ such that:

$$P^m V \leq \lambda V + b\mathbb{1}_C\,. \tag{2.16}$$

In that case, the function $V$ is called the drift or Lyapunov function. Assume also that $P$ is aperiodic and that $C$ is a small set. Then, $P$ is $V$-geometrically ergodic.

If this drift condition immediately gives us the $\lambda$ defining the speed of convergence $r_n = \lambda^{-n}$, it is a bit more complex in the subgeometric case.
Assume that there exist a function $V : X \to [1, \infty)$, a concave, monotone nondecreasing differentiable function $\phi : [1, \infty] \to (0, +\infty]$ with $\lim_{t \to \infty} \phi'(t) = 0$, a petite set $C$ and a finite constant $b$ such that:

$$PV + \phi \circ V \leq V + b\mathbb{1}_C\,. \tag{2.17}$$

Suppose also that there exists $x_0 \in \mathcal{X}$ such that $V(x_0) < \infty$. Then, there exists a unique invariant distribution $\pi$ such that $P$ is $\phi \circ V$-ergodic.

This time, the rate of convergence $r_n$ is more complex to determine. We will here follow the steps detailed in Douc et al. (2004). Define:

$$H_\phi(v) = \int_1^v \frac{dx}{\phi(x)}\,,$$

$$r_\phi(z) = \phi \circ H_\phi^{-1}(z)\,.$$

Then, if equation (2.17) holds with $\sup_{x\in C} V(x) < \infty$ and if $P$ is $\psi$-irreducible and aperiodic then $P$ is ergodic with, for all $x \in \{V < \infty\}$:

$$\lim_n r_\phi(n)||P^n(x,\cdot) - \pi||_{TV} = 0$$

To find the rate of convergence in $f$-norm for a certain function $f$, we need to introduce the notion of Young functions. We say that $\Psi_1$ and $\Psi_2$ are a pair of Young functions if they are ultimately nondecreasing with $\lim_{x\to\infty} \Psi_1(x) = \lim_{x\to\infty} \Psi_2(x) = \infty$ and, for all $x, y \in [1, \infty)$,

$$\Psi_1(x)\Psi_2(y) \leq x + y\,.$$

One of the most usual pair of Young function is $\Psi_1(x) = p^{1/p}x^{1/p}$ and $\Psi_2(x) = q^{1/q}x^{1/q}$ with $1/p + 1/q = 1$.
Then, if equation (2.17) holds with $\sup_{x\in C} V(x) < \infty$ and if $P$ is $\psi$-irreducible and aperiodic then $P$ is ergodic with, for all $x \in \{V < \infty\}$:

$$\lim_n \Psi_1(r_\phi(n))||P^n(x,\cdot) - \pi||_{\Psi_2(\phi\circ V)} = 0\,.$$

The choice of $\Psi_1$ and $\Psi_2$ expresses a compromise between the rate of convergence $\Psi_1(r_\phi(n))$ and the control function $\Psi_2(\phi \circ V)$.

In practice, it is those drift conditions (2.16) and (2.17) which are proved in order to verify the ergodicity of a Markov Chain.

### 2.5.2 Metropolis Hastings algorithm

When exact simulation of a random variable is impossible, one can use Markov Chain Monte Carlo (MCMC) methods. The idea of MCMC methods is to generate a Markov Chain whose invariant distribution is the law we want to sample from. Hence, we replace one sample from a complex distribution by the hopefully easiest iteration of a Markov Chain.

The most frequently used Markov Chain Monte Carlo method is the Metropolis Hastings algorithm introduced by Hastings (1970). This method has the particular advantage to only require knowledge of the target distribution up to a multiplicative constant. Hence, it is not necessary to know the normalization constant, often intractable. At each step, given the current step of the Markov Chain, it consists in proposing a new value and accepting it according to a certain rate. Given a target $\pi$ and a proposal $q$, the exact algorithm is described in Algorithm 1.

In most cases, the proposal is a Gaussian distribution centered in the previous value $x_{k-1}$. It is then symmetric and the acceptance rate has an easier form:

$$\alpha(x_k, x) = 1 \wedge \frac{\pi(x)}{\pi(x_{k-1})}\,.$$

**Algorithm 1:** Metropolis Hastings algorithm.

**Data:** $x_0$ an initial value, K a number of iterations.
**for** $1 \leq k \leq K$ **do**

**Proposition step:** Sample $x \sim q(\cdot, x_{k-1})$.

**Acceptation step:** Compute

$$\alpha(x_k, x) = 1 \wedge \frac{\pi(x)q(x_{k-1}, x)}{\pi(x_{k-1})q(x, x_{k-1})}$$

and set:

$$x_k = \begin{cases} x & \text{with probability } \alpha(x_{k-1}, x) \\ x_{k-1} & \text{with probability } 1 - \alpha(x_{k-1}, x) \end{cases}$$

In that case, it measures how the likelihood has evolved between the proposal and the last iteration. This is known as symmetric Random Walk Metropolis Hastings (SRW-MH) and was heavily studied.

A natural question is how to choose the variance $\sigma_q$ of the proposal. Indeed, this variance influences the acceptance rate and tuning it manually is most of the time impossible as it has to evolve along the simulation to better fit the target density. In practice, we try to reach an acceptance rate $\text{ar}_{\text{goal}} = 30\%$ by adapting it every $n_{\text{adapt}}$ iterations following the idea of Roberts et al. (1997). One can use the following formula:

$$\sigma_q \leftarrow \sigma_q \left(1 + \frac{1}{k^\delta} \frac{\bar{\text{ar}} - \text{ar}_{\text{goal}}}{(1 - \text{ar}_{\text{goal}}) \cdot \mathbb{1}_{\bar{\text{ar}} \geq \text{ar}_{\text{goal}}} + \text{ar}_{\text{goal}} \cdot \mathbb{1}_{\bar{\text{ar}} < \text{ar}_{\text{goal}}}}\right)$$

with $k$ the current iteration, $\delta > 0.5$ and $\bar{\text{ar}}$ the mean acceptance rates over the last $n_{\text{adapt}}$ iterations.

A different problem arises when sampling in a high dimension space. In that case, it is difficult to explore the whole set using the proposal and thus, the resulting Markov Chain will not explore the whole array of possible target values. To overcome this problem, one can combine the Metropolis Hastings algorithm with Gibbs samplers. Instead of proposing a new vector at each iteration, one can choose to update only one coordinate or block of coordinate, accept or reject it and then repeat this process for each coordinate (see Geman and Geman (1984)).

We finally interest ourselves in the ergodicity of this Markov Chain. It will be necessary to know its rate of convergence towards the target distribution when used in parallel with a stochastic approximation algorithm (see section 2.6). Jarner and Hansen (2000) show that geometric ergodicity of random-walk-based Metropolis algorithm is equivalent to the acceptance probability being uniformly bounded away from zero.

In particular, the authors give us conditions of geometric ergodicity if the target density is super-exponential, i.e. if it is positive and has continuous first derivatives such that

$$\lim_{x \to \infty} n(x) \cdot \nabla \log \pi(x) = -\infty \,,$$

where $n(x)$ denotes the unit vector $x/|x|$. This condition implies that for any $H > 0$, there exists $R > 0$ such that

$$\frac{\pi(x + an(x))}{\pi(x)} \leq \exp(-aH) \qquad \text{for all } |x| \geq R, a \geq 0\,.$$

This equation means that $\pi(x)$ is at least exponentially decaying along any ray with the rate $H$ tending to in infinity as $x$ goes to in infinity.
Then, under this assumption, the random-walk-based Metropolis algorithm is geometrically ergodic if

$$\limsup_{|x|\to\infty} n(x) \cdot m(x) < 0\,,$$

where $m(x) = \nabla\pi(x)/|\nabla\pi(x)|$ is the normalized gradient of $\pi$.

When leaving the class of super-exponential targets, there is no longer easy conditions ensuring geometric ergodicity of the Metropolis Chain. In particular, for heavy tails targets, this ergodicity will not always be verified. Such a behaviour has been highlighted in Fort and Moulines (2000, 2003) among others. In those papers, different conditions are presented resulting in a subgeometric ergodicity of the Markov Chain obtained from a Metropolis Hastings algorithm. We present here two sets of conditions resulting in subgeometric ergodic Markov Chains.

**(E1)** The target density $\pi$ is continuous and positive on $\mathbb{R}^d$ and there exist $m \in (0,1)$, $r \in (0,1)$, positive constants $d_i$, $D_i$, $i = 0,1,2$ and $R_0 < \infty$ such that, if $|x| \geq R_0$, $x \mapsto \pi(x)$ is twice continuously differentiable and

$$\left\langle \frac{\nabla\pi(x)}{|\nabla\pi(x)|}, \frac{x}{|x|} \right\rangle \leq -r$$

$$d_0|x|^m \leq -\ln\pi(x) \leq D_0|x|^m$$
$$d_1|x|^{m-1} \leq |\nabla \ln\pi(x)| \leq D_1|x|^{m-1}$$
$$d_2|x|^{m-2} \leq |\nabla^2 \ln\pi(x)| \leq D_2|x|^{m-2}\,.$$

**(E2)** There exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. Moreover, $q_\theta$ is symmetric, bounded away from zero in a neighborhood of zero, and is compactly supported. We also assume that there exist $C > 0$ and $\beta \in (0,1)$ such that for all $(\theta, \theta') \in \Theta^2$,

$$\int_X |q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz) \leq C|\theta - \theta'|^\beta\,.$$

**Remark 2.5.1.** *Among others, the Weibull distribution on* $\mathbb{R}_+$ $\pi : x \mapsto \beta\eta x^{\eta-1}\exp(-\beta x^\eta)$ *with* $\beta > 0$ *and* $\eta \in (0,1)$ *verifies those conditions.*
*The compactly supported condition could be relaxed with appropriate moment conditions.*

We then have the following proposition proved in Fort (2009):

**Proposition 2.5.1** *Assume (E1) and (E2). Then, there exists* $s, c > 0$ *such that:*

$$\lim_{n\to\infty} \exp\left(cn^{m/(2-m)}\right) ||P^n(x,\cdot) - \pi||_{\pi^{-s}} = 0$$

Hence, we have a first example of targets with heavy tails (see condition (E1)) making the Markov Chain obtained from the Metropolis Hastings algorithm subgeometric.

We now give another set of conditions that implies a polynomial rate of convergence.

**(E3)** $\pi$ is continuous on $\mathbb{R}$ and there exist some finite constants $\alpha > 1$, $M > 0$, $C > 0$ and a function $\rho : \mathbb{R} \to [0, \infty)$ verifying $\lim_{x\to\infty} \rho(x) = 0$ such that for all $|x| > M$, $\pi$ is strictly decreasing and, for all $y \in \{z \in \mathbb{R} \,|\, \pi(x + z) \leq \pi(x)\}$,

$$\left| \frac{\pi(x+y)}{\pi(x)} - 1 + \alpha y x^{-1} \right| \leq C|x|^{-1}\rho(x)y^2 \,.$$

**(E4)** There exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. Moreover, $q_\theta$ is symmetric and there exists $\xi \geq 1$ such that $\int |y|^{\xi+3} q_\theta(y)dy < \infty$.

**Remark 2.5.2.** *This class of distributions contains in particular the Pareto distributions ($\pi(x) \propto x^{-\alpha}$) as well as many heavy tail distributions.*

We can now give the following proposition proved in Fort and Moulines (2003):

**Proposition 2.5.2** *Assume (E3) and (E4) and set $s^* = \xi \wedge s$. Then, for all $r \in [0, s^* - 1)$ and $\gamma \in [0, (s^* - 1 - r)/2)$,*

$$\lim_{n\to\infty} (n+1)^\gamma ||P^n(x, \cdot) - \pi||_{1+|x|^r} = 0$$

Hence, as we can see, in some situations, we cannot assume that the resulting Markov Chain is geometric ergodic. This will be an issue when dealing with Stochastic Approximations where, to ensure theoretical convergence, one needed the geometric ergodicity of the Markov Chain. This particular framework is studied next section.

## 2.6 Stochastic Approximations

### 2.6.1 Presentation

Now that we have defined the necessary notions about Markov Chains, we can discuss the Stochastic Approximations framework. The goal is to estimate the parameter $\theta^*$ solution of the equation:

$$\mathbb{E}_\theta[H_\theta(X)] = 0 \,, \tag{2.18}$$

where $H$ is known, but the computation of the expectation of $H_\theta(X)$ is impossible (distribution unknown or computation too expensive). Moreover, the distribution of $X$ may also depend on $\theta$. In that case, we write $h(\theta) = E_\theta[H_\theta(X)]$ the mean field, and we are looking for a solution of $h(\theta^*) = 0$.

In the Stochastic Approximation framework, we do not have direct access to the cost function $\mathbb{E}_\theta[H_\theta(X)]$ but only to noisy observations: $H_\theta(X_n)$ with $X_n$ having the same law as $X$. The stochastic approximation then produces a sequence $(\theta_n)_{n\in\mathbb{N}}$ of the form

$$\theta_n = \theta_{n-1} + \gamma_n H_{\theta_{n-1}}(X_n) \tag{2.19}$$

and converging towards a solution of equation (2.18).

One of the first examples of this process is the case of the stochastic gradient descent. In that case, we want to find a solution to

$$\theta^* = \text{argmin}_\theta \, \mathbb{E}[Q(\theta, X)]]$$

and we observe independent samples $X_n$ with distribution $X$. Hence, the choice $H_\theta(y) = -\nabla_\theta Q(\theta, y)$ leads to process looking like a gradient descent algorithm where the expectation sign $\mathbb{E}$ has been omitted.

In general however, the variables $X_n$ are not independent. In the Markovian dynamic setting, $X_n$ is a random variable depending only on $(\theta_{n-1}, X_{n-1})$. Hence, the distribution of $X_n$ knowing $(\theta_{n-1}, X_{n-1})$ is given by a transition kernel $P_{\theta_n}(X_{n-1}, \cdot)$. This setting contains in particular the Stochastic Approximation Expectation Maximization algorithm presented section 2.7.4.

The convergence of the Stochastic Approximations in the Markovian dynamic case has in particular been studied in Andrieu et al. (2005). The authors present a set of hypotheses, recalled here, and ensuring the convergence of the stochastic approximation.

### 2.6.2 Convergence theorem in the Markovian dynamic case

In the following, we denote $\mathcal{X}$ the state space and $\Theta$ the parameter space that we assume to be an open subset of $\mathbb{R}^{n_\theta}$. Moreover, we suppose that both are equipped with countably generated $\sigma$-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$. We present the framework of a stochastic approximation producing a sequence of elements converging towards a solution of $h(\theta) = 0$ when there exist probability measures $\pi_\theta$ such that, for any $\theta \in \Theta$, $h(\theta) = \mathbb{E}_{\pi_\theta}(H_\theta(X))$ with $H_\theta : \mathcal{X} \mapsto \Theta$. $h$ is then called the mean field of the stochastic approximation.

Let $\Delta = (\Delta_n)_{n\in\mathbb{N}}$ be a non-increasing sequence of positive real numbers with $\Delta_0 \leq 1$ and set $\theta_c \notin \Theta$ and $x_c \notin \mathcal{X}$ two cemetery states. We then define a Markov chain $Y_n^\Delta = (X_n, \theta_n)$ on $\mathcal{X} \cup \{x_c\} \times \Theta \cup \{\theta_c\}$ by:

$$\theta_{n+1} = \begin{cases} \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}) & \text{and } X_{n+1} \sim P_{\theta_n}(X_n, .) & \text{if } \theta_n \in \Theta \\ \theta_c & \text{and } X_{n+1} = x_c & \text{if } \theta_n \notin \Theta \,. \end{cases} \tag{2.20}$$

Keeping notations and hypotheses labels from Andrieu et al. (2005), we put the following hypothesis on the transition probabilities $(P_\theta, \theta \in \Theta)$ and on the random vector field $H$:

**(A2)** For any $\theta \in \Theta$, the Markov kernel $P_\theta$ has a single stationary distribution $\pi_\theta$. In addition, $H : \Theta \times \mathcal{X} \to \Theta$ is measurable for all $(\theta, x) \in \Theta \times \mathcal{X}$ .

The existence and uniqueness of the invariant distribution can be verified under the classical conditions of irreducibility and recurrence Meyn and Tweedie (2012).

We assume the mean field $h$ satisfies the following hypothesis that amounts to the existence of a global Lyapunov function:

**(A1)** $h : \Theta \to \mathbb{R}^{n_\theta}$ is continuous and there exists a continuously differentiable function $w : \Theta \to [0, +\infty[$ such that:

(i) there exists $M_0 > 0$ such that

$$\mathcal{L} := \{\theta \in \Theta, \langle \nabla w(\theta), h(\theta) \rangle = 0\} \subset \{\theta \in \Theta, w(\theta) < M_0\} \ ,$$

(ii) there exists $M_1 \in (M_0, +\infty]$ such that $\mathcal{W}_{M_1} := \{\theta \in \Theta, w(\theta) \leq M_1\}$ is a compact set,

(iii) for any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla w(\theta), h(\theta) \rangle < 0$,

(iv) the closure of $w(\mathcal{L})$ has an empty interior.

We denote by $\mathcal{F} = \{\mathcal{F}_n, n \geq 0\}$ the natural filtration of the Markov chain $(X_n, \theta_n)$ and by $\mathbb{P}^{\Delta}_{x,\theta}$ the probability measure associated to the chain $(Y^{\Delta}_n)$ started from the initial conditions $(x, \theta) \in \mathcal{X} \times \Theta$. Finally, we denote by $Q_{\Delta_n}$ the sequence of transition probabilities that generate the inhomogeneous Markov chain $(Y^{\Delta}_n)$.

To ensure convergence of the sequence towards a root of $h$, the sequence $(\theta_n)_{n\in\mathbb{N}}$ is required to remain in a given compact set. This assumption is rarely satisfied. To alleviate this constraint, we introduce the usual trick which consists in reprojecting on increasing compact sets. It is then proved that the sequence will be projected only a finite number of times along the algorithm. Using this trick, the sequence $(\theta_n)_{n\in\mathbb{N}}$ now remains in a compact set of $\Theta$. We detail this process below.

We assume that there exists $(\mathcal{K}_n)_{n\in\mathbb{N}}$ a sequence of compact subsets of $\Theta$ such that

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta \qquad \text{and} \qquad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}) \, .$$

Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence of non-increasing positive numbers and $K$ be a subset of $\mathcal{X}$. Let $\Phi : \mathcal{X} \times \Theta \to K \times \mathcal{K}_0$ be a measurable function. We then define the stochastic approximation algorithm with adaptive truncation sets as a homogeneous Markov chain on $\mathcal{X} \times \Theta \times \mathbb{N} \times \mathbb{N}$ by

$$Z_n = (X_n, \Theta_n, \kappa_n, \nu_n) \tag{2.21}$$

with the following transition at iteration $n + 1$:

- If $\nu_n = 0$, then draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(\Phi(X_n, \theta_n), .)$. Otherwise, draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(X_n, \theta_n, .)$.
- If $|\theta_{n+1} - \theta_n| \leq \varepsilon_n$ and $\theta_{n+1} \in \mathcal{K}_{\kappa_n}$ then set $\kappa_{n+1} = \kappa_n$ and $\nu_{n+1} = \nu_n + 1$. Otherwise, set $\kappa_{n+1} = \kappa_n + 1$ and $\nu_{n+1} = 0$.

To summarize this process, if our parameter $\theta$ leaves the current truncation set $\mathcal{K}_{\kappa_n}$ or if the difference between two of its successive values is larger than a time dependent threshold $\varepsilon_n$, we reinitialize the Markov chain by a value inside $\mathcal{K}_0$: $\Phi(X_n, \theta_n)$ and update the truncation set to a larger one $\mathcal{K}_{\kappa_n+1}$ as well as the threshold to a smaller one: $\varepsilon_{n+1}$. Hence, $\kappa_n$ represents the number of re-initializations before the step $n$ while $\nu_n$ is the number of steps since the last re-initialization.

The idea behind this truncation process is to force the noise to be small in order for the drift $h(\theta)$ to dominate. We do so by forcing our algorithm to come back to the center of $\Theta$ whenever the parameters become too large.

We finally state two last hypotheses about the control of fluctuations before presenting the theorem proved in Andrieu et al. (2005).

We first define, for any compact $\mathcal{K}$ and any sequence of non-increasing positive numbers $(\varepsilon_k)_{k\in\mathbb{N}}$, $\sigma(\mathcal{K}) = \inf(k \geq 1, \theta_k \notin \mathcal{K})$ and $\nu_\varepsilon = \inf(k \geq 1, |\theta_k - \theta_{k-1}| \geq \varepsilon_k)$. Moreover, for $W : \mathcal{X} \to [1, \infty)$ and $g : \mathcal{X} \to \mathbb{R}^{n_\theta}$, we write

$$||g||_W = \sup_{x\in\mathcal{X}} \frac{|g(x)|}{W(x)} .$$

We can now present the hypothesis (A3):

**(A3)** For any $\theta \in \Theta$, the Poisson equation $g - P_\theta g = H_\theta - h(\theta)$ has a solution $g_\theta$. Moreover, there exist a function $W : \mathcal{X} \to [1, +\infty]$ such that $\{x \in \mathcal{X}, W(x) < +\infty\} \neq \emptyset$, constants $\alpha \in (0, 1]$ and $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(i) the following holds:

$$\sup_{\theta\in\mathcal{K}} ||H_\theta||_W < \infty \tag{2.22}$$

$$\sup_{\theta\in\mathcal{K}} ||g_\theta||_W + ||P_\theta g_\theta||_W < \infty \tag{2.23}$$

$$\sup_{\theta,\theta'\in\mathcal{K}} ||\theta - \theta'||^{-\alpha} \left(||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W\right) < \infty \tag{2.24}$$

(ii) there exist constants $\{C_k, k \geq 0\}$ such that, for any $k \in \mathbb{N}$, for any sequence $\Delta$ and for any $x \in \mathcal{X}$,

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\geq k}] \leq C_k W^p(x) \tag{2.25}$$

(iii) there exist a sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ and a constant $C$ such that for any sequence $\Delta$ and for any $x \in \mathcal{X}$,

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon\geq k}] \leq C W^p(x) . \tag{2.26}$$

This assumption concerns the existence and regularity of the Poisson equation associated with each of the transition kernel $P_\theta$. Finally, the last condition concerns the step size sequences:

**(A4)** The sequences $(\Delta_k)_{k\in\mathbb{N}}$ and $(\varepsilon_k)_{k\in\mathbb{N}}$ are non-increasing, positive and satisfy $\sum_{k=0}^{\infty} \Delta_k = \infty$, $\lim_{k\to\infty} \varepsilon_k = 0$ and

$$\sum_{k=1}^{\infty} \Delta_k^2 + \Delta_k \varepsilon_k^\alpha + (\varepsilon_k^{-1}\Delta_k)^p < \infty$$

where $p$ and $\alpha$ are defined in (A3).

We can finally state the theorem proved in Andrieu et al. (2005):

**Theorem 2.6.1** *Andrieu et al. (2005) Assume (A1)-(A4). Let $K \subset \mathcal{X}$ such that $\sup_{x\in K} W(x) < \infty$ and such that $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ (where $M_0$ and $\mathcal{W}_{M_0}$ are defined in (A1)) and let $Z_n$ be as defined in* (2.21)*. Then, for all $(x,\theta) \in \mathcal{X}\times\Theta$, we have $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$, $\mathbb{P}^{\Delta}_{x,\theta}$-a.s. where $\mathcal{L}$ is defined in (A1).*

Of the four conditions (A1) to (A4), (A3) is often the most difficult to verify and we need more practical conditions. In particular, in Andrieu et al. (2005), the authors show that a geometric ergodicity of the Markov Chain implies (A3). We recall this hypothesis here.

**(DRI)** For any $\theta \in \Theta$, $P_\theta$ is $\psi$-irreducible and aperiodic. In addition, there exist a function $V : X \to [1, +\infty)$, constants $p \geq 2$ and $\beta \in [0, 1]$ such that, for any compact subset $\mathcal{K} \subset \Theta$,

(DRI1) there exist an integer $m$, constants $0 < \lambda < 1$, $b$, $\kappa$, $\delta > 0$ and a probability measure $\nu$ such that:

$$\sup_{\theta\in\mathcal{K}} P^m_\theta V^p(x) \leq \lambda V^p(x) + b\mathbb{1}_C(x)\,,$$

$$\sup_{\theta\in\mathcal{K}} P_\theta V^p(x) \leq \kappa V^p(x) \qquad \forall x \in \mathcal{X}\,,$$

$$\inf_{\theta\in\mathcal{K}} P^m_\theta(x, A) \geq \delta\nu(A) \qquad \forall x \in C, \forall A \in \mathcal{B}(\mathcal{X})\,.$$

(DRI2) there exists $C$ such that, for all $x \in \mathcal{X}$,

$$\sup_{\theta\in\mathcal{K}} |H_\theta(x)| \leq CV(x)\,,$$

$$\sup_{\theta,\theta'\in\mathcal{K}} |\theta - \theta'|^\beta |H_\theta(x) - H_{\theta'}(x)| \leq CV(x)\,,$$

(DRI3) there exists $C$ such that, for all $(\theta, \theta') \in \mathcal{K}\times\mathcal{K}$,

$$||P_\theta g - P'_\theta g||_V \leq C||g||_V|\theta - \theta'|^\beta \qquad \forall g \in \mathcal{L}_V\,,$$

$$||P_\theta g - P'_\theta g||_{V^p} \leq C||g||_{V^p}|\theta - \theta'|^\beta \qquad \forall g \in \mathcal{L}_{V^p}\,.$$

where $\mathcal{L}_V := \{g : \mathcal{X} \to \mathbb{R}^{n_\theta} \,|\, ||g||_V < \infty\}$

The assumption (DRI1) is classical in the Markov chain literature as it implies the existence of a stationary distribution $\pi_\theta$ for all $\theta \in \Theta$ and $V^p$-geometric ergodicity as explained section 2.5.1.

Andrieu et al. (2005) then prove the following proposition:

**Proposition 2.6.1** *Assume (DRI). Then, (A2) and (A3) are verified for any $0 < \alpha < \beta$.*

If the condition (DRI) is easier to prove than (A3) in practice, it presents the serious drawback of asking the geometric ergodicity of the Markov kernel. Often, the distribution $Y_n$ is sampled using a Metropolis Hastings algorithm. But, as explained section 2.5.2, if the target has heavy tails, the kernel is only subgeometrically ergodic (see Douc et al. (2004); Fort and Moulines (2000, 2003); Jarner and Hansen (2000) among others). The assumption (DRI) will thus not be verified in that case and the proposition 2.6.1 will not be applicable. To overcome this prob-

lem, we will relax those conditions in the chapter 5 to allow subgeometric kernels with some hypotheses on their regularity and speed of convergence.

## 2.7 The Expectation Maximization algorithm and its variants

We now present Expectation Maximization algorithms that will allow us to estimate the parameters of the mixed effects models considered in this thesis. Some of them will in particular use the Stochastic Approximation framework presented above.

### 2.7.1 The Expectation Maximization algorithm

The Expectation-Maximization (EM) algorithm has first been introduced in Dempster et al. (1977). It proposes a general approach to iteratively compute maximum-likelihood estimates when the observations are viewed as incomplete data.
The term incomplete data means that we will consider two state spaces $\mathcal{Z}$ and $\mathcal{Y} \subset \mathbb{R}^n$. The observed data $y$ belongs to $\mathcal{Y}$. It depends of an unobserved variable $z \in \mathcal{Z}$, not observed directly but only indirectly through $y$.
The problem is then the following. We assume we have a family of complete densities $f(y, z, \theta)$ depending on a parameter $\theta$ with $y$ the observations and $z$ the latent, non observed, variables. We assume that the densities are integrable with respect to the measure $\mu$. From those, we derive the corresponding family of incomplete data densities:

$$g(y, \theta) = \int_{\mathcal{Z}} f(y, z, \theta)\, \mu(dz)\,.$$

The goal is then to find the parameter $\theta$ maximizing the observed likelihood $g(y, \theta)$ given the observations $y$. The EM algorithm allows us to compute this maximum using only the complete density $f$.
It consists in two different steps called the Expectation step (E-step) and Maximization step (M-step). To describe its operation, we introduce the following function:

$$Q(\theta|\theta') := \mathbb{E}\Big(\log f(y, z, \theta)\Big|y, \theta'\Big) = \int_{\mathcal{Z}} \log\Big(f(y, z, \theta)\Big) p(z|y, \theta')\, \mu(dz)\,,$$

where $p$ is the conditional distribution of $z$ given the observations $y$:

$$p(z|y, \theta) = \begin{cases} f(y, z, \theta)/g(y, \theta) & \text{if } g(y, \theta) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Dempster et al. (1977) then show that, for all $k \in \mathbb{N}$, $\log g(y, \theta_{k+1}) \geq \log g(y, \theta_k)$. Hence, the EM algorithm increases the value of the observed log likelihood at each iteration. It can however be observed that there is no guarantee that the EM algorithm will converge towards a global maximum of the observed likelihood.

The convergence properties of the EM algorithm have been studied by Wu (1983) in a general case.

Delyon et al. (1999) have then given more specific hypothesis in the case where the joint likelihood belongs to the curved exponential family (M1):

**Algorithm 2:** EM algorithm

**Data:** $\theta_0$ an initial value of the parameter.
**for** $1 \leq k \leq K$ **do**

**E-step:** Compute $Q(\theta|\theta_{k-1}) = \mathbb{E}(\log f(y, z, \theta)|y, \theta_{k-1})$.

**M-step:** Choose $\theta_k$ a value maximizing $Q(\theta|\theta_{k-1})$.

**(M1)** The parameter space $\Theta$ is an open subset of $\mathbb{R}^p$ with $p \in \mathbb{N}$. Moreover, for all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$ and $\theta \in \Theta$, the complete likelihood can be written as:

$$f(y, z, \theta) = \exp\left(\Psi(\theta) + \langle S(y, z), \Phi(\theta)\rangle\right) \tag{2.27}$$

where $S : \mathcal{Y} \times \mathcal{Z} \to \mathcal{S}$ is a Borel function taking its value in $\mathcal{S}$, an open subset of $\mathbb{R}^{n_s}$.
In that case, we say that $f$ belongs to the curved exponential family.
Moreover, we assume that the convex hull of $S(\mathbb{R}^l)$ is included in $\mathcal{S}$ and, for all $\theta \in \Theta$ for all $y \in \mathcal{Y}$,

$$\int_{\mathbb{R}^l} |S(y, z)|p(z|y, \theta)\mu(dz) < \infty$$

**(M2)** The functions $\Psi$ and $\Phi$ are twice continuously differentiable on $\Theta$.

**(M3)** The function $s : \Theta \to \mathcal{S}$ defined as:

$$s(\theta) = \int_{\mathbb{R}^l} S(y, z)p(z|y, \theta)\mu(dz)$$

is continuously differentiable on $\Theta$.

**(M4)** The observed likelihood $g$ is continuously differentiable on $\Theta$ and

$$\partial_\theta g(y, \theta) = \int_{\mathbb{R}^l} \partial_\theta f(y, z, \theta)\mu(dz)$$

**(M5)** There exists a function $\hat{\theta} : \mathcal{S} \to \Theta$ such that

$$\forall \theta \in \Theta, \forall s \in \mathcal{S}, L(s, \hat{\theta}(s)) \geq L(s, \theta)$$

with $L(s, \theta) = -\Psi(\theta) + \langle s, \Phi(\theta)\rangle$.
Moreover, $\hat{\theta}$ is continuously differentiable on $\mathcal{S}$.

Using the fact that the joint likelihood belongs to the curved exponential family, the EM algorithm can be written in an easier way. Indeed, in that case,

$$Q(\theta|\theta') = \Psi(\theta) + \langle \mathbb{E}\left(S(y, z)|y, \theta'\right), \Phi(\theta)\rangle \ .$$

Hence, at each iteration, we only need to compute $S_k = \mathbb{E}(S(y, z)|y, \theta_{k-1})$. Using (M5), the M-step can also be rewritten as $\theta_k = \hat{\theta}(S_k)$. This process is summarized in algorithm 3.

Then, we have the following convergence theorem:

**Algorithm 3:** EM algorithm when $f$ belongs to the curved exponential family.

**Data:** $\theta_0$ an initial value of the parameter.
**for** $1 \leq k \leq K$ **do**

**E-step:** Compute $S_k = \mathbb{E}(S(y,z)|y,\theta_{k-1})$.

**M-step:** Set $\theta_k = \hat{\theta}(S_k)$.

**Theorem 2.7.1** *Delyon et al. (1999)*
*Assume that (M1) to (M5) hold and that clos* $(\mathcal{L})$ *is a compact subset of* $\Theta$ *where*

$$\mathcal{L} := \{\theta \in \Theta | \partial_\theta g(y,\theta) = 0\} \ .$$

*Then, for any initial point* $\theta_0$*, the sequence* $(g(y,\theta_k))_{k\in\mathbb{N}}$ *is increasing and*

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

The Expectation Maximization algorithm has been used in a wide variety of applications such as maximum likelihood estimation of the parameters of mixture of densities (Titterington et al., 1985), of hidden Markov models (MacDonald and Zucchini, 1997) or maximum a posteriori estimation in censored data model (Little and Rubin, 1989) among many others.

It must be remarked that the EM algorithm can be difficult to execute in certain cases. Indeed, while the M-step is often achievable in closed form, it is not always the case. When this maximization step is not possible, one can replace the maximization by a single step of an approximate Newton's method leading to the EM gradient algorithm (Lange, 1995).

The E-step is often more problematic. Most of the time, the function $f$ is complex and the computation of the expectation not possible. To overcome this difficulty, different algorithms have been introduced. Those will be the subject of the next subsections with an emphasis being made on the SAEM and the MCMC-SAEM algorithms.

### 2.7.2 The Stochastic EM algorithm

When the E-step is intractable, Celeux (1985) proposes to replace the computation of the expectation of $f(y,z,\theta)$ with respect to the conditional distribution by the simpler sampling of $z_k$ with respect to this conditional distribution followed by the computation of $f(y,z_k,\theta)$. They then maximize the function $\theta \mapsto f(y,z_k,\theta)$. This leads to the algorithm 4.

This new algorithm has the advantage not to necessitate the computation of any expectation. Moreover, by introducing some randomness, it limits its dependence with respect to its initial parameter $\theta_0$ which can be a problem with the EM algorithm. However, contrary to the EM algorithm and the other variants we will see later on, its convergence is not proven almost surely but only in mean.

### 2.7.3 The Monte Carlo EM algorithm

In the same vein as the SEM, the Monte Carlo EM Wei and Tanner (1990) replaces the E-step by computing a Monte Carlo approximation of the expectation of $Q$ using a large amount of

**Algorithm 4:** Stochastic EM algorithm.

**Data:** $\theta_0$ an initial value of the parameter.
**for** $1 \leq k \leq K$ **do**

**S-step:** Sample $z_k$ from $p(.|y, \theta_{k-1})$.

**M-step:** Set $\theta_k \in$ argmax $f(y, z_k, \theta)$.

simulated missing data $z$. This is summarized algorithm 5.

**Algorithm 5:** Monte Carlo EM algorithm.

**Data:** $\theta_0$ an initial value of the parameter and $m$ the number of variables sampled in the S-step.
**for** $1 \leq k \leq K$ **do**

**S-step:** Sample $m$ variables $z_k^j$, for $j$ between $1$ and $m$ from $p(.|y, \theta_{k-1})$.

**E-step:** Compute the Monte-Carlo approximation of $Q$:

$$Q_k(\theta) = \frac{1}{m} \sum_{j=1}^{m} \log f(y, z_k^j, \theta)$$

**M-step:** Set $\theta_k \in$ argmax $Q_k(\theta)$.

It can be remarked that, if $m = 1$, we recover the previous algorithm. Similarly, $m = \infty$ corresponds to the initial EM. In order for $Q_k$ to approximate the expectation $Q$, we need $m$ to be big enough. However, taking $m$ big greatly increases the computation time. To mitigate this problem Wei and Tanner (1990) advise to begin with $m = 1$ and to gradually increase its value.

Fort et al. (2003) have then proved the almost-sure convergence of the MCEM algorithm in the case where the joint distribution $f$ belongs to the curved exponential family and when the number of simulations $m$ increases along iterations.
If this algorithm allows us to compute a maximum likelihood estimate when the E-step is intractable if we are able to simulate $z$, it must however be remarked that it can be computationally costly as one needs to sample more and more values of $z$. To overcome this problem, a new algorithm using stochastic approximations is presented next subsection.

### 2.7.4 The Stochastic Approximation EM algorithm

Delyon et al. (1999) propose to use the theory of the stochastic approximations presented section 2.6 to approximate the expectation of $Q$ using only one simulated value of the missing data $z$.
Once again, as we suppose that $f$ belongs to the curved exponential family, the SAEM algorithm can be written in the form given in algorithm 6, only involving the sufficient statistics.

**Algorithm 6:** SAEM algorithm.

**Data:** $\theta_0$ an initial value of the parameter.
**for** $1 \leq k \leq K$ **do**

**Simulation step:** Generate $z_k$ a realization of the hidden variables under the conditional density $p(z|y, \theta_{k-1})$

**Approximation step:** Update $S_k = S_{k-1} + \gamma_k(S(y, z_k) - S_{k-1})$.

**Maximization step:** Set $\theta_k = \hat{\theta}(S_k)$.

It can be remarked, that, contrary to the MCEM algorithm, the SAEM algorithm only requires us to sample one realization of the conditional density, greatly reducing the complexity of the problem.

To prove the convergence of this algorithm, the authors add the following hypothesis:

**(SAEM1)** For all $k \geq 0$, $0 \leq \gamma_k \leq 1$, $\sum_{i=1}^{\infty} \gamma_k = \infty$ and $\sum_{i=1}^{\infty} \gamma_k^2 < \infty$

**(SAEM2)** $l : \Theta \to \mathbb{R}$ and $\hat{\theta} : \mathcal{S} \to \Theta$ are $n_s$ times differentiable.

**(SAEM3)** For all positive Borel function $\phi$:

$$E(\phi(z_{k+1})|\mathcal{F}_k) = \int \phi(z)p(z|y, \theta_k)\mu(\mathrm{d}z)$$

where $z_k$ is the missing value simulated at step $k$ under the conditional density $p(z|y, \theta_{k-1})$ and $\mathcal{F}_n$ is the family of $\sigma$-algebra generated by the random variables $S_0, z_1, \ldots, z_n$.

**(SAEM4)** For all $\theta \in \Theta$, $\int_{\mathbb{R}^l} ||S(y, z)||^2 p(z|y, \theta)\mu(dz) < \infty$ and $\Gamma(\theta) := \mathrm{Cov}_\theta(S(z))$ is continuous with respect to $\theta$.

**(A)** With probability 1, clos$((S_k)_{k\geq 1})$ is a compact subset of $\mathcal{S}$.

**Remark 2.7.1.** *The assumption (A) can easily be relaxed without further hypotheses by projecting the sequence $(S_k)_{k\in\mathbb{N}}$ on increasing compacts. See Andrieu et al. (2005) for more details.*

We then have the following convergence theorem:

**Theorem 2.7.2** *Delyon et al. (1999)*
*Assume that (M1) to (M5), (SAEM1) to (SAEM4) and (A) are verified. Then, with probability 1,*

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0 \qquad \textit{where} \qquad \mathcal{L} = \{\theta \in \Theta | \partial_\theta g(\theta) = 0\}.$$

If this algorithm is implementable without having to compute the intractable E-step, it still presents two major drawbacks. First, we once again suppose that the joint probability belongs to the curved exponential family. However, this assumption is not verified in different situations: neither for heteroscedastic models (Dubois et al., 2011; Kuhn and Lavielle, 2005) nor with some more complex models (Bône et al., 2018a; Debavelaere et al., 2020; Lindsten, 2013; Meza

et al., 2012; Schiratti et al., 2015; Wang, 2007). The relaxation of this hypothesis will be the object of chapter 6.
The second restriction is the necessity to be able to draw from the conditional distribution $p(.|y,\theta)$ for any $\theta \in \Theta$. In many practical situation, it is in fact impossible to directly draw from this distribution. The next subsection answers this problematic.

### 2.7.5 The Monte Carlo Markov Chain SAEM algorithm

In Kuhn and Lavielle (2004), the authors propose to couple the Stochastic Approximation Scheme with a Monte Carlo Markov Chain method. Instead of sampling from the conditional distribution, they propose to compute one step of a Markov Chain whose invariant distribution is $p(.|y,\theta_k)$. More precisely, they suppose that, for all $\theta \in \Theta$, there exists a transition kernel $\Pi_\theta$ whose unique invariant distribution is $p(.|y,\theta)$ (see section 2.5.2 for an example of such a kernel). The MCMC-SAEM then takes the form described algorithm 7.

**Algorithm 7:** MCMC-SAEM algorithm.

**Data:** $\theta_0$ an initial value of the parameter.
**for** $1 \leq k \leq K$ **do**

**Simulation step:** Generate $z_k \sim \Pi_{\theta_{k-1}}(z_{k-1},.)$

**Approximation step:** Update $S_k = S_{k-1} + \gamma_k(S(y,z_k) - S_{k-1})$.

**Maximization step:** Set $\theta_k = \hat{\theta}(S_k)$.

The authors then replace the hypotheses (SAEM3) and (SAEM4) by the following:

**(SAEM3')**

1. The chain $(z_k)_{k\geq 0}$ takes its values in a compact subset of $\mathcal{E} \subset \mathcal{Z}$.
2. For any compact subset $V$ of $\Theta$, there exists a real constant $L$ such that, for any $(\theta,\theta') \in V^2$,
$$\sup_{(x,y)\in\mathcal{E}^2} |\Pi_\theta(x,y) - \Pi'_\theta(x,y)| \leq L|\theta - \theta'|\,.$$
3. The transition probabilities $\Pi_\theta$ generate a uniformly ergodic chain whose invariant probability is the conditional distribution $p(.|y,\theta)$:
$$\exists K_\theta \in \mathbb{R}^+, \exists \rho_\theta \in ]0,1[ | \forall z \in \mathcal{E}, \forall k \in \mathbb{N}, ||\Pi^k_\theta(z,.) - p(.|y,\theta)||_{TV} \leq K_\theta \rho^k_\theta$$
where $||.||_{TV}$ refers to the total variation norm, $\sup_\theta K_\theta < \infty$ and $\sup_\theta \rho_\theta < 1$.
4. The function S is bounded on $\mathcal{E}$.

Under these new hypotheses, one can state the following theorem:

**Theorem 2.7.3** *Kuhn and Lavielle (2004)*
*Assume that assumptions (M1) to (M5), (SAEM1), (SAEM2) and (SAEM3') and (A) are verified. Then, with probability 1,*

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0 \qquad \textit{where} \qquad \mathcal{L} = \{\theta \in \Theta \,|\, \partial_\theta g(\theta) = 0\}\,.$$

**Remark 2.7.2.** *Once again, it is possible to do without assumption (A) by projecting the sufficient statistics on increasing compacts and obtain convergence of the resulting algorithm without adding any other hypothesis (Allassonnière et al., 2010).*

It can also be highlighted that targeting the exact distribution by a Markov Chain is not always necessary. Indeed, in Allassonnière and Chevallier (2019), the authors show that sampling from an approximate distribution is enough under some assumptions. This, in particular, opens the possibility to use tempered distributions to improve the convergence.

Hence, instead of sampling from the exact conditional distribution, it is now possible to use Markov Chains targeting it. In part II we will use this algorithm coupled with Metropolis Hastings algorithms to compute maximum of likelihood estimates.

### 2.7.6 Restrictions of the Stochastic EM algorithms

In the part III, we will tackle two restrictions of those Stochastic EM algorithms.

First, the condition (SAEM3'-3.) forces us to only consider geometric ergodic chains. This can be a problem when the conditional probability has heavy tails (Weibull or Pareto distributions for instance). Indeed, in that case, a Metropolis Hastings Markov Chain targeting this conditional probability is only subgeometric ergodic. We will show chapter 5 that this condition can be relaxed by supposing only a subgeometric ergodicity of the Markov Chain with appropriate assumptions on the rate of convergence.

Moreover, both the SAEM and the MCMC-SAEM algorithms suppose that $f$ belongs to the curved exponential family while it is not verified in many practical examples. We will see chapter 6 how to deal with this assumption when it is not verified and we will propose a variant of the SAEM algorithm allowing a better estimation of the maximum of likelihood in that case.

## 2.8 Thesis outline

In this thesis, we will focus on some limitations of the existing models presented above. The manuscript is separated in two principal parts. In the first part, we focus on the modeling of medical data. Both longitudinal and cross-sectional data will be studied for two different purposes: clustering longitudinal trajectories of shapes with successive dynamics and identifying anomalies (such as tumors) in organs. In the second part, we study different theoretical properties of Stochastic Approximation and Expectation Maximization algorithms. The four different chapters constituting this manuscript are summarized below.

- **Chapter 3:** *Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories.*

  In this chapter, we choose to study data sets of longitudinal observations. The goal of a longitudinal atlas is to create a representative trajectory of the population as well as the deformations towards each subject. As explained section 2.4, longitudinal atlases as introduced in Schiratti et al. (2015) have the disadvantage of modeling the representative

trajectory as a geodesic. However, in lots of practical situations, this is not a valid hypothesis. It is for instance not valid in the case of chemotherapy where the tumor becomes resistant to the treatment after a certain time. To overcome this problem, we model the representative trajectory by a piecewise geodesic. This idea has first been introduced in Chevallier et al. (2017) where the authors proposed such a model for scalar data.
In this first chapter, we generalize this discussion in larger dimensions by introducing rupture times at which the representative trajectory goes from one dynamic to another. Modeling the representative trajectory as a piecewise geodesic also allows us to consider more complex data sets. Indeed, we can suppose that the population is separated in different clusters whose respective representative trajectories branch or join at different rupture times. This is in part our motivation to introduce unsupervised clustering in the model. This new model is presented chapter 3 and is applied on different data sets such as the RECIST score in the case of chemotherapy or meshes of the hippocampus in the case of the Alzheimer's disease.
This work has been presented at the International Conference on Medical Image Computing and Computer-Assisted Vision (Debavelaere et al., 2019) and published in the International Journal of Computer Vision (Debavelaere et al., 2020).

- **Chapter 4:** *Detection of anomalies using the LDDMM framework*.

  In this chapter, we are interested in the detection of anomalies, such as the presence of tumors in a medical image. More precisely, we place ourselves in the cross-sectional framework and assume that we have at our disposal a template of control subjects. We then define an anomaly as a structure that cannot be recovered as a diffeomorphic deformation of the control template.
  For example, in the case of tumor detection, the control template will have no tumor and the diffeomorphic deformations from this template will also have no tumor. We are therefore able to recover these tumors in the residuals of the deformation.
  We show that this process improves the reconstruction of the observations and indeed allows anomalies to be detected.
  In particular, our method has the advantage of not requiring large data sets or annotations by physicians. Moreover, it can be easily applied to any organ. To highlight these advantages, we apply this method to two different data sets: a data set of livers from patients with metastatic colorectal cancer and a data set of brains with gliomas.
  This chapter will be converted into a paper for submission.

- **Chapter 5:** *On the convergence of stochastic approximations under a subgeometric ergodic Markov dynamic.*

  As explained section 2.6, theorems ensuring convergence of stochastic approximations with Markovian dynamic require the Markov Chain to be geometrically ergodic. In the chapters 3 and 4, we will use stochastic approximations with a Markovian dynamic obtained from Metropolis Hastings algorithms. However, we know that, when targeting distributions with heavy tails, those Markov Chains can be subgeometric ergodic (Douc et al., 2004; Fort and Moulines, 2000, 2003; Jarner and Hansen, 2000). Thus, the theoretical guarantees on the convergence of those stochastic approximations are no longer met.
  Hence, in chapter 5, we choose to relax the condition of geometric ergodicity. We propose a more general set of hypotheses, under which we prove the convergence of stochastic

approximations with subgeometric Markovian dynamics. They are essentially about the rate of convergence of the Markov Chain and the regularity of its kernel. Most of the polynomial rates of convergence satisfy these assumptions. We then use this new set of hypotheses to prove the convergence of two stochastic approximations. The first one is a Metropolis Hastings algorithm where the variance of the proposal is adapted along iterations. In the second example, we consider the independent component analysis model where distributions with positive heavy tails lead to a subgeometric ergodic Markov Chain in a SAEM-MCMC algorithm.
The work done in that chapter has been published in the Electronic Journal of Statistics (Debavelaere et al., 2021).

- **Chapter 6:** *On the curved exponential family in the Stochastic Approximation Expectation Maximization Algorithm.*

  We have presented section 2.7 the different conditions ensuring the convergence of the SAEM and MCMC-SAEM algorithm. Among those hypotheses, one of the most restrictive is the necessity for the joint likelihood to belong to the curved exponential family. However, this hypothesis is not always verified, for instance for heteroscedastic models. In that case, Kuhn and Lavielle (2005) propose to transform the statistical model to make it exponential. Their solution consists at considering the parameters $\theta$ of the initial model as additional latent variables following a Normal distribution centered on a new parameter $\bar{\theta}$ and with fixed variance $\sigma^2$. Instead of estimating $\theta$ we then estimate its mean $\bar{\theta}$. If this method is often used, there is in fact no guarantee that $\bar{\theta}$ will be close to the parameter of the initial model.
  In chapter 6, we show that using this method can introduce a bias in the estimation of the maximum of likelihood. We then prove that this bias tends to zero when the variance $\sigma^2$ goes to zero and give an upper bound for $\sigma$ small. However, on a numerical example, we see that a compromise must be made between the error in the estimation and the computation time. Even worse, for very small values of $\sigma$ (and so, theoretically, small errors), the algorithm does not converge numerically. To overcome this problem, we propose in the last part of this chapter a new algorithm allowing a better estimation of the maximum likelihood with a reasonable computation time.
  This work has been submitted (Debavelaere and Allassonnière, 2021).

# Part II

# Atlases on Riemannian manifolds

# CHAPTER 3

## Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories

Given a longitudinal data set, this chapter introduces a new model allowing to learn a classification of the shapes progression in an unsupervised setting: we automatically cluster a longitudinal data set in different classes without labels. Our method learns for each cluster an average shape trajectory (or representative curve) and its variance in space and time. Representative trajectories are built as piecewise geodesics. This mixture model is flexible enough to handle independent trajectories for each cluster as well as fork and merge scenarios. This new formulation allows, for example, to consider subjects that deviate from a normal ageing at a certain rupture point.
The estimation of such non linear mixture models in high dimension is known to be difficult because of the trapping states effect that hampers the optimisation of cluster assignments during training. We address this issue by using a tempered version of the stochastic EM algorithm.
Finally, we apply our algorithm on different data sets. First, synthetic data are used to show that a tempered scheme achieves better convergence. We then apply our method to different real data sets: 1D RECIST score used to monitor tumors growth, 3D facial expressions and meshes of the hippocampus. In particular, we show how the method can be used to test different scenarios of hippocampus atrophy by using an heteregenous population of normal ageing individuals and mild cognitive impaired subjects.

This chapter uses notions of Riemannian manifolds and the LDDMM framework which are quickly presented. For more information on those notions, we refer to the sections 2.3 and 2.4 of the introduction.
This work has been published in the International Journal of Computer Vision (Debavelaere et al., 2020).

### Contents

## 3.1 Introduction

The emergence of large longitudinal data sets (subjects observed repeatedly at different time points) has allowed the construction of different models improving the understanding of biological or natural phenomenon. Longitudinal studies have numerous applications: understating of the differences of progression in neurodegenerative disease such as Alzheimer's, chemotherapy monitoring, facial recognition, etc.. Such medical studies enable to retrieve the global progression of the disease while explaining the inter subject variability. In particular, it would be interesting to highlight the influence of a disease on a normal ageing process and to be able to differentiate those two processes. Clinicians are also interested in the possibility to detect the moment when a disease begins to manifest itself, i.e. the moment at which a subject branches from the normal dynamic. For instance, in the case of the Alzheimer's disease, we still do not know if the disease has a very early genesis, leading to a specific aging pattern from an early age or if it is a sudden deviation from the normal ageing process. Another example is the monitoring of tumors along treatment. Indeed, it is well known that the whole population will not react the same way to a given drug. Therefore, clustering patients would enable a specific care. In both situations, the evolution may not be smooth in the sense that the disease can show variations in dynamics according to the stage of its development. To tackle those problems, we consider that populations can follow different dynamics over time. Moreover, in order to detect subgroups with specific patterns, we implement an unsupervised clustering of the dataset. Here, our populations are therefore heterogeneous but without prior knowledge on the sub-groups composing them, thus preventing from the use of supervised approaches.

We design our model such that it is able to detect a certain fixed number of different dynamics in the population and, for each of them, to estimate a representative trajectory of that population together with the inter subjects variability. The difficulty is in fact further increased in this spatiotemporal setting since clustering may take various forms: sub-groups may follow independent trajectories, or they may follow trajectories that fork or merge at specific time-points. The former case is relevant to discover pathological sub-types having different disease course. The latter is interesting for a disease that is seen as a progressive deviation from a normal aging scenario.

Usually, shape spaces are built by considering shape data as points on a Riemannian manifold (for instance, Kendall spaces (Kendall, 1984), currents (Vaillant and Glaunès, 2005) or varifolds (Charon and Trouvé, 2013)). In such shape spaces, descriptive (Donohue et al., 2014) or generative (Jedynak et al., 2012; Durrleman et al., 2013; Allassonnière et al., 2015) models have been constructed. To deform the shapes, different frameworks can be used, among others diffeomorphic demons (Vercauteren et al., 2009) or the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework. We will here use the last. It allows us to compute the deformation from one shape to the other by coding deformations as geodesics on a Riemannian manifold and using flows of deformations (Miller et al., 2006). Given a data set of shapes, it is then possible to construct an atlas. An atlas is composed of a shape that is representative of the population, as well as the spatial variability within this population (Fletcher, 2013; Allassonnière and Kuhn, 2010; Lorenzen et al., 2005; Su et al., 2014). The next logical step is to handle longitudinal data sets. Once again, the trajectory of a shape from one time point to the other will be constructed by using flows of diffeomorphisms (Bône et al., 2018a; Lorenzi et al., 2011; Singh et al., 2016; Muralidharan and Fletcher, 2012; Kim et al., 2017; Chakraborty et al., 2017). In this framework, a longitudinal atlas consists of a representative trajectory, or template, and of the spatiotemporal variability of the population. The representative trajectory is a long-term

scenario of changes informed by sequences of short-term individual data. It can be seen as a geodesic (Bône et al., 2018a; Schiratti et al., 2017) or a piecewise geodesic (Allassonniere et al., 2017) curve on the manifold. For instance in the case of a sphere, a geodesic on the manifold is just a great circle. Spatial and temporal deformations are then considered to generate subjects from this representative trajectory. In particular, the temporal reparametrization can be considered as a general diffeomorphism (Su et al., 2014) or as an affine reparametrization combining acceleration and offset coefficients (Bône et al., 2018a).

All these methods however assumed that observations are drawn from an homogeneous population that may be summarized by a single representative trajectory. Several clustering methods have already been proposed to create atlases from cross sectional datasets in an unsupervised way (Allassonnière and Kuhn, 2010; Srivastava et al., 2005) or for longitudinal datasets of continuous trajectories in a supervised way (Abdelkader et al., 2011). However, if Hong et al. (2015) proposes a test to detect if there is one cluster or more in a longitudinal population, there is, to our knowledge, no paper proposing a method to detect those clusters in an unsupervised way in the longitudinal framework while also creating the corresponding atlases. This will be one of the goals of this paper. Our algorithm should be able to detect sub populations that could be different from those expected and so highlight unexpected dynamics. Such a behaviour can be interesting to test different models or to highlight in a population some characteristics that were previously considered without influence on the phenomenon under study.

In this chapter, we tackle the case where the population is supposed to contain a certain fixed number of unknown clusters. To tackle this problem, we construct a mixed-effect generative model. To estimate the different parameters, we choose to use a variant of the Expectation-Maximization algorithm called the Markov Chain Monte Carlo Stochastic Approximation Expectation Maximization algorithm (MCMC-SAEM) (Delyon et al., 1999; Allassonnière et al., 2010). However, using those algorithms in a clustering context leads to the problem of trapping states: changing class assignment is often more costly than adjusting the parameters of the current clusters, resulting in very few updates of class assignment during optimization. Solutions have already been presented in the case of cross sectional data sets analysis but at very high computational costs (Allassonnière and Kuhn, 2010). Here, we choose to introduce temperate distributions in our Expectation-Maximization algorithm to avoid being trapped in the initial labelling.

In this paper, we will first explain in section 3.2 the geometrical framework allowing us to compute the representative trajectories and deformations towards the subjects. Because this framework allows us to define our model by a finite number of parameters, we will present in section 3.3 the statistical model and the algorithm used to estimate those parameters. Finally, we will apply our work to different data sets. We will quantitatively validate it on simulated 2D data. We will then perform experiments on real data: we will work with 1D RECIST score used to monitor the growth of a tumor (Therasse et al., 2000), with a data set of 3D faces expressing different expressions and with a 3D data set of hippocampi of patients with or without Alzheimer's disease.

## 3.2 Geometrical model

We will first present the geometrical model that allows us to compute the representative trajectory of each of our clusters as well as the deformations towards the subjects.

### 3.2.1 Construction of the representative trajectory

In the following, we consider a longitudinal data set of $n$ subjects, each being observed $k_i$ times: $(y_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$ at time $(t_{i,j})_{1\leq i\leq n, 1\leq j\leq k_i}$, where each observation $y_{i,j}$ is a point of $\mathbb{R}^d$, $d \in \mathbb{N}$.

We first want to explain how to construct a longitudinal trajectory in a set of shapes that will, later on, define our group average. We choose to use the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework to define our shape deformations. Therefore, we can deform an initial shape using the flow of a velocity $v_t \in V$ for $t \in [0,1]$ and for $V$ a fixed Hilbert space:

$$\begin{cases} \dfrac{\partial \phi_t^v}{\partial t} = v_t \circ \phi_t^v \\ \phi_0^v = Id\,. \end{cases} \tag{3.1}$$

Given velocities $(v_t)_{t\in[0,1]}$, this equation creates diffeomorphisms $(\phi_t^v)_{t\in[0,1]}$ that will deform the ambient space and so, in particular, our initial shape $y_0$. Hence, given velocities $(v_t)_{t\in[0,1]}$, $(\phi_t^v(y_0))_{t\in[0,1]}$ will define a longitudinal trajectory of shapes.

Each of those diffeomorphism $\phi_t^v$ belongs to the set $\mathcal{G} = \{\phi_1^v | v \in V\}$. This group of deformation maps is provided with a right invariant metric via

$$d(Id, \phi) = \sqrt{\inf\left\{\int_0^1 ||v_t||_V^2 dt | \phi = \phi_1^v\right\}}\,.$$

This exactly states that $\mathcal{G}$ is given the structure of a manifold on which distances are computed as the length of minimal geodesic paths connecting two elements. Given this structure, we will no longer allow any diffeomorphism to be our group average but only diffeomorphisms such that $t \mapsto \phi_t^v$ follows a geodesic path in $\mathcal{G}$.

We need now to ask ourselves how to choose velocities verifying this condition. Since we only study discrete shapes, we can place ourselves in the finite dimensional setting and suppose that our velocities $(v_t)_{t\in[0,1]}$ belong to a Reproducing Kernel Hilbert Space $V$ with kernel $K_g$. $V$ is in fact the set of squared integrable functions regularized by the convolution by the kernel $K_g$. A vector $v$ in $V$ can then be written using a set of $n_{cp}$ control points $(c_i)_{1\leq i\leq n_{cp}}$ and momentum vectors $(m_i)_{1\leq i\leq n_{cp}}$ in $\mathbb{R}^d$: for $x \in \mathbb{R}^d$,

$$v(x) = \sum_{i=1}^{n_{cp}} K_g(c_i, x) m_i\,. \tag{3.2}$$

The value of $v$ at a point $x$ is obtained as the interpolation of the momenta at the control points. Hence, to create a longitudinal trajectory, we now need to choose an initial shape and a set of control points and momenta defining the velocities $(v_t)_{t\in[0,1]}$ such that $(\phi_t)_{t\in[0,1]}$ defines a geodesic in $\mathcal{G}$.

It has been shown in Miller et al. (2006) that if the initial velocity field $v_0$ is the interpolation of momentum vectors at control points as in Eq. (3.2), then the velocity field defining a geodesic path in $\mathcal{G}$ keeps the same form:

$$v_t(x) = \sum_{i=1}^{n_{cp}} K_g(c(t)_i, x) m(t)_i\,. \tag{3.3}$$

Moreover, $m(t)$ and $c(t)$ are then time dependent momenta and control points solutions of the Hamiltonian equations:

$$\begin{cases} \dot{c}(t) = K_g(t)m(t) \\ \dot{m}(t) = \nabla_{c(t)} \left( m(t)^T K_g(t) m(t) \right) \end{cases} \tag{3.4}$$

with initial conditions $m(0) = (m(0)_k)_{1 \leq k \leq n_{cp}}$, $c(0) = (c(0)_k)_{1 \leq k \leq n_{cp}}$ and where $K_g(t)$ is the $n_{cp} \times n_{cp}$ kernel matrix $(K_g(c_i(t), c_j(t)))_{1 \leq i,j \leq n_{cp}}$.

To sum up, to define our longitudinal trajectory of shapes, we now only need to set an initial shape and an initial set of momenta and control points. By integrating the Hamiltonian equations (3.4), one can compute the evolution of those control points and momenta over time and obtain the velocity vector at any time $t$ (Eq. (3.3)). By integrating the flow equation (3.1), we obtain diffeomorphisms $(\phi_t))_{t \in [0,1]}$ deforming the ambient space. By applying this diffeomorphism at a point cloud or mesh $y_0$, we are finally able to deform it.

We finally note $\mathcal{E}xp_{c_0,t_0,t}(m_0) = \phi_t^v$ the diffeomorphism obtained above with the initial condition $\phi_{t_0}^v = Id$. This deformation process involving the Riemaniann Exponential is showed on an example figure 3.1.

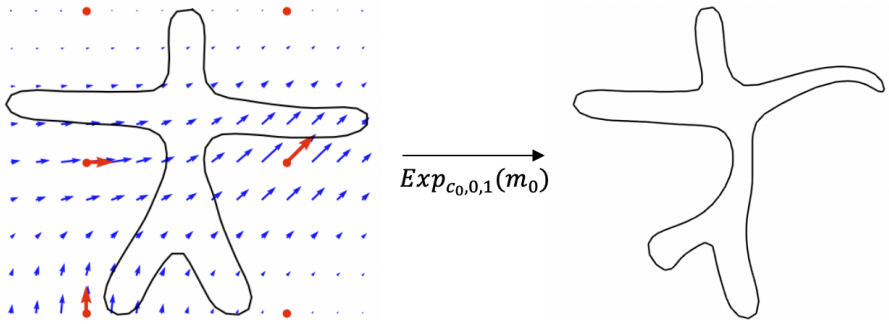



**Figure 3.1** The initial control points are the red points, the initial momenta, the red vectors. The blue vector field is created using the initial momenta and control points. Finally, we compute the deformation of the initial shape by this vector field.

However, in order to deal with possible change of dynamics in the population, we do not only want to consider geodesics but piecewise geodesics. Hence, we will modelize our group trajectories as a combination of $K$ different geodesics following each other, generalizing the work done in Allassonniere et al. (2017) in dimension $1$. In particular, each of the geodesics defining $\gamma_0$ describes a dynamic of the population on a particular time segment, different from the others. The time at which the group average goes from one dynamic to the other will be called rupture times. The component of the piecewise geodesic following a rupture time will then be defined using the Exponential operator defined previously, applied at the value of the trajectory at that rupture time.

We now formalize this: we introduce a subdivision of $\mathbb{R}$: $(t_{R,1} < ... < t_{R,K-1} < t_{R,K} := +\infty)$ where $(t_{R,k})_{1 \leq k \leq K-1}$ are called rupture times i.e. times when the representative curve switches from one geodesic to another. It is at those times that the population switches from one dynamic

to the other. Given a set of initial control points $c^1 \in \mathbb{R}^{n_{cp} \times d}$, of rupture times $t_R \in \mathbb{R}^{K-1}$, an initial shape $x^1$ and $K$ momenta $(m^0, m^1, ..., m^{K-1})$, we define the representative trajectory as:

$$\begin{cases} \gamma(t)(x^1) = \mathcal{E}xp_{c^1, t_{R,1}, t_{R,1}-t}(m^0) \cdot x^1 \mathbb{1}_{t \leq t_{R,1}} \\ \qquad + \sum_{k=1}^{K-1} \mathcal{E}xp_{c^k, t_{R,k}, t-t_{R,k}}(m^k) \cdot x^k \mathbb{1}_{t_{R,k} \leq t \leq t_{R,k+1}} \\ \text{with, for } k \geq 2 : \\ \qquad c^k = \mathcal{E}xp_{c^{k-1}, t_{R,k-1}, t_{R,k}-t_{R,k-1}}(m^{k-1}) \cdot c^{k-1} \\ \qquad x^k = \mathcal{E}xp_{c^{k-1}, t_{R,k-1}, t_{R,k}-t_{R,k-1}}(m^{k-1}) \cdot x^{k-1} \end{cases}$$

Here, the $c^k$ and $x^k$ are respectively the position of the control points and the value of the representative curve at times $t_{R,k}$. For $k \geq 2$, they are fixed to assure the continuity of the trajectory. It can be noticed that the first rupture time has a particular role as we must define a geodesic before it, determining the trajectory from $-\infty$ to the first rupture time and another after it, determining the trajectory from the first rupture time to the second. The control points $c^1$ and momenta $m^0$, $m^1$ are used to compute the velocities at the time $t_{R,1}$ defining the geodesic before and after it. The other momenta $m^2, ..., m^{K-1}$ and control points $c^2, ..., c^{K-1}$ define the subsequent geodesics.

The construction of a piecewise geodesic is applied on an example figure 3.2.

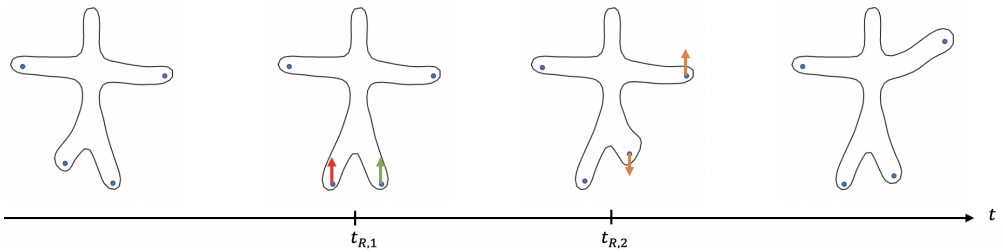



**Figure 3.2** Example of a piecewise geodesic with 3 parts. At the first rupture time $t_{R,1}$, the blue control points and red momenta code the exponential before it. The green momenta codes the exponential after the first rupture time. Both the control points and the shape are transported by this diffeomorphism until the second rupture time $t_{R,2}$. It is this transported shape and those transported control points that will be used, along with the orange set of momenta, to compute the deformation after the second rupture time.

### 3.2.2 Deformations towards the subjects

We now know how to construct a longitudinal trajectory that will play the role of a representative trajectory. From this representative trajectory featuring the group characteristic path, we want to generate individual trajectories following different behaviours. To achieve this goal, we take into account both temporal and spatial differences by introducing a time reparametrization and a diffeomorphic spatial deformation.

### Time reparametrization

Each individual can follow its own rhythm of progression, different from the representative curve and varying from one time segment to another, hence the need to introduce time reparametrizations.
For each subject $i$, let $\xi_{i,0}, ...\xi_{i,K-1}$ be acceleration coefficients and $\tau_{i,0}, ..., \tau_{i,K-1}$ time shifts. We write for every subject $i$:

$$\psi_{i,0}(t) = t_{R,1} - e^{\xi_{i,0}} (t_{R,1} - t + \tau_{i,0}) \tag{3.5}$$

and, for each time segment $k \geq 1$,

$$\psi_{i,k}(t) = t_{R,k} + e^{\xi_{i,k}} (t - t_{R,k} - \tau_{i,k}) \ . \tag{3.6}$$

$\psi_{i,k}$ codes the temporal reparametrization of the subject $i$ on the time segment $k$. Once again, a first time reparametrization must be defined before the first rupture time.
The time shifts $\tau_{i,k}$ are offsets that allow the subjects to be at different stage of evolution while the acceleration factors $\xi_{i,k}$ allow an inter-subject variability in the pace of evolution on each geodesic (quicker evolution if $\xi_{i,k} > 0$, slower if $\xi_{i,k} < 0$). Both of those factors allow us to represent behaviors in the population observed by clinicians.

Different conditions must be verified to assure the continuity of the time reparametrizations. First, as the representative trajectory goes through a change of dynamics at the rupture times, each subject has its own rupture times $t_{R,i,k}$ such that $t_{R,k} = \psi_{i,k}(t_{R,i,k})$ i.e. $t_{R,i,k} = t_{R,k} + \tau_{i,k}$. Before the individual rupture time $t_{R,i,k}$, the time reparametrization is computed using $\psi_{i,k-1}$ and after it, using $\psi_{i,k}$. Hence, to assure the continuity of the global time reparametrization at each of those rupture times, we also fix all the time shifts but $\tau_{i,0}$ by continuity conditions: we impose for all $k$ $\psi_{i,k-1}(t_{R,i,k}) = \psi_{i,k}(t_{R,i,k})$, i.e.: $\tau_{i,0} = \tau_{i,1}$ and, for $k \in [|2, K-1|]$,

$$\tau_{i,k} = \tau_{i,k-1} + (t_{R,k} - t_{R,k-1})(e^{-\xi_{i,k-1}} - 1) \, . \tag{3.7}$$

From now on, we note $\tau_i = \tau_{i,0}$.

It can be remarked that the choice of this particular temporal reparametrization simplifies the computations needed to assure the continuity of the final trajectory at each of the rupture time. Indeed, if we had chosen, on each component, a diffeomorphic temporal reparametrization without constraint (as done in Su et al. (2014) in the geodesic case), more complex equalities should have been imposed at each of the individual rupture times. This reparametrization has also the advantage to be easily interpreted.
Finally, we set:

$$\psi_i(t) = \psi_{i,0}(t)\mathbb{1}_{t \leq t_{R,i,1}} + \sum_{k=1}^{K-1} \psi_{i,k}(t)\mathbb{1}_{t_{R,i,k} \leq t \leq t_{R,i,k+1}} \, .$$

To summarize, those equations mean that the subject $i$ at the instant $t$ is obtained from the representative trajectory shifted by $\tau_i$ and accelerated on each time segment by $e^{\xi_{i,k}}$. The time reparametrization process is summarized figure 3.3.

### Space deformations

Concerning the space deformations, as proposed in Bône et al. (2018a), we will account the space variability by using exp-parallelizations, i.e. the generalization of parallelism to geodesically complete manifolds (Schiratti et al., 2015). More precisely, we introduce for each subject

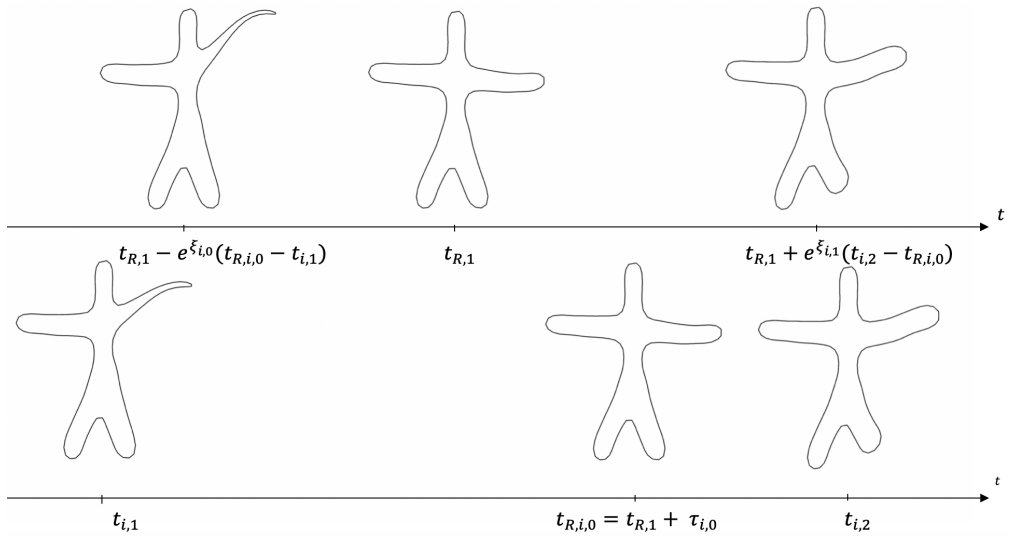



**Figure 3.3** Example of a time reparametrization. At the top, the representative trajectory. At the bottom, a time reparametrization towards the subject $i$ observed at two times: $t_{i,1}$ and $t_{i,2}$. The individual rupture time of the subject $i$ is obtained as a translation of the rupture time by $\tau_{i,0}$, here chosen positive. On the first time segment, $\xi_{i,0}$ is negative and the progression is slower than the one of the representative trajectory. On the second time segment, $\xi_{i,1}$ is positive and the progression is quicker.

$i$ a space-shift momentum $w_i$. We note $P_\gamma(w)$ the parallel transport which transports any vector $w \in \mathbb{R}^{n_{cp} \times d}$ along the trajectory $\gamma$. Practically, we compute it using the fanning scheme (Louis et al., 2017). Then, to code the deformation field at a time $t$, we transport the momentum $w$ along the curve $\gamma(t)$ and then compute the flow given by this new momentum. The given trajectory is the exp-parallelization of $\gamma$ by $w_i$. Hence, we define:

$$\eta_t(w) = \mathcal{E}xp_{\gamma(t)(c^1),0,1}(P_{\gamma(t)}(w)) \,.$$

Finally, given $x^1$ the value of the representative curve at the first rupture time, the deformation of the representative curve $\gamma$ by the space shift $w$ is given by:

$$\gamma_w(t) = \eta_t(w) \circ \gamma(t) \circ x^1 \,.$$

We give examples of the space deformation process first on Fig. 3.4 by computing the exp-parallelization of a trajectory on a sphere and then on Fig 3.5 by presenting an example in a space of shapes.

We model this space shift as a linear combination of $n_s$ sources: we suppose that $w = As$ with $A$ a $n_{cp} \times n_s$ matrix called the modulation matrix and $s \in \mathbb{R}^{n_s}$ the sources. This matrix plays the role of the source separation matrix also known as the modulation matrix in the Independent Component Analysis. This helps to reduce the dimension by highlighting the principal sources of deformation. By projecting all the columns of $A$ on $(m^0, ..., m^{K-1})^\perp$ for the metric $K_g$, we impose orthogonality between the deformations towards the subjects and the velocity field defining our representative trajectory. It has been shown in Schiratti et al. (2017) that this condition is necessary to assure the identifiability of the model by preventing the algorithm to

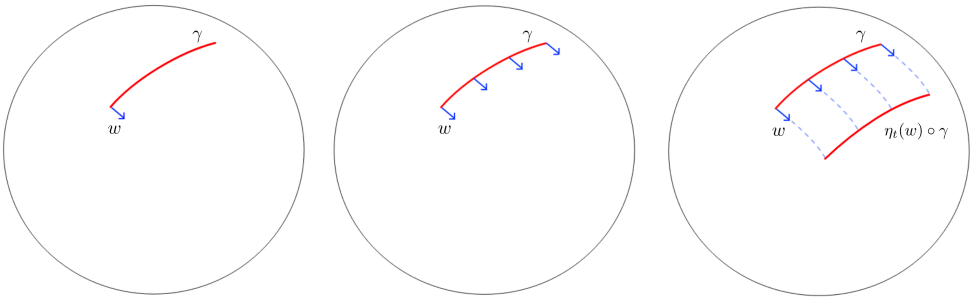



**Figure 3.4** Example of parallel transport on a sphere. On the left, we draw a trajectory $\gamma$ and the momenta to transport $w$. On the center, we transport $w$ along $\gamma$. On the right, we compute the exp-parallelization of $\gamma$ by $w$.

consider an acceleration with respect to the representative trajectory as a space shift. Finally, we deform the template $\gamma(t)(x^1)$ by setting:

$$\gamma_i(t) = \gamma_w(\psi_i(t)) \,.$$

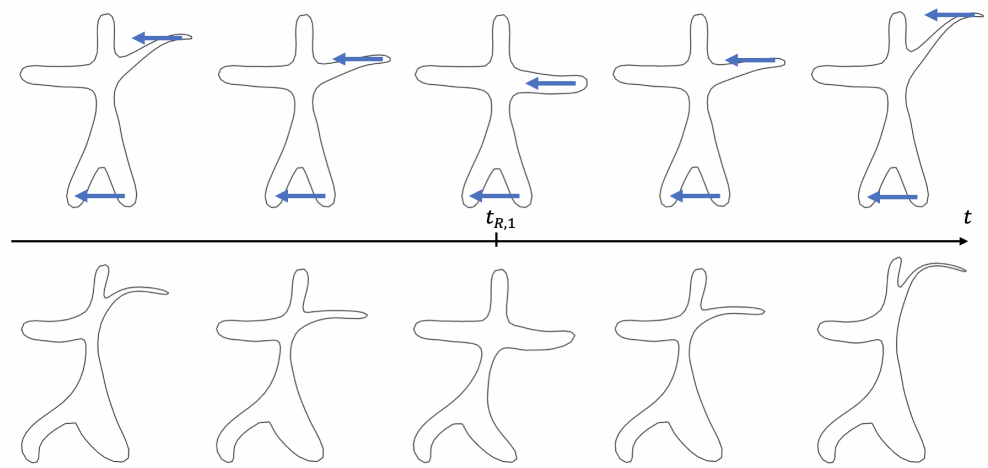



**Figure 3.5** Samples from a piecewise geodesic (top) and a parallel deformation (bottom). The blue momenta is first defined at the rupture time $t_{R,1}$. It is then transported along the piecewise geodesic and defines the deformation frame towards a subject.

### 3.2.3 Mixture and branching process

This construction builds a piecewise geodesic model of progression. Until now, it can only process homogeneous populations. We propose an extension for the analysis of heterogeneous populations. More precisely, we suppose there exists $N$ different representative curves in a given population, each of the subjects $i$ being in the cluster $cl(i)$ defined by the particular representative curve $\gamma^{cl(i)}$. This representative curve comes with its own set of rupture times $(t_{R,1}^{cl(i)} <$

$... < t_{R,K-1}^{cl(i)}$), initial shape $x^{1,cl(i)}$, control points $c^{1,cl(i)}$, momenta $(m^{0,cl(i)}, ..., m^{K-1,cl(i)})$ and modulation matrix $A^{cl(i)}$.

This mixture framework enables to compare and test hypothesis on the clusters. For instance, some of the time segments can be shared by several clusters. This imposes the representative curves of these clusters on these time segments to be the same. In particular, if we want some of the clusters to be equal on the first time segment, we impose $t_{R,1}^k$, $x^{1,k}$, $c^{1,k}$ and $m^{0,k}$ to be the same for these clusters. This allows us to handle populations forking or merging at the rupture times. The rupture times are then not only times when a change of dynamic occurs but also times when populations fork or merge.

Hence, we have presented a complex geometrical model allowing us to compute global trajectories and the deformations towards subjects. Those global trajectories can take a wide variety of forms. But, in all cases, our model is parameterized by a finite number of parameters. Hence, the next step is to construct a statistical model to estimate the unknown variables. We will need to estimate the parameters defining the template as well as the clusters and the parameters defining the deformations towards the subjects. This is the goal of the next section: in section 3.3.1, we will present the statistical model considered while in section 3.3.2 we will explain how to estimate the parameters defining it.

## 3.3 Statistical Model and estimation

### 3.3.1 Statistical Model

We define a mixed effects statistical model allowing us to estimate those different parameters. We note:

$$z_{pop}^r = \left((m^{k,r}, t_{R,k}^r)_{0 \leq k \leq K-1}, x^{1,r}, c^{1,r}, A^r\right)$$

the population parameters of the cluster $r$ and

$$z_i = ((\xi_{i,k})_{0 \leq k \leq K-1}, \tau_i, s_i)$$

the deformation parameters of the subject $i$ with $\xi_{i,k}$ the acceleration parameters, $s_i$ the sources and $\tau_i$ the first time shift. As for the other time shifts they are fixed by continuity conditions (cf Eq. (3.7)).

We suppose that the subject $i$ is obtained as a noisy deformation of the representative curve $\gamma^{cl(i)}$: $\forall i \in [|1, n|]$, $\forall j \in [|1, k_i|]$,

$$y_{i,j} | cl(i), z_{pop}^{cl(i)}, z_i \sim \mathcal{N}(\gamma_i(t_{i,j}), \sigma^2 Id)\,.$$

Such a notation implies that we are able to compute the distance between two different shapes. Depending on the application, the points constituting the shape will be labeled or not. In the first case we will be able to use a landmark distance. In the other, we will use the current (Vaillant and Glaunès, 2005) or varifold (Charon and Trouvé, 2013) distances.

We also suppose that the deformation parameters $z_i$ verify:

$$z_i | cl(i) \sim \mathcal{N}(0, \Sigma^{cl(i)})$$

where for all cluster $r$, $\Sigma^r$ is a positive-definite matrix.

The cluster $r$ is drawn with a probability $p^r$ i.e.

$$cl(i) \sim \sum_{r=1}^{N} p^r \delta_r \,.$$

To estimate the parameters of our model, we choose to apply a SAEM algorithm. However, this algorithm requires the joint distribution to belong to the curved exponential family (see section 2.7 of the introduction). This is not the case if we suppose $z_{pop}$ to be a parameter. To overcome this problem, we apply a trick first proposed in Kuhn and Lavielle (2005) and suppose that $z^r_{pop} \sim \mathcal{N}(\bar{z}^r_{pop}, v_{pop})$ where $v_{pop}$ are small fixed variances. This new model now belongs to the curved exponential family. We thus consider the population variables as random effects and estimate their mean $\bar{z}_{pop}$. This trick and its influence on the maximum of likelihood returned by the SAEM algorithm is further discussed chapter 6.

Finally, our model is defined with parameters $\theta = \big((\Sigma^r, p^r, \bar{z}^r_{pop})_{1\leq r\leq N}, \sigma\big)$.

For effectiveness in the high dimension low sample size setting, we work in the Bayesian framework and set the usual conjugate priors:

$$\begin{cases} \Sigma^r \sim \mathcal{W}^{-1}(V, m_\Sigma) \\ p \sim \mathcal{D}(\alpha) \\ \bar{z}^r_{pop} \sim \mathcal{N}(\bar{\bar{z}}^r_{pop}, \bar{v}_{pop}) \\ \sigma \sim \mathcal{W}^{-1}(v, m_\sigma) \end{cases} \tag{3.8}$$

where $\mathcal{W}$ is the inverse Wishart distribution, $\mathcal{D}$ is the Dirichlet distribution and $V$, $m_\Sigma$, $v$, $m_\sigma$, $\alpha$, $\bar{\bar{z}}^r_{pop}$ and $\bar{v}_{pop}$ are hyperparameters of the model.

Finally, if we note $\Lambda$ the dimension of the space in which the residuals $\|y_{i,j} - \gamma_i(t_{i,j})\|^2$ are computed, the complete log-likelihood writes:

$$\begin{aligned} \log q(y, z_{pop}, z, c, \theta) = &- \sum_{i=1}^{n} \left( \sum_{j=1}^{k_i} \frac{1}{2\sigma^2} \|y_{i,j} - \gamma^{cl(i)}(t_{i,j})\|^2 - \frac{\Lambda k_i}{2} \log(\sigma^2) \right) \\ &- \frac{1}{2} \sum_{r=1}^{N} (z^r_{pop} - \bar{z}^r_{pop})^T v^{-1}_{pop} (z^r_{pop} - \bar{z}^r_{pop}) \\ &- \frac{1}{2} \sum_{i=1}^{n} \left( z_i^T (\Sigma^{cl(i)})^{-1} z_i - \log |\Sigma^{cl(i)}| \right) \\ &+ \sum_{i=1}^{n} \log p^{c(i)} + \sum_{r=1}^{N} \alpha \log p^r + \sum_{r=1}^{N} \left( \frac{m_\Sigma}{2} (\log |V| - log|\Sigma_r|) - tr(V\Sigma_r^{-1}) \right) \\ &+ m_\sigma \log\left(\frac{v}{\sigma}\right) - \frac{m_\sigma}{2} \left(\frac{v}{\sigma}\right)^2 - \frac{1}{2} \sum_{r=1}^{N} (\bar{z}^r_{pop} - \bar{\bar{z}}^r_{pop})^T \bar{v}^{-1}_{pop} (\bar{z}^r_{pop} - \bar{\bar{z}}^r_{pop}) + cste \,. \end{aligned} \tag{3.9}$$

It is important to note that our model belongs to the curved exponential family and so allows us to define sufficient statistics. For any class $r$, wet set:

$$\begin{cases} S_1^r(y,z,z_{pop}) = z_{pop}^r \\ S_2^r(y,z,z_{pop}) = \sum_{i=1}^{n} \mathbb{1}_{cl(i)=r} \\ S_3(y,z,z_{pop}) = \sum_{i=1}^{n} k_i \\ S_4^r(y,z,z_{pop}) = \sum_{i=1}^{n} \mathbb{1}_{cl(i)=r} z_i^t z_i \\ S_5(y,z,z_{pop}) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} \|y_{i,j} - \gamma_i(t_{i,j})\|^2 \end{cases} \tag{3.10}$$

It will then be possible, in the next section, to estimate the parameters of our algorithm using only those sufficient statistics.

### 3.3.2 Estimation

To estimate the parameters $\theta$, we want to compute a maximum a posteriori estimator by using a stochastic version of the Expectation Maximization algorithm known as MCMC-SAEM (Allassonnière and Kuhn, 2010). It consists in the following steps:

(i) Simulation of $(z, z_{pop}, cl)$.

(ii) Stochastic approximation of the sufficient statistics of the curved exponential model.

(iii) Maximization using the updated stochastic approximation.

Concerning the sampling in step (i), we simulate $(z, z_{pop}, cl)$ as an iterate of an ergodic Monte Carlo Markov Chain with stationary distribution $q(z_{pop}, z, cl|y, \theta)$. More precisely, we use a symmetric random walk Monte-Carlo Markov Chain within Gibbs sampler with adapted variance. Once those variables are sampled, it is then possible to compute the sufficient statistics and to obtain the parameters maximizing the posterior distribution in a closed form, as explained below.

In step (ii), we compute a stochastic approximation of the sufficient statistics using Eq. (3.10) and a decreasing positive sequence of step size $(\Delta_k)_{k\in\mathbb{N}}$: if $m$ is the current iteration of our algorithm, $1 \leq r \leq N$ and $1 \leq k \leq 5$, we compute $s_{m,k}^r = s_{m-1,k}^r + \Delta_{m-1}(S_k^r(y,z,z_{pop}) - s_{m-1,k}^r)$.

Finally, in step (iii), the update of the parameters $\theta$ in the maximization step of the MCMC-SAEM at iteration $m$ can be derived: for all $1 \leq r \leq N$,

$$\begin{cases} \bar{z}_{pop}^r = \dfrac{\bar{v}_{pop} s_{m,1}^r + v_{pop} \bar{\bar{z}}_{pop}^r}{v_{pop} + \bar{v}_{pop}} \\ \Sigma_r = \dfrac{s_{m,4}^r + m_\Sigma V}{s_{m,2}^r + m_\Sigma} \\ \sigma^2 = \dfrac{s_{m,5}^r + m_\sigma v}{\Lambda s_{m,3}^r + m_\sigma} \\ p^r = \dfrac{s_{m,2}^r + \alpha}{n + \alpha N} \end{cases} . \tag{3.11}$$

However, using the algorithm as presented above yields to bad results in exploring the support of the conditional probability distribution. This issue is known as trapping states: once a label is given to an observation, the probability of changing to another is almost zero. This leads to no change of cluster after a few iterations. This problem has already been encountered in the clustering case, for instance in Allassonnière and Kuhn (2010) and Srivastava et al. (2005). In the first case, the authors chose to compute deformations from each template towards each subject leading to very high computational cost. In the second paper, the authors used tempered distributions but only determine the clusters without the associated representative curve and inter-subjects variability.
Here, to solve this problem, we use a tempered version of the MCMC-SAEM. Instead of targetting $q(c|y,\theta)$ in the MCMC step, we rather sample from an ergodic Markov Chain with density $\frac{1}{C(T_k)}q(c|y,\theta_k)^{\frac{1}{T_k}}$ where $k$ is the current iteration of the algorithm, $T_k$ is a sequence of temperature converging towards $1$ and $C(T_k)$ is the normalizing constant. The higher the temperature, the flatter the distribution and the more the clusters are likely to explore the entire set.

Finding a good distribution of temperatures such that meaningful representative curves are found without immediately fixing the clusters nor forcing them to move throughout the whole algorithm is quite difficult. Several choices have been proposed in Allassonnière and Chevallier (2019) but we choose here a distribution that takes into account the current state of the algorithm. For each subject $i$ and each cluster $k$, we set $\tau_i^k = \log\left(\frac{q(cl_(i)=j)}{q(cl(i)=k)}\right)$ where $cl(i)$ is the cluster of the subject $i$, $j$ the index of that cluster during the previous iteration and $q$ is the complete log likelihood. $\tau_i^k$ is in fact the logarithm of the acceptance rate of the MCMC-SAEM algorithm for the subject $i$ to go from the cluster $j$ to the cluster $k$. We then take:

$$T = \begin{cases} \dfrac{\text{Median}(\tau)}{\lceil \text{iter}/10 \rceil}\dfrac{5-\text{iter}\%10}{5} + 1 - \dfrac{5-\text{iter}\%10}{5} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if iter}\%10 < 5 \\ 1 \text{ otherwise} \end{cases} \tag{3.12}$$

where $\%$ is the modulo operator and iter is the current iteration.

Such a distribution of temperature allows the representative curves to fix themselves when iter$\%10 \geq 5$ while forcing the clusters to explore the whole space when iter$\%10 < 5$. Indeed, such a temperature distribution allows us to directly influences the acceptance rate of the clusters.

If this temperature scheme allows us to observe meaningful clusters, as showed later in section 3.4, it must be remarked that it depends of the acceptance rate $\tau$ and so of the previous state of the algorithm. The convergence of tempered SAEM algorithms has already been proven in Allassonnière and Chevallier (2019) but not for a state-dependent temperature, nor for the MCMC-SAEM algorithm. A generalization of this work would be needed to obtain the theoretical convergence of our algorithm.

The process is summarized on algorithm 8.

**Algorithm 8:** MCMC-SAEM algorithm

**Data:** $(y_{i,j})$, $(t_{i,j})$, total number of iterations $K$, $s_0 = 0$ and $(\Delta_k)_{k\in\mathbb{N}}$ a decreasing positive step size sequence

**for** $1 \leq k \leq K$ **do**

Sample $(z_{pop}, z)$ using a single step of a Symmetric Random-Walk Metropolis Hastings within Gibbs sampler targeting the posterior distribution $q(z_{pop}, z|y, \theta_k)$.

Compute $T_k$ using Eq. 3.12 and sample $c$ using a single step of a Symmetric Random-Walk Metropolis Hastings within Gibbs sampler targeting the posterior distribution $\frac{1}{T_k}q(c|y, \theta_k)$.

Compute the stochastic approximation $s_k = s_{k-1} + \Delta_{k-1}(S(z, z_{pop}, y) - s_{k-1})$ where $S$ are the sufficient statistics.

Update the parameters $\theta_k$ to maximize the posterior likelihood $q(\theta|y)$: $\theta_k = \hat{\theta}(s_k)$.

### 3.3.3 Initialization and influence of the hyperparameters

Now that we have presented the algorithm estimating $\theta$, we interest ourselves in its initialization and in the influence of the choice of the hyperparameters.

Concerning the initialization, all the representative trajectories of the different clusters are chosen equally by building a constant trajectory equal to the first observation of the first subject at all times. Similarly, we initialize the individual parameters such that there is no initial deformation towards the subjects. Hence, at first, all individual trajectories are equals.

The different hyperparameters defining the priors influence the update of $\theta$ at each iteration. Indeed, all those updates can in fact be seen as barycenters between a quantity defined by the sufficient statistics and another depending on the prior. For instance, $\bar{z}^r_{pop}$ is the barycenter between a sufficient statistic and $\bar{\bar{z}}^r_{pop}$ with respective weight $\frac{\bar{v}_{pop}}{\bar{v}_{pop}+v_{pop}}$ and $\frac{v_{pop}}{\bar{v}_{pop}+v_{pop}}$. Hence, we can choose the prior to influence the final value of $\bar{z}^r_{pop}$ and also choose the weight given to this a priori. Similar remarks can be done with all parameters.

Finally, we must also choose the kernel used to compute the deformations. Here, we take a Gaussian kernel: $K_g(x, y) = \exp\left(\frac{\|x-y\|_2^2}{\sigma_g^2}\right)$. We choose the kernel width $\sigma_g$ in the range of the distance between the control points such that the whole shape can be deformed smoothly.

## 3.4 Results

### 3.4.1 2D simulated data

#### Creation of the dataset

We first test our algorithm on simulated data mimicking the shape of a dancing man. We create 100 subjects by deforming a branching piecewise-geodesic representative curve with two

components. More precisely, we begin by creating the two branching representative trajectories by drawing three sets of random momenta that we apply on 16 control points equally spaced. We first apply one set of momenta on a fixed shape to obtain the first common component and then we apply the two other sets of momenta on the same fixed shape to obtain the two distinct components forking at the rupture time, set as 70. We then create our 100 individuals by sampling random accelerations, time shifts and space shifts from a gaussian distribution as well as random number of observation times before and after the rupture time. Those observation times are sampled using an exponential distribution. Finally, we add a gaussian noise of variance $0.02$ to each subject, use the varifold distance and choose a kernel width equals to the distance between two adjacents control points.

### Estimation of the parameters

We apply our algorithm to find the representative curves and the spatiotemporal deformations towards the data sequence of each subject, asking for two branching clusters. Results in Fig. 3.6 show that there is only little differences between the true and estimated representative trajectories (left), and no noticeable differences between the true and reconstructed observations. To quantify the reconstruction error, we compute the varifold norm of the errors for all subjects along the iterations on Fig. 3.7 (left).

$97\%$ of the subjects are classified in their right cluster. As for the others subjects, in most cases, no measurement is done after the rupture time or the second acceleration coefficient is so small that the shape practically does not vary after the rupture time, which explains why the algorithm cannot find the right cluster. We also show the necessity of using tempered distributions by plotting the error of classification with and without temperature on Fig. 3.7 (right). The oscillations we see on those figures are due to the oscillating evolution of the temperature. We can see that the classification and hence the final reconstructions are better with tempered distributions.

Finally, we launch the algorithm on the same data set 10 times to compute the errors on the estimation of the different parameters. On the table 3.1, we display the relative errors of the individual parameters. In particular, we do not show the error on the time shifts $\tau_i$ but on the individual rupture times $t_{R,i,0}$ (obtained from $\tau_i$) since it is this value which will be of interest for the clinicians. All those errors are below $10\%$, with particular good estimation for the individual rupture times. The high standard deviation observed is in fact due to the badly classified subjects. Indeed, for those subjects, the individual parameters often take absurd values: practically null accelerations, large rupture times, etc..
On the table 3.2, we present the errors of reconstruction for the varifold norm. We can remark that both the subjects and the templates are very well reconstructed. The error on the template is a bit higher due to the repercussion of the small errors in the temporal reparametrization. Indeed, the small errors in accelerations can cause the time lines between the real template and the estimated one to differ causing small errors when comparing them at the same time point.

We also present the errors on our parameters table 3.3. Here, we can remark the very poor estimation of $\Sigma$. Once again, this is due to the presence of badly classified subjects having absurd individual parameters. Those outliers then induce a very high variance in the estimated individual parameters. However, if we try to compute the estimated $\Sigma$ taking into account only the subjects in the correct cluster, we then find more correct results: an error of $8.12\%$ with a

standard deviation of $3.97$. Hence, it seems impossible to have a correct estimation of $\Sigma$ here.

| $\xi_{i,0}$ | $\xi_{i,1}$ | $t_{R,i,0}$ |
|---|---|---|
| $5.89\%_{\pm 7.01}$ | $8.60\%_{\pm 10.7}$ | $0.76\%_{\pm 1.61}$ |

**Table 3.1** Mean and standard deviation of the relative errors for the temporal parameters.

| Subjects | Templates |
|---|---|
| $1.23\%_{\pm 1.96}$ | $5.56\%_{\pm 2.60}$ |

**Table 3.2** Mean and standard deviation of the errors of reconstruction for the subjects and templates.

| $\Sigma$ | $\sigma$ | $p$ |
|---|---|---|
| $160\%_{\pm 223}$ | $7.19\%_{\pm 4.01}$ | $2\%$ |

**Table 3.3** Mean and standard deviation of the errors on the parameters $\theta$.

#### Prediction of new data

Here, we test the ability of our model to predict new data by using cross validation. We create 100 new subjects deformed from the same representative curve as before. We then ask our algorithm to classify and reconstruct the trajectories while fixing the parameters $\theta$ and the representative curve by those learned previously. This time, $91\%$ of the subjects are well classified and the error of reconstruction is only $0.84\%$ with a standard deviation of $1.93$. Hence, our model can process new data without a problem, proving that we have no problem of overfitting or selection bias.

#### Comparison of the clustering with a baseline

We now want to test the performance of the clustering of our model against a baseline. To do so, for each of the subjects, we compute the trajectory minimizing the distance with the observations using a geodesic regression. We obtain, for each subject, a set of momenta defining its trajectory. We then use the kmeans algorithm on the set of all momenta to classify the subjects. This algorithm will not create representative trajectories nor compute the variability of the population but will only classify the subjects without any time reparametrization.

In this easy example where only the global movement of the shapes is important in the clustering, the baseline gives us a perfect classification of the subjects. However, it is easy to create cases where our algorithm will outperform the baseline. Indeed the baseline only takes into account space deformations. Hence, it is unable to distinguish two different objects deformed the same way. For instance, a geodesic regression will give us the same set of momenta for

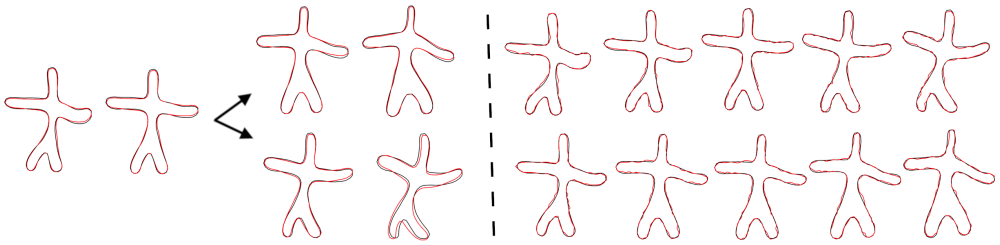


**Figure 3.6** In red, the exact simulated data, in black, the results given by our algorithm. On the left, the representative curves that split up at a certain rupture time. On the right side, two subjects given with their reconstructions.

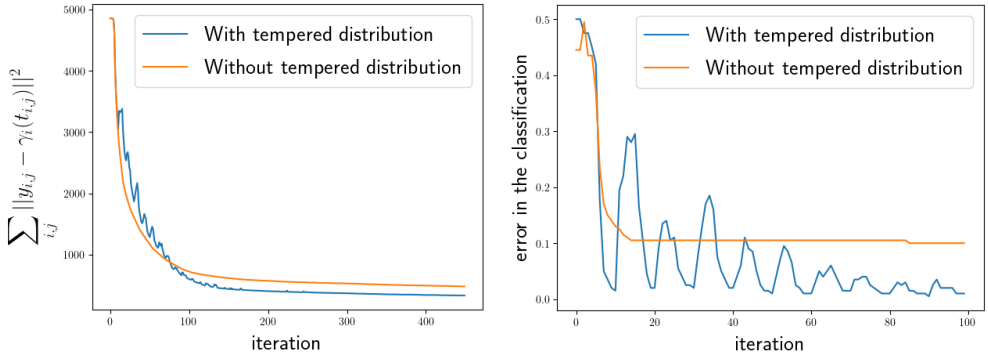



**Figure 3.7** Left: evolution of the varifold distances between the subjects and their reconstructions. Right: percentage of error in the classification along the first 100 iterations. With tempered distribution, the oscillating temperature coerces a lot of subjects to change classes. After 500 iterations, the error is $31.3\%$ smaller.

squares and spheres following the same movement. Hence, the baseline will not be able to distinguish two different clusters. In contrast, our algorithm also takes into account the mean shape of each cluster and so is able to separate two such clusters.
Moreover, no time reparametrization is taken into account by the baseline. To highlight this fact, we create a new dataset of "dancing men" with two clusters, each containing 100 subjects. We obtain those subjects from the same representative curve but, for one cluster, the mean acceleration of the subjects $e^{\xi}$ is 1.3 while the other has a mean acceleration of 0.7. This time, the baseline is unable to distinguish the two clusters as the momenta obtained by geodesic regression for the different trajectories are all collinear. All the subjects but 6 are placed in the same cluster and so only $51\%$ of the subjects well classed. On the other hand, our algorithm is more successful in this clustering task: subjects are indeed classified according to their speed of progression: $84\%$ of the subjects are classified as expected. As for those badly classified, their acceleration is close to 1.

Finally, when the only distinction between clusters is based on their space deformation, the baseline seems as precise as our algorithm. However, it is not able to distinguish differences in time and is more limited than our model. Those observations will be confirmed in the next examples.

**Test of an hypothesis on the model**

We want now to test hypothesis about the heterogeneity of the population. We run our algorithm on the dataset created section 3.4.1, supposing first that the two representative trajectories are different. We then run it again supposing that their first component is the same and that they fork at the rupture time. To select the model, we first compute the log-likelihood ratio test. However, in this case, this test is not enough to determine which model to choose. Indeed, with two independent representative curves, the algorithm can reconstruct the subjects as precisely as with branching representative curves. Hence, the difference between the likelihoods of the two models is too small to conclude and the test unstable between runs. To overcome this problem, we choose to compute the Bayesian Information Criterion (BIC):

$$\text{BIC} = ln(n)m - 2ln(q(y, z, \theta))$$

where $m$ is the total number of parameters involved in the model and $n$ the number of subjects. This criterion takes into account the complexity of the model by adding a penalty proportional to the number of parameters involved. Hence, we will penalize the model with two independent trajectories (as it involves more parameters) even if the reconstruction is similar. This time, there is a difference of $2.98\%$ between the two BIC criterions, leading us to choose, as expected, the model with branching representative curves.

### 3.4.2 1D RECIST scores

We test here the algorithm on a real 1D dataset. We consider a database of patients suffering from the meta-
static kidney cancer and taking antiangiogenic drugs. They come on a regular basis at the hospital to check the tumor evolution. Two behaviours are expected in the population: for all patients, the tumor first regresses. But then, for some, it stabilizes while for others the tumor size increases again forcing to change the treatment. The RECIST score is a feature that measures the tumor size and is used in the majority of clinical trials evaluating cancer treatments for objective response in solid tumors. Our dataset consists in the evaluation of the RECIST score for 176 patients with an average of 7 visits per subject and an average duration of 90 days between consecutive visits.

In this 1D case, shapes are just curves on $\mathbb{R}$ and we work with a logistic metric. The parallel transport is just a translation of the geodesic. That is why we rather considerate another space reparametrization, as done in Allassonniere et al. (2017): for all classes $i$ and all components $l$, we set:

$$\phi_{i,l}(x) = \gamma^{cl(i)}(t_R^{cl(i)}) + e^{\rho_i^l}\left(x - \gamma^{cl(i)}(t_R^{cl(i)})\right) + \delta_i^l\,.$$

$\rho_i^l$ is a dilatation factor and $\delta_i^l$ is a translation factor. As with the time reparametrization, all the translation factors but the first one are fixed by continuity conditions and we note $\delta_i^0 = \delta_i$. Finally, our individual curve is defined by deforming spatially each component of $\gamma^{cl(i)}$ by $\phi_{i,l}$ and temporally by the same $\psi_{i,l}$ as previously.

With only two components, the piecewise geodesics for the logistic metric can be parame-

terized, for any class $r$, by:

$$\begin{cases} \gamma_1^r(t) = \dfrac{\gamma_{init}^r + \gamma_{escap}^r e^{a_r t + b_r}}{1 + e^{a_r t + b_r}} \\ \gamma_2^r(t) = \dfrac{\gamma_{fin}^r + \gamma_{escap}^r e^{-(c_r t + d_r)}}{1 + e^{-(c_r t + d_r)}} \\ \gamma^r(t) = \gamma_1^r(t)\mathbb{1}_{]-\infty, t_R^r]} + \gamma_2^r(t)\mathbb{1}_{]t_R^r, +\infty[}\,, \end{cases} \tag{3.13}$$

with $\gamma_{init}^r, \gamma_{escap}^r, \gamma_{fin}^r \in \mathbb{R}$. We fix $a_r$, $b_r$, $c_r$ and $d_r$ by asking the geodesics $\gamma_{0,r}^1$ and $\gamma_{0,r}^2$ to be $\nu$-near their geodesics at an initial time $t_0^r$, at the rupture time $t_R^r$ and at a final time $t_1^r$ (see Allassonniere et al. (2017) for more details). Hence, rather than sampling momenta and control points, we will sample $z_{pop}^r = (\gamma_{init}^r, \gamma_{escap}^r, \gamma_{fin}^r, t_0^r, t_1^r)$. This whole process is summarized Fig. 3.8.

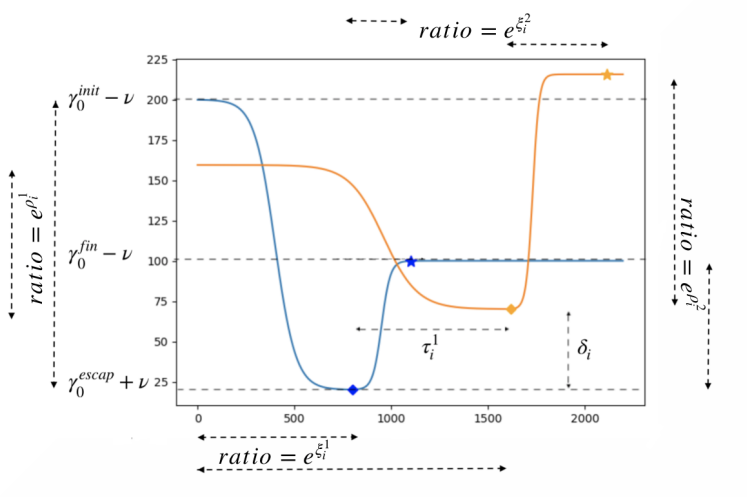



**Figure 3.8** Model description. In blue, the template with the different parameters defining it and in orange one subject obtained by deforming it. Here, $t_0 = 0$, the rupture points are represented by diamonds and the final times $t_1$ by stars.

First, we launch our algorithm looking for two different representative curves. The result is displayed on the first line of figure 3.9. Our algorithm is indeed able to explain the variability of the population. However, it seems that our algorithm favours size over response dynamic as a clustering feature: small initial tumors (blue curve, $28\%$ of the patients) are separated from big initial tumors (orange curve, $72\%$ of the patients). For example, the orange reconstructed trajectory (top right plot) is classified with the blue template (top left plot) even if the treatment stays effective.

To overcome this trivial differentiation based on the tumor initial size, we ask the two templates to be the same until the rupture time using a branching process. This time, on the second line of figure 3.9, we really see two different behaviours: for one of the template, the RECIST score increases a lot more (blue curve, $37\%$ of the patients) than for the other (orange curve, $63\%$ of the patients). As for the clustering, we see indeed that the subjects whose RECIST score do not increase after the rupture time are pooled together (green, red, orange and blue curves).

Hence, we are able to separate the patients whose tumor becomes resistant to the treatment from the others. It can also be remarked that we have fewer time points for patients whose tumor becomes resistant because the clinicians change the treatment when this resistance is remarked and so the record of score for this patient stops.

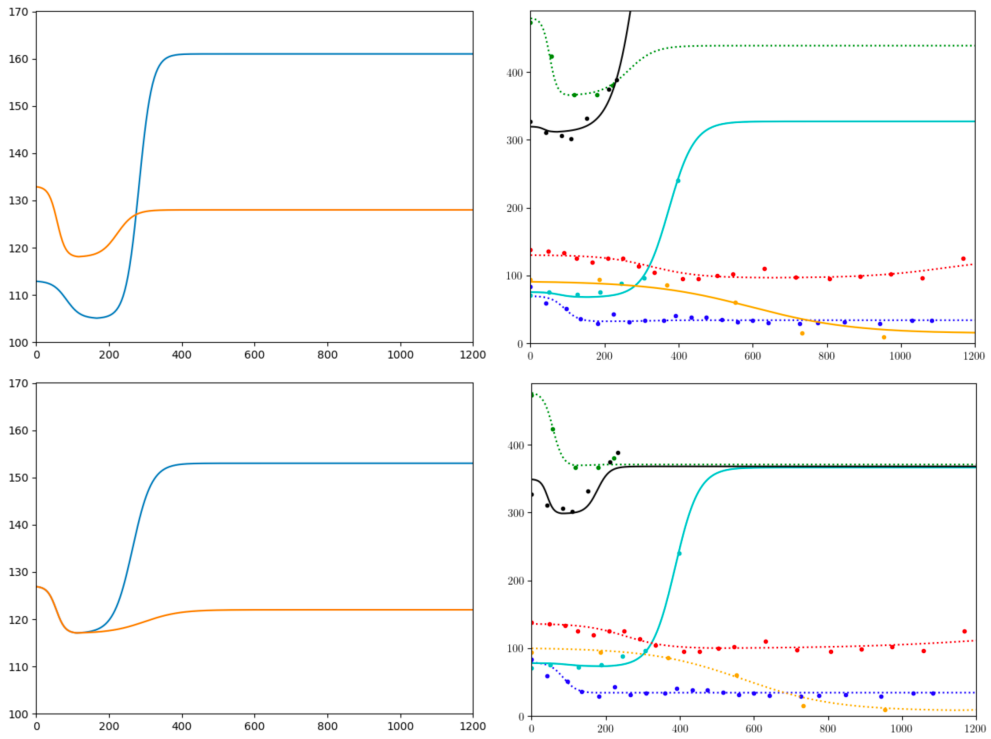


**Figure 3.9** At the top, the results given with two different templates, at the bottom, with two templates whose first component is the same. To the left, our templates. To the right, 6 subjects and their reconstructed trajectories. In dotted lines, subjects in the cluster of the orange template. In plain lines subjects in the cluster of the blue template.

### 3.4.3 3D faces

We now obtain shapes of subjects expressing different facial expressions from the Birmingham University 3D dynamic facial expression database (Yin et al.). This real database contains short videos from 101 subjects expressing happiness or surprise. We uniformly extract 8 frames, from the first to the 36-th one, which correspond to a subsampling of the first 1.4 seconds of each video. We do not work directly with the texture video, but with a set of $75$ semi-automatically extracted landmarks, which were readily available along with this data set. Every set of 3D landmarks is registered to a reference one by Procrustes alignment.

We apply our algorithm, once again with the varifold distance, to find two clusters, with only one component geodesic for each template. As we can see Fig. 3.10 and 3.11, the faces are well reconstructed and we can recognize the two expressions of surprise and happiness on the two templates. In particular, for the surprise cluster, the mouth is more widely open, while the

eyes are wide open and the eyebrows higher.

Hence, we can ask ourselves if the algorithm has really detected those two expressions or if another characteristic has been detected to distinguish two sub populations. In fact, $68.5\%$ of the subjects are classified as expected (i.e. surprised subjects in the cluster with the template looking surprised and happy subjects in the one looking happy). There are different explications about the subjects classified differently. First, we can remark that some of them have a non neutral expression at the first image, for example smiling at the beginning while they should express surprise. For others, it is just really difficult (even for a human) to determine if they express happiness or surprise (see Fig. 3.12). Finally, we can also remark figure 3.10 that the left eyebrow is quite different from one template to another. And indeed, we find that same difference in several subjects misclassified. However, even if the clustering can be surprising, the algorithm fulfilled his role: we have been able to highlight two different dynamics in the population that can be explained by differences in the subjects considered.

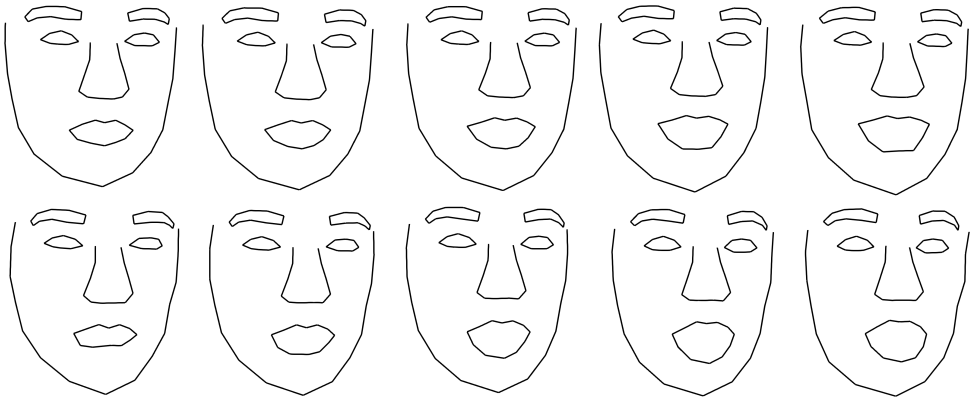


**Figure 3.10** Results of the algorithm when applied to a dataset of surprised or happy visages. At the top, the evolution of the template of the happiness cluster, at the bottom, the evolution of the template of the surprised cluster, one component for each template.

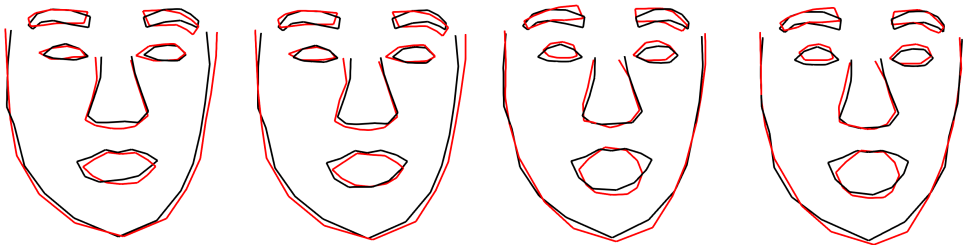


**Figure 3.11** Reconstitution of a subject expressing surprise. In red, the exact data, in black the reconstitution.

Concerning the baseline, we have a better classification in this case: $88\%$ of the subjects are classified as expected. This better classification can be explained by the fact that the movement of the lips and eyebrows is the principal feature separating the two clusters. By not taking into

**Figure 3.12** Evolution of subject that has been asked to express happiness but seems to express surprise. It is indeed classed in the template looking surprised by our algorithm.

account the initial shape of the subjects but only the deformation, the baseline is able to obtain a better classification result. In this case, if we are interested in separating the happy subjects from the surprised ones, it would thus be preferable to first compute the clusters using the baseline and only after to run our algorithm in a supervised way with the fixed clusters obtained previously to obtain the representative trajectories and the variability in each cluster.

### 3.4.4 Hippocampi dataset

We finally test the algorithm on 100 subjects obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). 50 of those subjects are control patients (CN) and 50 are Mild Cognitive Impairment subjects eventually diagnosed with Alzheimer's disease (MCIc). Meshes of the right hippocampus is segmented from the rigidly registered MRI.

We first run our algorithm with a forking model: we look for two clusters that fork at a certain rupture time. As there is no reason for the control subjects to have two different dynamics, we also ask one of the cluster (i.e. one of the evolution scenario) to follow the same geodesic before and after the rupture time. Finally, we choose to use the varifold distance. Our algorithm splits the patients in two clusters, one of them presenting a quicker and different pattern of atrophy (Fig. 3.15 and left side of Fig. 3.13 where the hippocampi volume is plotted along time). Moreover, $72\%$ of the subjects are classified as expected: the CN in the cluster with a single dynamic showing a slower atrophy and the MCIc in the cluster with a faster atrophy after the rupture time.

We have also studied the relation between our rupture time and the age of diagnosis. The individual rupture times are strongly correlated to the diagnostic age, indicating that we have been able to detect a change of behaviour correlated with the date of diagnosis (Fig. 3.14).

We run again the algorithm, this time looking for two clusters with separate trajectories, one of them with only one dynamic. The results are presented Fig. 3.16 and on the right side of Fig. 3.13 for the hippocampi volumes evolution. It is interesting to remark that the cluster with only one dynamic also presents a slower atrophy, as expected with a normal ageing. We can also detect different patterns of atrophy before and after the rupture time for the cluster with two dynamics. This time, $70\%$ of the subjects are classified as expected: CN in the cluster with one dynamic and MCIc in the cluster with two dynamics and a quicker rate of atrophy.

As we are given two possible evolution scenarii, it is natural to try to quantify the goodness of

fit of each of them, allowing for a choice of a better explanation of the disease. As for synthetic data, we use the Bayesian Information Criterion. We find a difference of $2.92\%$ between the two BIC values leading to choose the branching model. Hence, this suggests that the MCI subjects first follow a normal aging scenario but deviate from it at the rupture time. It must however be remarked that our model is quite complex with a lot of high dimensional variables, making model selection quite difficult.

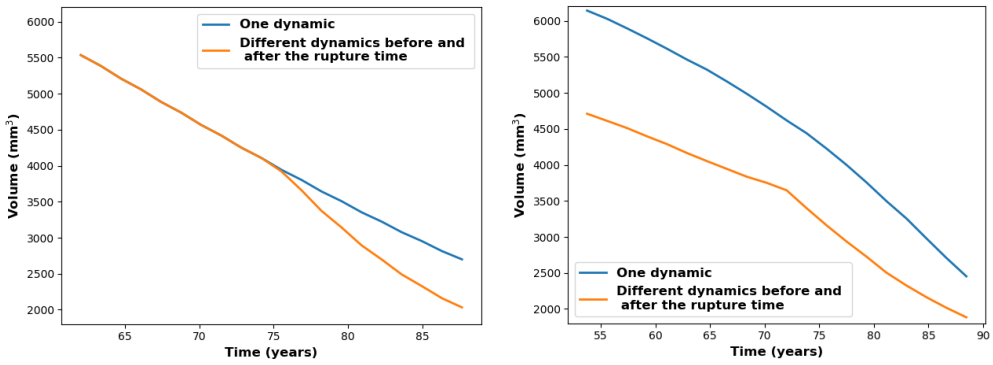



**Figure 3.13** Left: volume evolution for two branching clusters. Right: volume evolution for two clusters with separate trajectories.

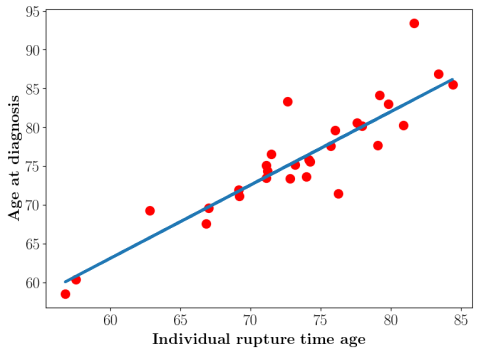



**Figure 3.14** Comparison of the age at diagnosis with the individual rupture time for the MCIc patients in the case of the branching model, $R^2 = 0.91$

Once again, we compare those results with the baseline. However, in this case, the difference between the two clusters is largely coded by the speed of atrophy and not the global dynamic. Hence, it is not surprising to note that practically all the subjects are grouped in the same cluster by the baseline and so, only $52\%$ of the subjects are well classified. Thus, in this example, our algorithm has to be used to cluster the subjects.

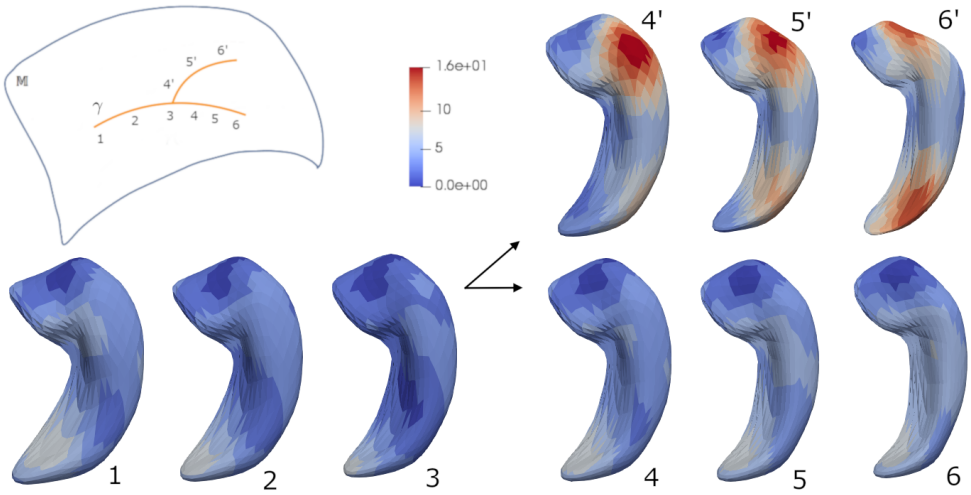



**Figure 3.15** Representative shape evolution at the ages $63.4$y, $68.8$y, $74.2$y (i.e. rupture time), $75.5$y, $80.9$y and $86.3$y. Bottom shapes: cluster with one dynamic. Top shapes : cluster with change of dynamic after rupture time. The color map gives the norm of the velocity field $\|v_t\|$ on the meshes.

## 3.5 Conclusion

We proposed a mixture model for longitudinal shape data sets where representative trajectories take the form of piecewise geodesic curves. Our model can be applied in a wide variety of situations to test whether sub-populations are independant from each other or fork or merge at different time-points. We showed on simulated examples that our tempered optimization scheme is key to achieve convergence of such a mixed effect model combining discrete variables with continuous variables of high dimension. It has also been noticed that taking only into account the individual trajectories is not always enough to obtain a meaningful clustering of the population. We have shown the versatility of our model by applying it to a lot of different cases: trajectories with one or several dynamics, branching or not after a rupture time, with one part of the population still following the same dynamic or not after the rupture time. Its application on 1D data allowed us to present results of the same model in another setting while the application with 3D faces showed that we can highlight different meaningful dynamics in a same population. Finally, the hippocampi data set allowed us to investigate the relationship between normal and pathological ageing.

Different questions still have to be answered. In particular, our scheme of temperature depends of the current state of the algorithm and a proof of convergence should be provided in this situation. Moreover, specific model selection criterion should be devised in this complex longitudinal setting. Those criterion should in particular help us to detect the optimal number of clusters and rupture times. Finally, one may ask what is the consequence to consider the population parameters $z_{pop}$ as random variables. This particular question is the subject of the chapter 6 of the manuscript.

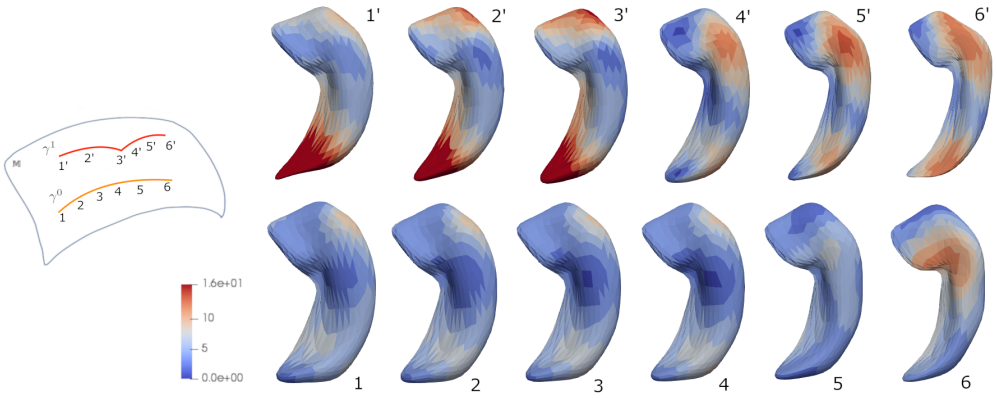



**Figure 3.16** Model with two separate clusters, one of them following only one dynamic. Representative shape evolution at the ages $64.7$y, $68.3$y, $74.3$y (i.e. rupture time for the cluster with two dynamics), $77.5$y, $80.9$y and $86.3$y. Top shapes : cluster with change of dynamic after rupture time. Bottom shapes: cluster with one dynamic. The color map gives the norm of the velocity field $\|v_t\|$ on the meshes.

# CHAPTER 4

# Detection of anomalies using the LDDMM framework

In the previous chapter, we have created a new longitudinal model that allows to consider subjects with change of dynamics at certain time points. We have in particular applied it to RECIST scores in the case of chemotherapy monitoring. The intuitive next application would be to consider images of organs with tumors.
However, we are not able to apply this model directly. Indeed, in that case, each observation consists of the organ with a different number of lesions on it. But, each reconstruction, obtained as a diffeomorphic deformation of a template, will have the same number of lesions as this template. We hence need to create a new model tackling this issue.
This is the subject of this chapter where we propose to model observations as diffeomorphic deformations of a template (to be estimated or not) and structured residuals. The residuals are composed of a sparse matrix coding for the lesions and an independent additional noise. The goal is to be able to highlight the tumors in these sparse matrices. We will see that this new approach has two advantages. First, it localizes the tumors without any labeled training sample (avoiding for clinicians pre-processing). It also reduces the error of reconstruction while being efficient in the detection of anomalies. This enables to better evaluate the size and shape of the organ as a whole. Last, this method is not organ dependent, allowing for multiple use in oncology.

## Contents

## 4.1 Introduction

In this chapter, we are interested in the detection of anomalies taking advantage of the Large Diffeomorphic Deformation Metric Mapping (LDDMM) framework. The motivation behind the introduction of this new model comes from a medical issue faced by interventional radiologist when dealing with a data set of patients with colorectal cancer. In most cases, hepatic metastases of colorectal cancer are not resectable at the time of diagnosis. Intra-arterial chemotherapy is one of the techniques developed recently and aims at providing the patients with a new therapy: patients can receive high doses of chemotherapy administered locally, with fewer systemic doses, thus reducing overall toxicity. Several of them will benefit from secondary curative options such as surgery or classical intravenous chemotherapy. A 2020 clinical trial considered using these techniques as a first-line treatment for colorectal cancer liver metastases (Pernot et al., 2020). However, this treatment is not efficient for all patients. This led the interventional radiologist Medical Doctors to ask for the possibility to predict the progression-free survival at 9 months i.e. the patients still alive and whose disease has not worsened after 9 months of treatment. This could lead to better patient selection, and would prevent patients from receiving inefficient treatment in a setting where many other options exist. As a first step, localizing the lesions and quantifying the global liver shape is required. This encouraged us to propose the following contribution. This model has been designed to address this problem but can also be generalized to other organs. Therefore, two applications are presented below.

We are given a data set of scanners (i.e. of 3D images) of 57 patients, each of them being observed between 2 and 11 times. An example of such a scanner is given Figure 4.1. Tumors can be seen as dark spots on the liver.
We are also considering the BraTS 2018 data set of neuro-oncology where we are given 50 MRIs of brains with glioma (Bakas et al., 2017, 2018; Menze et al., 2014).
As a first step, we will just consider the first observation of each patient and interest ourselves in the creation of a cross-sectional atlas.

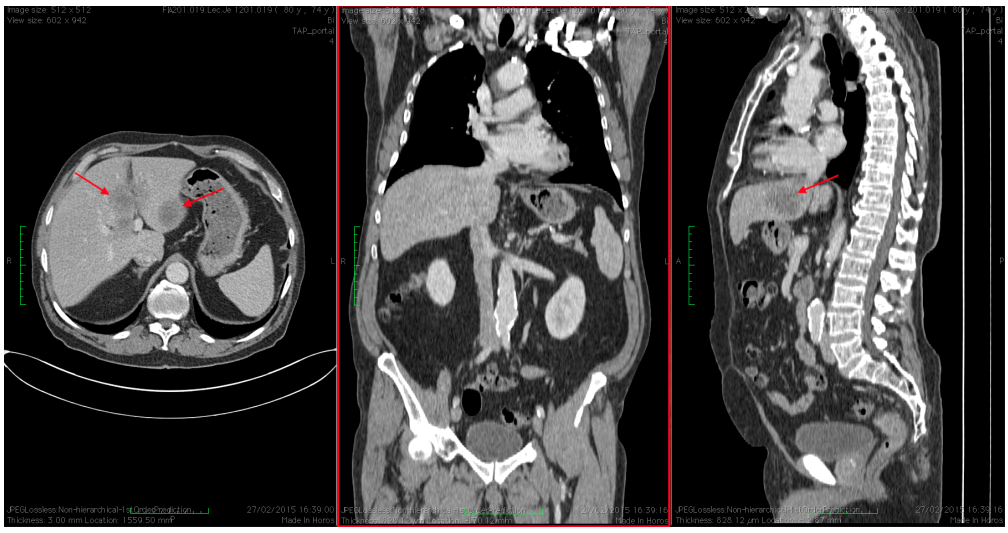


**Figure 4.1** Example of the scanner of a patient. The red arrows indicate tumors.

The presence of tumors prevents us from directly applying the methods developed previously, which use the LDDMM framework to account for deformations. Indeed, each subject can have a different number of lesions which leads to a different topology. If we directly apply the LDDMM framework, the reconstruction of each subject is obtained as a diffeomorphic deformation of the template. If this template contains $k$ dark spots, each deformation will also have $k$ dark spots, even if the targeted subject does not have this number of lesions. It may even force the model to try to make some dark spots appear or disappear by using strong deformations in areas where the liver shape should not change. Therefore, a model driven by a template with lesion may not be the right one. One would rather think the template as a typical organ and the lesions as additional elements not concerned by the global shape of each subject. Hence, we need to transform the model.

We will in fact use this diffeomorphic constrain to automatically detect anomalies on a set of images. More generally, we suppose that we have at our disposal a data set of subjects without anomaly and another with anomalies. The goal is to immediately detect the presence, or not, of those anomalies.

Different methods already exist for the segmentation of anomalies such as tumors. Most of them use deep learning and algorithms derived from U-net (Ronneberger et al., 2015) (see among many other works Bakas et al. (2018); Dong et al. (2017); Seo et al. (2019)). They have however the disadvantage to require hundreds of images first annotated by oncologists. As the methods depend on clinicians and large data set, they cannot be easily generalized to any organ. Moreover, manual segmentation is expensive and time-consuming to obtain, greatly limiting clinical application and the scale of available labelled data. To overcome this lack of labeled data, another possibility is to study the residuals of an auto-encoder network trained on control subjects (Baur et al., 2018; You et al., 2019; Pawlowski et al., 2018; Yu et al., 2019). This method is quite similar to the one we will develop but still necessitates lots of images to learn the parameters of the auto-encoder.

The method we propose has the advantage not to require any annotation nor large data sets and can straightforwardly be generalized to any organ. We suppose that we have a data set of control patients from which we are able to create a control template $\bar{y}$. In practice, obtaining a data set of control patients is often easier, for example by considering images from patients suffering from any other pathology with no impact on the organ considered. Moreover, estimating an atlas using the LDDMM framework does not require many training data. Indeed, we are not estimating parameters of a blind neural network but rather parameters of a statistical model mimicking the data generation. This control template will characterize a control population and hence, will have no anomaly. With this template fixed, anomalies on a new subject are then defined as what cannot be obtained as a diffeomorphic deformation of this control template.

To extract these anomalies, we model the residuals (i.e. the difference between the deformed template and the observation) as a sparse matrix in addition to an independent noise. What cannot be reconstructed as a diffeomorphic deformation of the template is hence put in this matrix and classified as an anomaly. The goal is to obtain these anomalies in the matrix and separate them from the noise. The idea to study the residuals is further motivated by Durrleman et al. (2011b) where the authors showed that the residuals still contain information on the variability of the population. Another advantage of our model is that it retrieves the deformations from the control template towards the patients. This deformation can be interesting in the prediction of the outcome of a treatment where the environment around the anomalies sometimes

plays a decisive role.

One could imagine to first estimate the deformation from the template towards the observation and only at the end of the estimation, estimate the anomalies from the residuals. However, we will show that estimating both deformations and anomalies at the same time improves the results by reducing the error of reconstruction. Indeed, taking into account the possibility of having additional elements in the organ enables to see the organ globally and therefore to have a more accurate deformation.

Although the method seems to rely on a template of control patients, it can actually be estimated as well. We will propose a way to derive a template without anomaly from the sick patients only using a single observation of a control patient, called hypertemplate.

In the following, we will first present the mathematical framework, the statistical model and its estimation algorithm. Then, we apply the model on a simulated data set and show that we obtain better reconstructions of our objects and an accurate localization of the lesions. We then show its versatility by using it on two different real data sets. First, we apply it on a data set of brains with lesions and then on a data set of livers.
At the time of writing of this manuscript, this work is still in progress.

## 4.2 Detection of anomalies using residuals

### 4.2.1 Presentation of the model

The model we propose aims at highlighting the anomalies of patients with respect to a control template. As we are modeling anomalies, we make the assumption that these features are sparse in the volume of the organ. Apart from these anomalies, the rest of the organ has to be similar to a control one. Therefore, we propose the following model.

We write $(y_i)_{1\leq i\leq n}$ the $n$ subjects and $\bar{y}\in\mathbb{R}^d$ the template of control subjects obtained using the methods of Section 2.4. We suppose that our images all have the same size $\prod_{j=1}^{d} n_i$ where $d=2$ or $3$ for $2d$ or $3d$ images.

We assume that each observation can be written as:

$$y_i = \mathcal{E}xp_{0,1}(v_i)(\bar{y}) + A_i + \varepsilon_i\,,$$

where the Riemannian exponential is defined Section 2.3.1. This equation means that we obtain the subject $y_i$ as the diffeomorphic deformation of the template $\bar{y}$: $\mathcal{E}xp_{0,1}(v_i)(\bar{y})$ to which we add a matrix $A_i$ containing the anomalies and a noise $\varepsilon_i$. As explained in Section 2.4.4, $v_i$ is a velocity field that we obtain as the interpolation of momenta $\alpha_i\in(\mathbb{R}^d)^{n_{cp}}$ at $n_{cp}\in\mathbb{N}$ control points $c_i\in(\mathbb{R}^d)^{n_{cp}}$ using a Gaussian kernel as follows: $\forall x\in\mathbb{R}^d$:

$$v_i(x) = \sum_{j=1}^{n_{cp}} K_V(c_{i,j},x)\alpha_{i,j}\,. \tag{4.1}$$

In order to enforce the expected sparsity of the anomalies, we add a $\mathcal{L}^1$ regularization on $A_i$. This is also a way to prevent it from including the noise, modeled as following a centered

normal distribution.

Given this model, our goal is to estimate jointly the deformations from the given template and the anomaly matrices for each subject. Here, as the control template has already been estimated, we can process each subject separately and we want to minimize the following functions:

$$J_i(c_i, \alpha_i, A_i) = \frac{1}{2\sigma^2}||y_i - \mathcal{E}xp_{0,1}(v_i)(\bar{y}) - A_i||_2^2 + \lambda||A_i||_1 + \frac{1}{2}||v_i||_V^2 \,, \tag{4.2}$$

where $||.||_V$ is defined Section 2.4.4 and $v_i$ is obtained using equation (4.1). The first term of $J_i$ measures the distance between the observation and the reconstruction while the other terms measure the sparsity of $A_i$ and the regularity of the diffeomorphic deformation.

To minimize $J_i$, as $||.||_1$ is not differentiable, we implement a proximal gradient descent algorithm, using the Pytorch package to automatically compute the gradients of the differentiable part of $J_i$.

### 4.2.2 Computation of the template using a hypertemplate

In some cases, it can be difficult to obtain a data set of control subjects, necessary to create the control template $\bar{y}$. It is for instance the case for brains where we barely get a MRI scan from a control patient. In that case, it is possible to create the template $\bar{y}$ along the estimation of the anomaly matrices using only the image of one control patient $y_0$. To do so, we consider $\bar{y}$ to be the diffeomorphic deformation of a control hypertemplate $y_0$:

$$\bar{y} = \mathcal{E}xp_{0,1}(v_0)(y_0) \,, \tag{4.3}$$

where $v_0$ is obtained as the interpolation of momenta $\alpha_0$ at control points $c_0$, as explained above. As $y_0$ is a control subject, it has no anomaly, and its diffeomorphic deformation $\bar{y}$ will have no anomaly either.
Hence, in that case, we will not only estimate the velocities $(v_i)_{1\leq i\leq n}$ and sparse matrices $(A_i)_{1\leq i\leq n}$ but also the velocity $v_0$ by minimizing:

$$\begin{aligned}\tilde{J}(c_0, (\alpha_i)_{0\leq 1\leq n}, (A_i)_{1\leq i\leq n}) = &\sum_{i=1}^{n}\left(\frac{1}{2\sigma^2}||y_i - \mathcal{E}xp_{0,1}(v_i)(\bar{y}) - A_i||_2^2 + \lambda||A_i||_1 + \frac{1}{2}||v_i||_V^2\right) \\ &+ \frac{1}{2}||v_0||_V^2 \,,\end{aligned} \tag{4.4}$$

where $\bar{y}$ is obtained using the velocity $v_0$ and $v_i$ is obtained as the interpolation of the momenta $\alpha_i$ at the control points $c_0$.
Note that, as the template is a reference image for the whole population, the energy to minimize is not separable and involves the contributions of all subjects.
Once again, this optimization is done by proximal gradient descent.

**Remark 4.2.1.** *Here, we use the same control points $c_0$ for all velocity fields $(v_i)_{0\leq i\leq n}$ in order to reduce the computation time.*

**Remark 4.2.2.** *In practice, we sometimes have a computational problem to estimate the parameters of those models. Indeed, the gradient descent tends to stay blocked in local minima. By including the errors of reconstruction in the anomaly matrix, the algorithm often chooses not*

*to improve the reconstruction and stays blocked in a local minimum. To solve this problem, we prevent the anomaly matrix to take any value outside of the reconstruction of the object for the first 100 iterations. This allows to improve the diffeomorphic reconstruction and to reach a more relevant area of interest of the energy landscape.*

### 4.2.3 Comparison to other models

In the following, we will compare this model to the usual cross-sectional atlas, estimating the template $\bar{y}$ and the deformations towards the subjects without the use of anomaly matrices nor hypertemplate. This writes as minimizing the following energy:

$$J_0(\bar{y}, c_0, (\alpha_i)_{1\leq i\leq n}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n} ||y_i - \mathcal{E}xp_{0,1}(v_i)(\bar{y})||_2^2 + \frac{1}{2}\sum_{i=1}^{n} ||v_i||_V^2 \,. \tag{4.5}$$

We will also compare it to the cross-sectional atlas when the template is obtained via a control hypertemplate. As above, $v_0$ will be obtained as the interpolation of momenta $\alpha_0$ at control points $c_0$. The functional to minimize is now:

$$J_1(c_0, (\alpha_i)_{0\leq i\leq n}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n} ||y_i - \mathcal{E}xp_{0,1}(v_i)(\bar{y})||_2^2 + \frac{1}{2}\sum_{i=0}^{n} ||v_i||_V^2 \,, \tag{4.6}$$

where $\bar{y}$ is defined by equation (4.3).
The estimation of those two models can be done using a usual gradient descent.

## 4.3 Simulated example

### 4.3.1 Data set

To test the model, we create a simulated data set of 500 subjects deformed from a common template to which we add a random number of dark spots (between 1 and 5) and some Gaussian noise. This template is created as the deformation of an ellipse (hypertemplate). We also create a "control" data set of 100 subjects without dark spot from the same template to be able to estimate a control template $\bar{y}$ and use it in the case of the model (4.2). The template and five subjects with dark spots are presented on the first line of Figure 4.2.

### 4.3.2 Application of the models presented section 4.2

We first apply the model where the template is directly estimated, without the use of a hypertemplate nor sparse matrices (Equation (4.5)). As can be seen on the second line of Figure 4.2, a dark shadow is created on the estimated template. Similarly, this dark shadow is reported on the reconstructions of each subject. Moreover, as can be seen on the last two columns, to minimize $J_0$ the algorithm sometimes badly estimates the deformed object in order not to include a dark spot.

On the third line of Figure 4.2, we apply the model where the template is obtained from an ellipse hypertemplate but without any anomaly matrix (model (4.6)). This time, as expected, there is no dark shadow on the template but the reconstruction of the last two subjects is still

bad.

We also test the model (4.2). To do so, we begin by estimating a template from the 100 "control" subjects. We then fix this template in the minimization of Equation (4.2). This estimated template and some reconstructions are represented on the fourth line of figure 4.2. This time, the subjects are better reconstructed and the dark spots retrieved. Their intensity is a bit weaker than initially due to the proximal gradient descent applying a soft threshold on the residuals. Moreover, the shapes of the dark spots are well identified and so is their position.

Then, we apply the model (4.4) where the template is estimated from a hypertemplate. The results are presented on the last line of Figure 4.2. Once again, the subjects are well reconstructed and the dark spots are retrieved in the anomaly matrix. Shapes, positions and volumes of the dark spots are captured.

Finally, we compare the four models by computing the mean error of registration when one only considers the form (i.e. the error between the observations without their dark spots and the reconstructions without their anomaly matrix) on table 4.1. As expected, the error is smaller when considering the models (4.2) and (4.4).

| | Model (4.5) | Model (4.6) | Model (4.2) | Model (4.4) |
|---|---|---|---|---|
| Error of registration | $19,8\%_{\pm 0.06}$ | $16.5\%_{\pm 0.10}$ | $11.9\%_{\pm 0.07}$ | $12.9\%_{\pm 0.03}$ |

**Table 4.1** Mean and standard deviation of the error of registration when one does not consider the dark spots.

Hence, we have been able to reconstruct the subjects and to retrieve their anomalies using the models (4.2) and (4.4). Moreover, we have been able to highlight an improvement of reconstruction when we estimate both anomaly matrix and deformations.

### 4.3.3 On the choice to estimate anomaly matrix and deformation at the same time

The choice to estimate both the deformation $v_i$ and the anomaly matrix $A_i$ at the same time, and not one after the other, comes from an effort to improve the reconstruction of the object. Indeed, we can see on the Table 4.1 that the errors of reconstruction are smaller when we estimate both at the same time. It can particularly be seen on the third line of Figure 4.2 where, for the last two columns, the residuals contain whole parts of the observations. Those parts would hence be retrieved in the anomaly matrix.

We present a last example to emphasize this need to estimate both the deformation and the anomaly matrix at the same time. The different images of that example can be seen Figure 4.3. This time, we add a little black line on the control template, mimicking for example a vein in a liver or a gyrus in a brain slice. From this template we create one subject to which we add a dark spot. We try to reconstruct this subject from the template, either with or without anomaly matrix. Without anomaly matrix, the model chooses to heavily deform the black line to create the dark spot. In particular, a part of the line would here be in the residual and the black spot would not entirely be in it. The estimation of an anomaly matrix from this residual would hence be bad. However, if we choose to estimate both deformation and anomaly matrix at the same

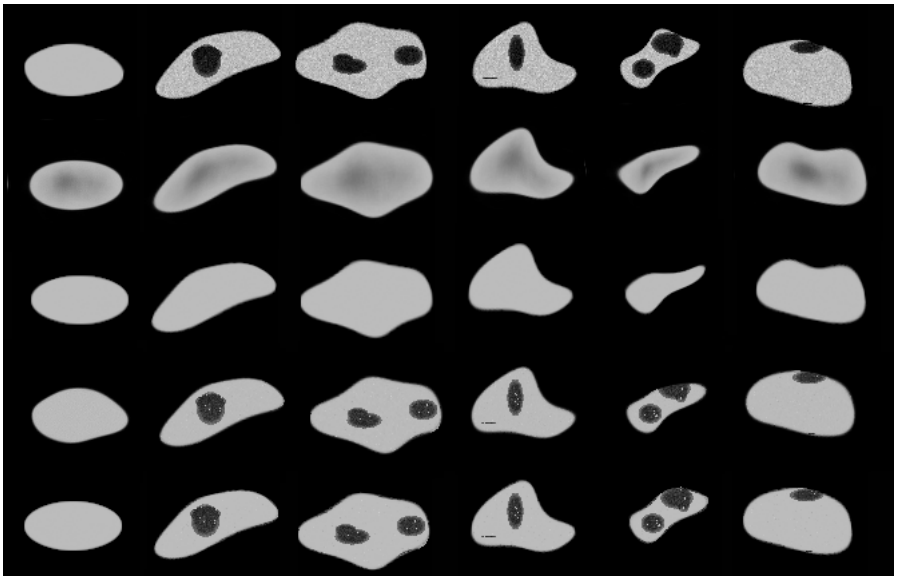


**Figure 4.2** On the top line, the template and five observations. On the second line, the estimated template and reconstructed subjects without the use of hypertemplate nor anomaly matrix (model (4.5)). On the third line, the results when one uses the hypertemplate but no anomaly matrix (model (4.6)). On the fourth line, we first estimate a template (first column) from control subjects and then reconstruct the other subjects using an anomaly matrix (model (4.2)). On the last line, the results with sparse matrices when the estimation of a template is done at the same time, from a hypertemplate (model (4.4))

time, the black line is well registered from the template to the individual and only the black spot is retrieved in the anomaly matrix.

Those two observations confirm the need to couple the estimation of anomaly matrices and deformations.

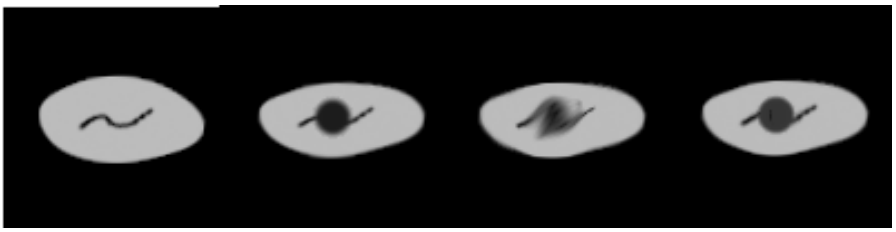


**Figure 4.3** From left to right, the fixed template, the subject, the reconstruction without anomaly matrix and the reconstruction with anomaly matrix. Without anomaly matrix, the algorithm uses the black line to recreate the dark spot, creating a fuzzy black zone.

## 4.4 Application to a data set of brains with tumors

### 4.4.1 Presentation of the data set

We choose to first apply our model to a data set of brains with tumors obtained from the BraTS 2018 data set (Bakas et al., 2017, 2018; Menze et al., 2014). More precisely, the data set is composed of 50 post-contrast T1-weighted MRI scans of glioblastoma and lower grade glioma. We do not dispose of a data set of control subjects but only of the observation of one control subject. We will hence use the model (4.4) to estimate the template using this control subject as hypertemplate.

The goal is to reconstruct the brain of each subject from a control template and to obtain the tumors in the anomaly matrix. Ideally, the diffeomorphic deformation will register the brain folds (called gyri) and ventricles. What cannot be retrieved in the diffeomorphic deformation should hence only be the tumors.

### 4.4.2 Results

We here present the results of our model applied to this data set. On figure 4.4, we present the template estimated by our algorithm.

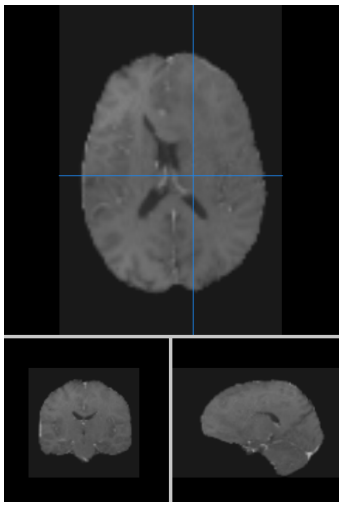


**Figure 4.4** The template estimated using a control patient as a hypertemplate.

On Figure 4.5, we show the results for four subjects. As can be seen, the lesions are retrieved in the anomaly matrix with only small errors of reconstruction. In particular, we can see that the gyri have been well registered and are not present in the anomaly matrix. As for the ventricles, if a part of one is present in the bottom right image, they are also well registered for the other images. In fact, for the bottom right patient, its right ventricle is quite different from the template, causing our algorithm difficulties to register a part of it. But, in all cases, the use of the LDDMM framework has allowed to register most of the parts of the brain without anomaly and so to obtain a clean anomaly matrix.

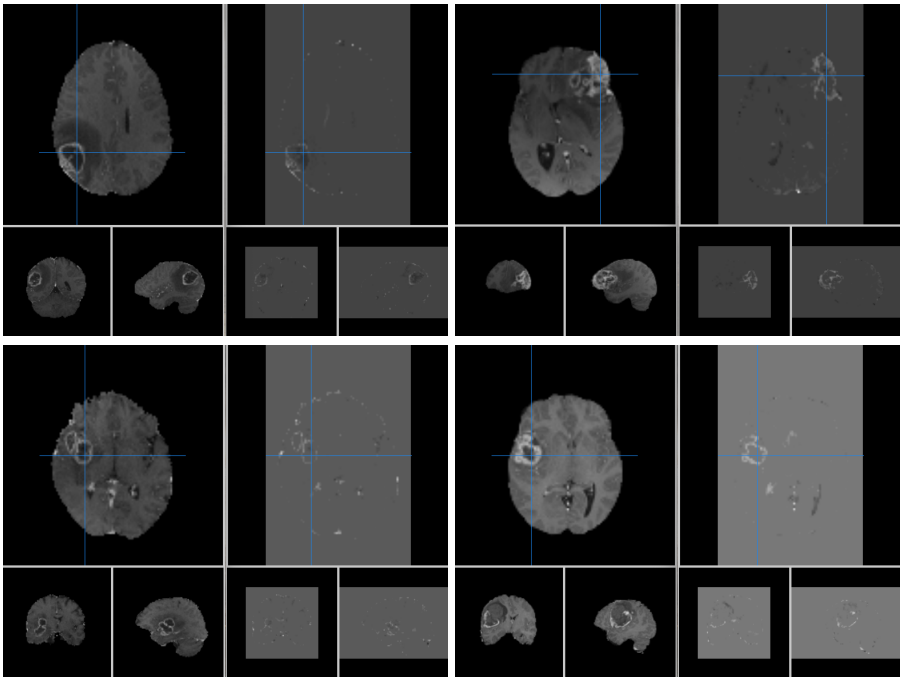


**Figure 4.5** The results for four different subjects. Each time, we put on the left, the observation and on the right, the anomaly matrix estimated. Each time, the lesions are well retrieved.

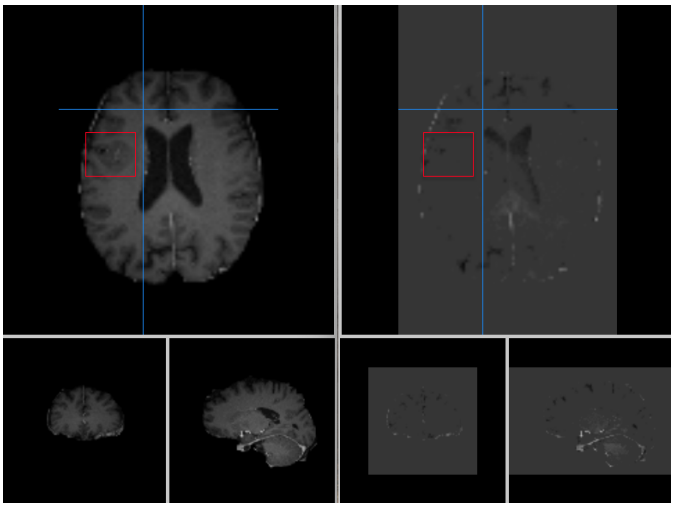


**Figure 4.6** The patient and the anomaly matrix in a case where the tumor (red zone) is not detected.

Here, the important choice of parameter in Equation (4.4) is in fact the sparsity constant $\lambda$. One could choose to take a smaller $\lambda$. This would allow to include the peritumoral edema (dark area around the tumor) in the anomaly matrix. However, we would then also include more reconstruction errors. Here, we have chosen $\lambda$ in order to visually detect the tumors but with as

little reconstruction errors as possible, even if the whole tumor is not in the anomaly matrix. It fulfils our goal as we wanted to be able to inform the doctors of possible anomalies, which are indeed included in the anomaly matrix.

In fact, of the 62 tumors in the data set, 59 are visible in the anomaly matrix ($95\%$). As for the tumors not visible, they are small lesions in a zone of high variability of the brain. We show such a case in Figure 4.6 where the tumor is less easy to distinguish and in the middle of the brain gyri.

### 4.4.3 Application with only one control and one sick subject

Our approach here is particularly interesting as it does not require a large data set to be applied and there is no annotation on the position of the tumors required. To highlight this advantage, we apply the exact same algorithm to only one brain with tumors. As for the template, it is fixed as a brain without tumor. Hence, we only use two different brains to try to detect an anomaly. In particular, we compare the results with the anomaly matrix estimated for this patient in the previous section where a template was estimated alongside (see Figure 4.7).

The tumor is once again retrieved in the anomaly matrix with small errors of reconstruction, particularly on the border and top of the brain. The errors on the top, in particular, are not present when one estimates a template alongside the anomaly matrix. In fact, the variability between the control subject and the one with tumor is bigger there, causing bigger errors of reconstruction. But, estimating a template, even with few subjects as done section 4.4.2, prevents this issue.

Even if more errors are included in the sparse matrix, it must be emphasized that it has been produced using only two subjects and that it shows that our method can yield usable results even without access to more than a few subjects.

## 4.5 Application to the liver data set

### 4.5.1 Pre-processing

We now apply our algorithm to the liver data set. The goal is to reconstruct the liver of each patient while recovering the tumors in the anomaly matrix. As we do not have access to a full data set of control patients, we cannot use the model (4.2) and we need to estimate the template from a hypertemplate using model (4.4). The hypertemplate is chosen as a patient, not in the database, and without tumor.

A first problem that arises from this data set is that the images contain the whole abdomen while we only interest ourselves in the liver. If we try to directly create an atlas using the framework presented before, the algorithm will find the template and deformations by minimizing a function measuring the distance between the subjects and the deformed template. However, most of the variability will come from differences of size, position and form of organs other than the liver and the registration of the liver could be bad. To prevent this problem, a collaboration with Philips has been set up allowing us to segment the liver. They provided us with a software (Intellispace Portal pre-release AI Liver segmentation version 3.0.5, date 2020) allowing the

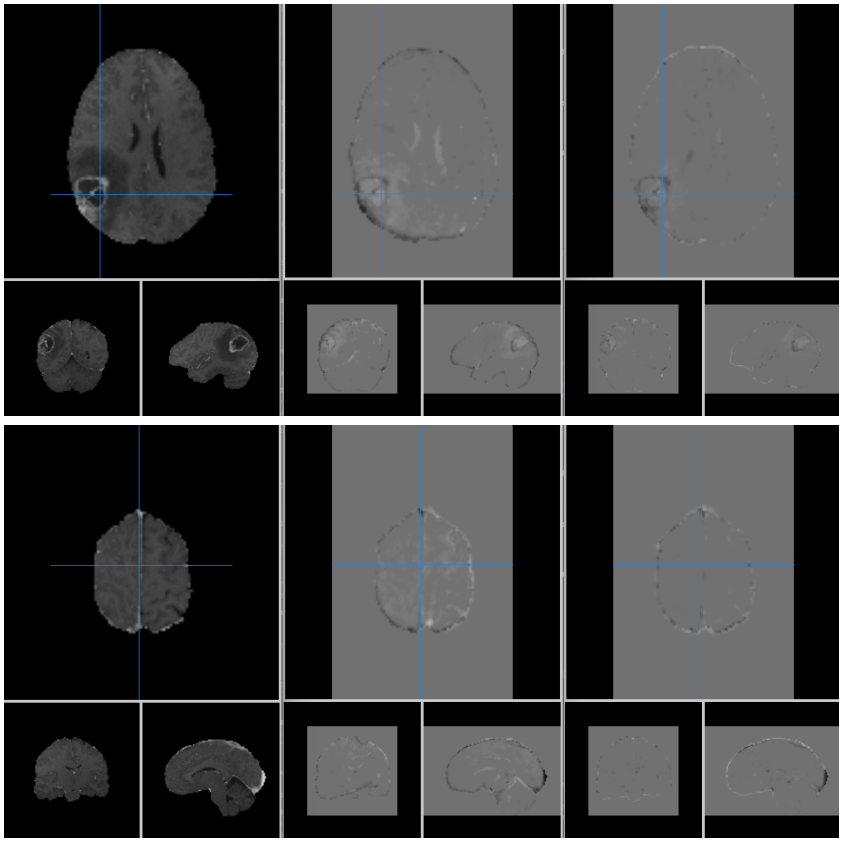


**Figure 4.7** On the left, the observation. In the center, the anomaly matrix estimated with the template fixed as a control subject. On the right, when one estimates the template, as done section 4.4.2. The results are showed for two different slices: at the position of the tumor (top images) and at the top of the brain (bottom images). In both cases, the tumor is retrieved. More reconstruction errors are included when one does not estimate a template.

segmentation of livers using deep learning. An example of such a segmentation is presented figure 4.8. On that figure, the dark spots are tumors.

Two different structures appear on the livers: tumors as dark spots and vessels as white structures. Hence, it is those two structures we will recover in the anomaly matrix. If one only wants to retrieve the tumors, it is easy to separate them from the vessels according to their intensity. Moreover, from one subject to the other, the noise level can be totally different. If we do not preprocess the data set, we would not be able to find a sparsity constant $\lambda$ efficient for each subject and, for those with the highest level of noise, this noise would be recovered in the anomaly matrix. To prevent this phenomenon, we decide to first convolve our observations with a Gaussian kernel. The resulting images can be seen Figure 4.9. This smoothing allows to have a robust algorithm for this population.

Moreover, because all images do not have the same pixel spacing, we need to down-sample some of the images. We also include all of them in a black box of the same size.

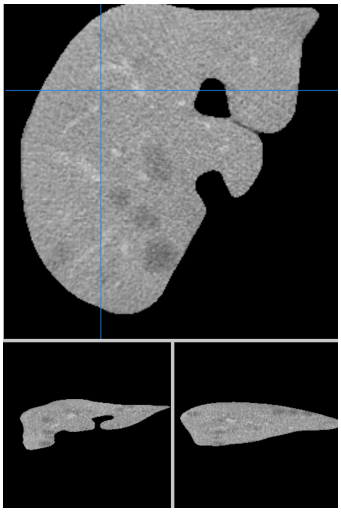


**Figure 4.8** Example of a segmented liver.

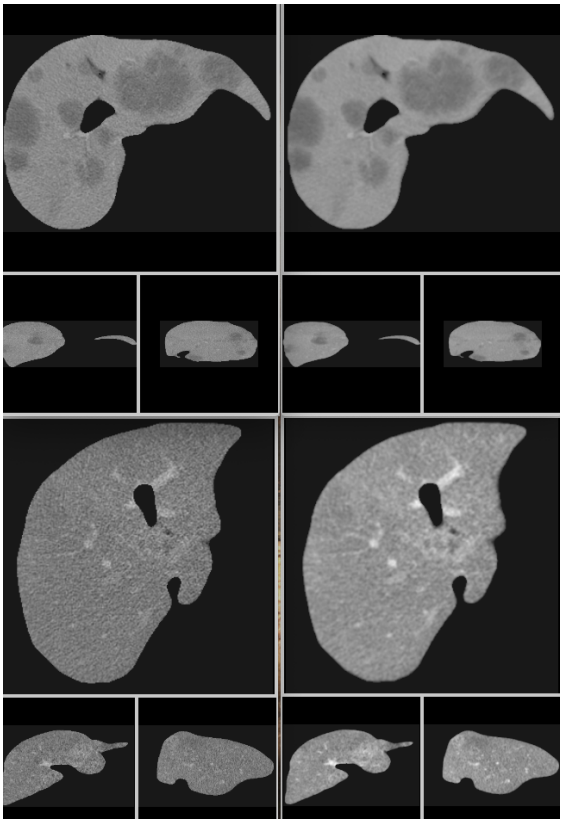


**Figure 4.9** On the left, the initial images. On the right, the same subjects after convolution with a Gaussian kernel.

In the following, we present the results for different subjects.

We also choose, as a post process, to put the coefficients of the anomaly matrix to $0$ outside of the reconstructions and targets. If one does not make this choice, the errors of reconstruction are reported in the anomaly matrix. In particular, one would find white zones in the anomaly matrix at voxels where the diffeomorphic deformation has not been able to recreate a liver part and black zones where the diffeomorphic deformation has created a liver part at a place there should not be one.

### 4.5.2 Presentation of the results

We begin by presenting the estimated template $\bar{y}$ in figure 4.10. This estimation would particularly benefit from the use of a whole data set of control patients. Here, because it is derived from a particular control subject, the vessels of this particular subject are still present in the final template and can influence the future estimation of the anomaly matrix. Having a template of control subject would also surely reduce the errors of reconstruction and allow a better detection of the anomalies.

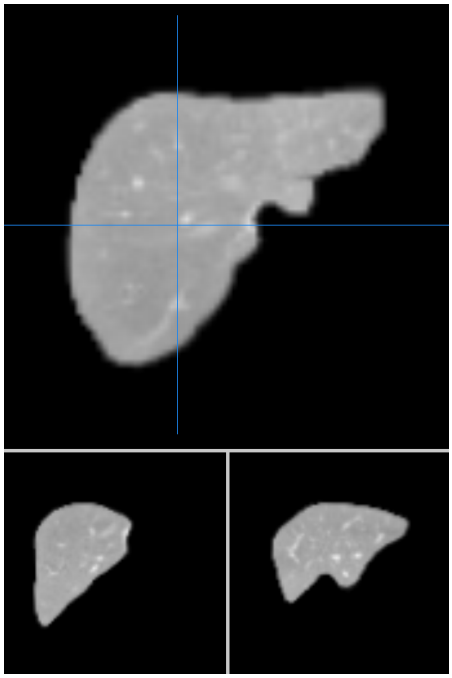


**Figure 4.10** Template estimated as a diffeomorphic deformation of a control patient.

Then, on Figure 4.11, we show the results for a subject without any vessel visible on the scanner. The algorithm is able to retrieve the tumors in the anomaly matrix. The outline of the registration is also present in the anomaly matrix as it is used to obtain a better final reconstruction.

We then present a subject with vessels visible in Figure 4.12. As can be seen, not only the

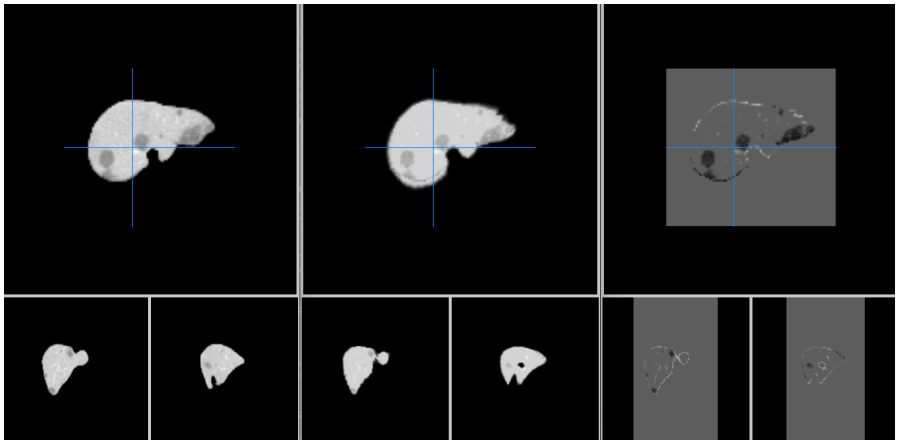


**Figure 4.11** On the left, the observation. In the center, the reconstruction. On the right, the anomaly matrix.

tumors are retrieved but also the vessels. If one wants to only find the tumors, a first possibility is to look at the negative values of the anomaly matrix. Further investigation on a post process would be required to only retrieve the anomalies without the small errors of reconstruction.

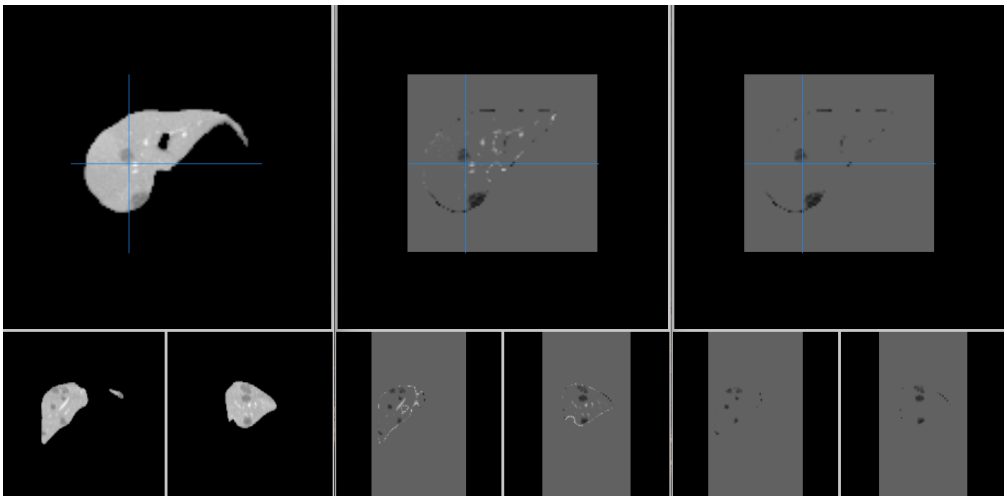


**Figure 4.12** From left to right, the observation, the anomaly matrix, and the negative values of the anomaly matrix.

Finally, we show the importance to apply a Gaussian convolution to the data set before estimating the parameters of the model figure 4.13. If one does not perform this preprocessing, the noise is retrieved in the anomaly matrix and the tumors are not visible. This problem is indeed solved after convolution and the tumor (at the top left of the liver) is retrieved.

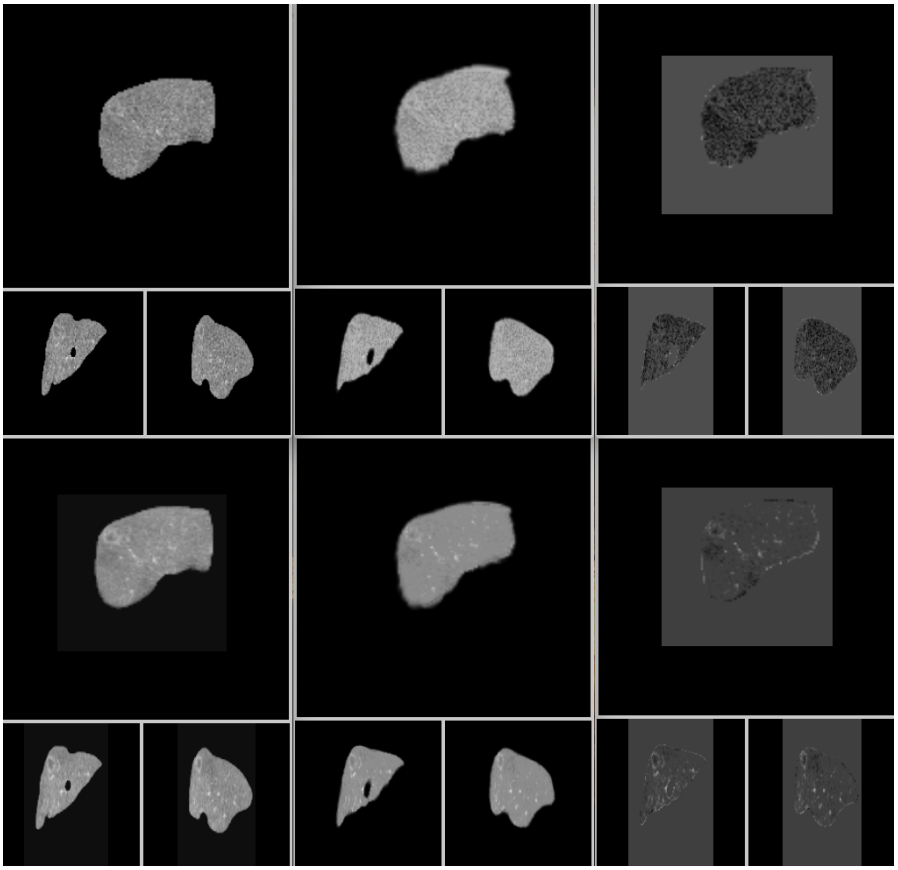


**Figure 4.13** On the top line, the results if we do not convolve the data set with a Gaussian kernel beforehand. On the bottom, with preprocessing. On the left, the observation. In the center, the reconstruction. On the right, the anomaly matrix.

### 4.5.3 Quality of the detection

To measure the quality of the detection, we asked a MD Radiologist to segment the tumors of 10 patients. This led to a total number of 133 tumors segmented. Here, we will only look at the negative coefficients of the sparse matrices to measure the quality of detection as tumors are dark spots on the liver.

We choose not to evaluate the segmentation but the detection. In fact, here, the dice score would be average as our algorithm rarely segments the whole tumor. However, here, we do not want to segment the exact tumor but only to inform the doctor of a possible anomaly and particularly detect very small lesions.

On the 133 tumors segmented, our algorithm detects 125 of them ($94\%$). As for the tumors which are not detected, there are two possibilities. Sometimes, the difference between the tumor intensity and the noise is really small and the algorithm is not able to separate them, in particular for some subjects for which the noise is still high. The other possibility is when the diffeomorphic registration of the liver is not perfect and the tumor is outside of it. In that case,

the tumor will in fact be in the anomaly matrix but it will be lost in the error of reconstruction.

Finally, not only the tumors are retrieved in the anomaly matrix. As showed above, small errors of reconstruction can be present on the boundary. But our algorithm also plays its part by detecting other anomalies than lesions. In particular, on several subjects, some slightly dark spots are retrieved and are in fact due to a perfusion disorder.

## 4.6 Conclusion

In this chapter, we have showed that we can use the residuals of the diffeomorphic deformation from a control template to detect and segment lesions in an organ. Moreover, we showed that it can even improve the diffeomorphic reconstruction of the observations. This method has the advantage not to require a large data set of sick patients nor annotations from medical doctors. It is hence particularly suited to the detection of anomalies, in particular for specific treatment protocols where it is often impossible to obtain large data sets. We have showed the efficiency of this method on a data set of brains with glioma and on a data set of livers. In particular, in the former case, our algorithm has been able to register the gyri and the ventricles of the brains while retrieving the tumors in the anomaly matrix.

Several problems still need to be solved. First, if this method detects the tumors it also detects small errors of reconstruction. One would need to post process the image to only obtain the lesions. A first idea is to use openings but this poses the risk to lose some little anomalies in the post process. Moreover, a better treatment of the errors of reconstruction is needed as a tumor not in the reconstruction of the object is lost. Further post-processing of the anomaly matrix would be needed to retrieve them.

Moreover, with this data set of patients with colorectal cancer, the goal was to predict the progression-free survival at 9 months. With only 57 patients, this is not an easy task. Because the position of the tumors could be of importance, one cannot use usual techniques of rotation and translation to augment the data. It could however be possible to use Variational Auto Encoders to create new synthetic subjects. Another possibility is to further inform our model by only giving it the segmented tumors transported on the template: $\left(\mathcal{E}xp_{0,1}(v_i)^{-1}(A_i)\right)_{i\leq 1\leq n}$. This would allow to inform the network of the tumors while erasing the variability due to differences of livers form and size between subjects. Compared to learning on the whole image, one however loses information on the environment around the tumors.

Finally, we do not only have access to the first exam but also to subsequent ones. It could be interesting to add a longitudinal model on the segmentation to study the evolution of the lesions. This would however not be an easy task as we cannot directly apply the model presented in chapter 3. Indeed, from one time point to the other, tumors can appear or disappear. Hence, the longitudinal trajectory of the tumors is not the diffeomorphic deformation of an initial one.

At the time of writing of this manuscript, those problems are still being studied.

# Part III

# On the convergence properties of Stochastic Approximation algorithms

# CHAPTER 5

## On the convergence of stochastic approximations under a subgeometric ergodic Markov dynamic

In the chapter 3, we have created a statistical model for longitudinal data sets. To estimate the parameters of this model, we used the Monte Carlo Markov Chain Stochastic Approximation Expectation Maximization (MCMC SAEM) algorithm. This algorithm in particular lies on the theoretical foundations of Stochastic Approximations (SA) with Markovian dynamic and its theoretical convergence requires the geometric ergodicity of the chain considered. This can however be a problem in practice as this assumption is not always verified.

In this chapter, to solve this issue, we extend the framework of the convergence of stochastic approximations. Such a procedure is used in many methods such as parameters estimation inside a Metropolis Hastings algorithm, stochastic gradient descent or stochastic Expectation Maximization algorithm. It is given by

$$\theta_{n+1} = \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}) \,,$$

where $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables following a parametric distribution which depends on $(\theta_n)_{n\in\mathbb{N}}$, and $(\Delta_n)_{n\in\mathbb{N}}$ is a step sequence. The convergence of such a stochastic approximation has already been proved under an assumption of geometric ergodicity of the Markov dynamic. However, in many practical situations this hypothesis is not satisfied, for instance for any heavy tail target distribution in a Monte Carlo Metropolis Hastings algorithm. In this chapter, we relax this hypothesis and prove the convergence of the stochastic approximation by only assuming a subgeometric ergodicity of the Markov dynamic. This result opens up the possibility to derive more generic algorithms with proven convergence. As an example, we first study an adaptive Markov Chain Monte Carlo algorithm where the proposal distribution is adapted by learning the variance of a heavy tail target distribution. We then apply our work to the Independent Component Analysis when a positive heavy tail noise leads to a subgeometric dynamic in an Expectation Maximization algorithm.

This work has been published in the Electronic Journal of Statistics (Debavelaere et al., 2021).

## Contents

## 5.1 Introduction

A common problem across scientific fields is to find the roots of a non-linear function $h : \Theta \to \mathbb{R}$. Numerical schemes such as Newton's methods have been developed to provide a numerical solution to this equation. In statistics, the problem is further increased by the fact that $h$ is not known, but only noisy values of it, or of its gradient. This problem appears across different domains such as stochastic optimization (Mandt et al., 2017; Spall et al., 1992), Expectation Maximization algorithms (Allassonnière et al., 2010; Kuhn et al., 2019) or reinforcement learning (Abounadi et al., 2002; Borkar and Meyn, 2000) for instance. In all cases, solutions to this problem often take the form of an iterative sequence $(\theta_n)_{n\in\mathbb{N}}$ that converges towards a point $\theta^*$ in the set of solutions of $h(\theta) = 0$. The general class of stochastic approximation methods, such as Robbins-Monro methods, falls within this framework. These methods produce a sequence of the form:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}\zeta_{n+1}\,,$$

where $\zeta_{n+1}$ is a noisy observation of $h(\theta_n)$: $\zeta_{n+1} = h(\theta_n)+\xi_{n+1}$ with $\xi_{n+1}$ a sequence of random variables. In that case, $h$ is called the mean field. This procedure, first developed in Robbins and Monro (1951), has been studied under various sets of hypotheses, see Abounadi et al. (2002); Benveniste et al. (2012); Borkar and Meyn (2000); Chen (2006); Chen et al. (1987); Duflo (2013); Kushner and Yin (2003) among many other works.

In this paper, we focus on the case of a state-dependent noise with a Markovian dynamic. The sequence $(\zeta_n)_{n\in\mathbb{N}}$ takes the form of $(H_{\theta_n}(X_n))_{n\in\mathbb{N}}$, with $h(\theta_n)$ being the expectation of $H_{\theta_n}$:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}H_{\theta_n}(X_{n+1})\,. \tag{5.1}$$

The sequence $(X_n, \theta_n)_{n\in\mathbb{N}}$ is a Markov chain on $\mathcal{X} \times \Theta$. For all $\theta \in \Theta$, $H_\theta$ is a function from the state space $\mathcal{X}$ to the parameter space $\Theta$.

The assumption of state-dependent noise is met for instance in stochastic gradient descent or Metropolis Hastings algorithms. Eq. (5.1) is also used as a step in stochastic optimization algorithms where the parameter to estimate is a function of $\theta_n$. These algorithms include the Stochastic Approximation Expectation Maximization Markov Chain Monte Carlo (SAEM MCMC) algorithm (Allassonnière et al., 2015, 2010; Delyon et al., 1999). Eq. (5.1) also appears in some adaptive MCMC algorithms where the proposal distribution depends on a parameter $\theta$. They are used to adapt the variance of the proposal across iterations for better sampling (Andrieu et al., 2005; Andrieu and Robert, 2001; Haario et al., 2001; Roberts and Tweedie, 1996).

The convergence of stochastic approximation algorithms has been studied in Andrieu et al. (2005) for state-dependent noise. Conditions to ensure convergence include control of the fluctuations of the Markov Chain and of the regularity of the solution of a Poisson equation. These conditions are difficult to verify in practice. Authors introduce then a more restrictive, but more practical condition: the Markov chain must satisfy drift conditions implying a *geometric ergodicity* of the chain. This condition amounts to assuming the convergence of the kernel of the Markov Chain towards its invariant distribution at a geometric rate. Further developments lead to prove the convergence of the SAEM MCMC algorithm (Allassonnière et al., 2010), some adaptive MCMC algorithms (Andrieu et al., 2005) and mini-batch MCMC (Kuhn et al., 2019) under the same conditions.

Nevertheless, the ergodicity condition is a limiting factor in practice. For instance, the sequence $(X_n)_{n\in\mathbb{N}}$ is often sampled using a Metropolis Hastings algorithm. The ergodic condition is not met if one targets heavy tail distributions such as Weibull or Pareto distribution (Douc et al.,

2004; Fort and Moulines, 2000, 2003; Jarner and Hansen, 2000). The models for independent component analysis presented in Allassonniere et al. (2012) with non-Gaussian distributions of the sources or the noise do not meet the condition either. These examples show that these methods may be used in practice without any theoretical guarantee of convergence.

This situation leads us to study the convergence of such stochastic algorithms for Markov chains with a relaxed assumption of subgeometric ergodicity. The convergence of adaptive MCMC algorithms under subgeometric constraints has been studied in Atchadé et al. (2010, 2012); Rosenthal and Roberts (2007); Yang (2008). To the best of our knowledge, there are no results on the convergence for subgeometric Markovian dynamic in the general case.

In this paper, we propose a general set of hypotheses, under which we prove the convergence of stochastic approximations with subgeometric Markovian dynamics. Our hypotheses are essentially about the rate of convergence of the Markov Chain and the regularity of its kernel. Most of the polynomial rates of convergence satisfy these hypotheses. Furthermore, the proof shows the regularity of the solution of the Poisson equation under the same subgeometric conditions. We use this result to prove two corollaries. The first corollary proves the convergence of a stochastic approximation used to adapt the variance of the proposal within a Metropolis Hastings algorithm. We prove this convergence for two different classes of heavy tail target distributions including the Weibull and the Pareto distributions among others. The second corollary is about the independent component analysis model where distributions with positive heavy tails lead to a subgeometric ergodic Markov Chain in a Stochastic Approximation Expectation Maximization Monte Carlo Markov Chain (SAEM MCMC) algorithm.

## 5.2 Stochastic approximation framework with Markovian dynamic

In this section, we summarize the stochastic approximation procedure in the case of a Markovian dynamic with adaptive truncation sets. This procedure was first described in Andrieu et al. (2005). In the following, we denote $\mathcal{X}$ the state space and $\Theta$ the parameter space that we assume to be an open subset of $\mathbb{R}^{n_\theta}$. Moreover, we suppose that both are equipped with countably generated $\sigma$-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$.

In the next subsection, we present the framework of a stochastic approximation producing a sequence of elements converging towards a solution of $h(\theta) = 0$ when there exist probability measures $\pi_\theta$ such that, for any $\theta \in \Theta$, $h(\theta) = \mathbb{E}_{\pi_\theta}(H_\theta(X))$ with $H_\theta : \mathcal{X} \mapsto \Theta$.

### 5.2.1 Markovian dynamic

Let $\Delta = (\Delta_n)_{n\in\mathbb{N}}$ be a non-increasing sequence of positive real numbers with $\Delta_0 \leq 1$ and set $\theta_c \notin \Theta$ and $x_c \notin \mathcal{X}$ two cemetery states. We also set, for all $\theta \in \Theta$ the vector field $H_\theta : \mathcal{X} \mapsto \Theta$. We then define a Markov chain $Y_n^\Delta = (X_n, \theta_n)$ on $\mathcal{X} \cup \{x_c\} \times \Theta \cup \{\theta_c\}$ by:

$$\theta_{n+1} = \begin{cases} \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}) & \text{and } X_{n+1} \sim P_{\theta_n}(X_n, .) \quad \text{if } \theta_n \in \Theta \\ \theta_c & \text{and } X_{n+1} = x_c \quad \text{if } \theta_n \notin \Theta\,. \end{cases} \tag{5.2}$$

Keeping notations and hypotheses labels from Andrieu et al. (2005), we put the following hypothesis on the transition probabilities $(P_\theta, \theta \in \Theta)$ and on the random vector field $H$:

**(A2)** For any $\theta \in \Theta$, the Markov kernel $P_\theta$ has a single stationary distribution $\pi_\theta$. In addition, $H : \Theta \times \mathcal{X} \to \Theta$ is measurable for all $(\theta, x) \in \Theta \times \mathcal{X}$ .

The existence and uniqueness of the invariant distribution can be verified under the classical conditions of irreducibility and recurrence (Meyn and Tweedie, 2012). We also set $h(\theta) = \int_{\mathcal{X}} H_\theta(x)\pi_\theta(dx)$ the mean field of the stochastic approximation. This allows us to recognize the usual stochastic approximation procedure:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}(h(\theta_n) + \xi_{n+1})$$

where $\xi_{n+1} = H_{\theta_n}(X_{n+1}) - h(\theta_n)$ is the noise sequence.

We assume the mean field $h$ satisfies the following hypothesis that amounts to the existence of a global Lyapunov function:

**(A1)** $h : \Theta \to \mathbb{R}^{n_\theta}$ is continuous and there exists a continuously differentiable function $w : \Theta \to [0, +\infty[$ such that:

(i) there exists $M_0 > 0$ such that

$$\mathcal{L} := \{\theta \in \Theta, \langle \nabla w(\theta), h(\theta)\rangle = 0\} \subset \{\theta \in \Theta, w(\theta) < M_0\} \ ,$$

(ii) there exists $M_1 \in (M_0, +\infty]$ such that $\mathcal{W}_{M_1} := \{\theta \in \Theta, w(\theta) \leq M_1\}$ is a compact set,

(iii) for any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla w(\theta), h(\theta)\rangle < 0$,

(iv) the closure of $w(\mathcal{L})$ has an empty interior.

We denote by $\mathcal{F} = \{\mathcal{F}_n, n \geq 0\}$ the natural filtration of the Markov chain $(X_n, \theta_n)$ and by $\mathbb{P}^{\Delta}_{x,\theta}$ the probability measure associated to the chain $(Y^{\Delta}_n)$ started from the initial conditions $(x, \theta) \in \mathcal{X} \times \Theta$. Finally, we denote by $Q_{\Delta_n}$ the sequence of transition probabilities that generate the inhomogeneous Markov chain $(Y^{\Delta}_n)$.

### 5.2.2 Truncation process

To ensure convergence of the sequence towards a root of $h$, the sequence $(\theta_n)_{n\in\mathbb{N}}$ is required to remain in a given compact set. This assumption is rarely satisfied. To alleviate this constraint, we introduce the usual trick which consists in reprojecting on increasing compact sets. It is then proved that the sequence will be projected only a finite number of times along the algorithm. Using this trick, the sequence $(\theta_n)_{n\in\mathbb{N}}$ now remains in a compact set of $\Theta$. We detail this process below.

We assume that there exists $(\mathcal{K}_n)_{n\in\mathbb{N}}$ a sequence of compact subsets of $\Theta$ such that

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta \qquad \text{and} \qquad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}) \, .$$

Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence of non-increasing positive numbers and $K$ be a subset of $\mathcal{X}$. Let $\Phi : \mathcal{X} \times \Theta \to K \times \mathcal{K}_0$ be a measurable function. We then define the stochastic approximation

algorithm with adaptive truncation sets as a homogeneous Markov chain on $\mathcal{X} \times \Theta \times \mathbb{N} \times \mathbb{N}$ by

$$Z_n = (X_n, \Theta_n, \kappa_n, \nu_n) \tag{5.3}$$

with the following transition at iteration $n+1$:

- If $\nu_n = 0$, then draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(\Phi(X_n, \theta_n), .)$. Otherwise, draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(X_n, \theta_n, .)$.

- If $|\theta_{n+1} - \theta_n| \leq \varepsilon_n$ and $\theta_{n+1} \in \mathcal{K}_{\kappa_n}$ then set $\kappa_{n+1} = \kappa_n$ and $\nu_{n+1} = \nu_n + 1$. Otherwise, set $\kappa_{n+1} = \kappa_n$ + 1 and $\nu_{n+1} = 0$.

To summarize this process, if our parameter $\theta$ leaves the current truncation set $\mathcal{K}_{\kappa_n}$ or if the difference between two of its successive values is larger than a time dependent threshold $\varepsilon_n$, we reinitialize the Markov chain by a value inside $\mathcal{K}_0$: $\Phi(X_n, \theta_n)$ and update the truncation set to a larger one $\mathcal{K}_{\kappa_n+1}$ as well as the threshold to a smaller one: $\varepsilon_{n+1}$. Hence, $\kappa_n$ represents the number of re-initializations before the step $n$ while $\nu_n$ is the number of steps since the last re-initialization.

The idea behind this truncation process is to force the noise to be small in order for the drift $h(\theta)$ to dominate. We do so by forcing our algorithm to come back to the center of $\Theta$ whenever the parameters become too large.

### 5.2.3 Control of the fluctuations and main convergence theorem

In this section, we state two last hypotheses about the control of fluctuations before presenting the theorem proved in Andrieu et al. (2005). In that paper, the authors present several conditions (A1 to A4) that imply the convergence of the stochastic approximation algorithm. It is those conditions that we will, in the next section, verify under subgeometric ergodicity of the Markov chain.

We first define, for any compact $\mathcal{K}$ and any sequence of non-increasing positive numbers $(\varepsilon_k)_{k\in\mathbb{N}}$, $\sigma(\mathcal{K}) = \inf(k \geq 1, \theta_k \notin \mathcal{K})$ and $\nu_\varepsilon = \inf(k \geq 1, |\theta_k - \theta_{k-1}| \geq \varepsilon_k)$. Moreover, for $W : \mathcal{X} \to [1, \infty)$ and $g : \mathcal{X} \to \mathbb{R}^{n_\theta}$, we write

$$||g||_W = \sup_{x\in\mathcal{X}} \frac{|g(x)|}{W(x)}.$$

We can now present the hypothesis (A3):

**(A3)** For any $\theta \in \Theta$, the Poisson equation $g - P_\theta g = H_\theta - h(\theta)$ has a solution $g_\theta$. Moreover, there exist a function $W : \mathcal{X} \to [1, +\infty]$ such that $\{x \in \mathcal{X}, W(x) < +\infty\} \neq \emptyset$, constants $\alpha \in (0, 1]$ and $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(i) the following holds:

$$\sup_{\theta \in \mathcal{K}} ||H_\theta||_W < \infty \tag{5.4}$$

$$\sup_{\theta \in \mathcal{K}} ||g_\theta||_W + ||P_\theta g_\theta||_W < \infty \tag{5.5}$$

$$\sup_{\theta, \theta' \in \mathcal{K}} ||\theta - \theta'||^{-\alpha} \left( ||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W \right) < \infty \tag{5.6}$$

(ii) there exist constants $\{C_k, k \geq 0\}$ such that, for any $k \in \mathbb{N}$, for any sequence $\Delta$ and for any $x \in \mathcal{X}$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k) \mathbb{1}_{\sigma(\mathcal{K}) \geq k}] \leq C_k W^p(x) \tag{5.7}$$

(iii) there exist a sequence $(\varepsilon_k)_{k \in \mathbb{N}}$ and a constant $C$ such that for any sequence $\Delta$ and for any $x \in \mathcal{X}$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k) \mathbb{1}_{\sigma(\mathcal{K}) \wedge \nu_\varepsilon \geq k}] \leq C W^p(x) \,. \tag{5.8}$$

This assumption concerns the existence and regularity of the Poisson equation associated with each of the transition kernel $P_\theta$. In Andrieu et al. (2005), the authors show that those conditions are verified under the hypothesis of geometric ergodicity of the Markov chain. In the next sections, we will relax this ergodicity condition to be able to consider subgeometric ergodic chains.

Finally, the last condition concerns the step size sequences:

**(A4)** The sequences $(\Delta_k)_{k \in \mathbb{N}}$ and $(\varepsilon_k)_{k \in \mathbb{N}}$ are non-increasing, positive and satisfy $\sum_{k=0}^{\infty} \Delta_k = \infty$, $\lim_{k \to \infty} \varepsilon_k = 0$ and

$$\sum_{k=1}^{\infty} \Delta_k^2 + \Delta_k \varepsilon_k^\alpha + (\varepsilon_k^{-1} \Delta_k)^p < \infty$$

where $p$ and $\alpha$ are defined in (A3).

We can finally state the theorem proved in Andrieu et al. (2005):

**Theorem 5.2.1** *Andrieu et al. (2005) Assume (A1)-(A4). Let $K \subset \mathcal{X}$ such that $\sup_{x \in K} W(x) < \infty$ and such that $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ (where $M_0$ and $\mathcal{W}_{M_0}$ are defined in (A1)) and let $Z_n$ be as defined in* (5.3)*. Then, for all $(x, \theta) \in \mathcal{X} \times \Theta$, we have $\lim_{k \to \infty} d(\theta_k, \mathcal{L}) = 0$, $\mathbb{P}^{\Delta}_{x,\theta}$-a.s. where $\mathcal{L}$ is defined in (A1).*

Of the four conditions (A1) to (A4), (A3) is often the most difficult to verify and we need more practical conditions. In particular, in Andrieu et al. (2005), the authors show that drift conditions imply (A3). However, those drift conditions are only true for geometric ergodic Markov chains. In a lot of cases, this ergodicity is not satisfied. To tackle this problem, we will, in the next section, state subgeometric drift conditions and hypotheses on the rate of convergence that are sufficient to ensure the validity of (A3). The new theorem then allows us to verify the convergence in a

broader range of cases, some of them being presented in sections 5.5 and 5.6.

## 5.3 Convergence of the stochastic approximation sequence under subgeometric conditions

In this section, we state the drift conditions and hypotheses under which we will work to prove the validity of (A3). Denote, for $V : \mathcal{X} \to [1, \infty)$, $\mathcal{L}_V = \{g : \mathcal{X} \to \mathbb{R}^{n_\theta}, ||g||_V < \infty\}$.

**(DRI)** For any $\theta \in \Theta$, $P_\theta$ is $\psi$-irreducible and aperiodic. In addition, there exist a function $V : \mathcal{X} \to [1, \infty)$ and a constant $p \geq 2$ such that, for any compact subset $\mathcal{K} \subset \Theta$, there exist constants $b$, $\delta_0 > 0$, a probability measure $\nu$, a concave, increasing function $\phi : [1, \infty) \to (0, \infty)$, continuously differentiable such that $\lim_{v\to\infty} \phi'(v) = 0$ and a subset $\mathcal{C}$ of $\mathcal{X}$ with

$$\sup_{\theta\in\mathcal{K}} P_\theta V^p(x) + \phi \circ V^p(x) \leq V^p(x) + b\mathbb{1}_{\mathcal{C}}(x) \qquad \forall x \in \mathcal{X} \tag{5.9}$$

$$\inf_{\theta\in\mathcal{K}} P_\theta(x, A) \geq \delta_0 \nu(A) \qquad \forall x \in \mathcal{C}, \forall A \in \mathcal{B}(\mathcal{X}) \,. \tag{5.10}$$

**Remark 5.3.1.** *We could consider the following, more general, drift condition: there exists* $m \in \mathbb{N}^*$ *such that*

$$\sup_{\theta\in\mathcal{K}} P_\theta^m V^p(x) + \phi \circ V^p(x) \leq V^p(x) + b\mathbb{1}_{\mathcal{C}}(x) \qquad \forall x \in \mathcal{X}$$

$$\inf_{\theta\in\mathcal{K}} P_\theta^m(x, A) \geq \delta_0 \nu(A) \qquad \forall x \in \mathcal{C}, \forall A \in \mathcal{B}(\mathcal{X}) \,.$$

*The results we present in the following sections would still be verified under such a drift condition. To adapt the proofs (and more precisely, the proof of the lemma 5.4.6), we would then need to use the lemma B.3. of Andrieu et al. (2005).*

Under the condition (DRI), $\mathcal{C}$ is a small set and the Markov kernel $P_\theta$ verifies a subgeometric drift condition (Douc et al., 2018). In particular, it implies the existence of a stationary distribution $\pi_\theta$ for all $\theta \in \mathcal{K}$ as well as a uniform subgeometric ergodicity on all compacts of $\Theta$. Hence, for all $\theta \in \Theta$, there exist a constant $C_\theta$ and a sequence $(r_{\theta,k})_{k\in\mathbb{N}}$ such that, $\forall q, s > 0$ with $1/q + 1/s = 1$ and $\forall f \in \mathcal{L}_{(\phi\circ V^p)^{1/s}}$,

$$r_{\theta,k}^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi\circ V^p)^{1/s}} \leq C_\theta ||f||_{(\phi\circ V^p)^{1/s}} \,.$$

Moreover, it has been showed in Douc et al. (2004) that, under such a subgeometric ergodicity condition, we can choose a rate of convergence $(r_k)_{k\in\mathbb{N}}$ that only depends on the function $\phi$ and so only on the fixed compact $\mathcal{K}$. Similarly, it has been proved that the constant $C_\theta$ is bounded on all compact $\mathcal{K}$. Hence, there exist a constant $C_\mathcal{K}$ and a sequence $(r_k)_{k\in\mathbb{N}}$ such that, for all $f \in \mathcal{L}_{(\phi\circ V^p)^{1/s}}$ and for all $\theta \in \mathcal{K}$,

$$\sup_{\theta\in\mathcal{K}} r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi\circ V^p)^{1/s}} \leq C_\mathcal{K} ||f||_{(\phi\circ V^p)^{1/s}} \,. \tag{5.11}$$

We will see in the following that several hypotheses must be made on that rate of convergence $(r_k)_{k\in\mathbb{N}}$ for the condition (A3) to be satisfied.

**Remark 5.3.2.** *In general, we can consider any pair $\Psi_1$ and $\Psi_2$ of inverse Young functions i.e. two strictly increasing continuous functions on $\mathbb{R}_+$ verifying for all $x, y$ in $\mathbb{R}_+$, $\Psi_1(x)\Psi_2(y) \leq x+y$. Under the subgeometric drift condition, we then have, for all $f \in \mathcal{L}_{\Psi_2(\phi \circ V^p)}$:*

$$\Psi_1(r_k)||P_\theta^k f - \pi_\theta(f)||_{\Psi_2(\phi \circ V^p)} \leq C_{\mathcal{K}}||f||_{\Psi_2(\phi \circ V^p)} \,.$$

*In order to simplify the notations, we will only consider in the following the pair of inverse Young functions $\Psi_1(x) = qx^{1/q}$ and $\Psi_2(x) = sx^{1/s}$. The same reasoning could be carried out for any other pair of Young functions by adapting the hypotheses (H1) and (H2).*

We now state several hypotheses that we will need in order to prove the condition (A3). The first one concerns the choice of the inverse Young functions with respect to the rate of convergence and the regularity of $H_\theta$. With $p$ as defined in (DRI), we suppose:

**(H1)** For any compact $\mathcal{K}$, there exist $q > 0$ and $s \geq p$ with $1/q + 1/s = 1$ such that:

$$\sum_{k \geq 0} \frac{1}{r_k^{1/q}} < \infty \qquad \text{and} \qquad \sup_{\theta \in \mathcal{K}} ||H_\theta||_{(\phi \circ V^p)^{1/s}} < \infty \,.$$

**Remark 5.3.3.** *We will show in section 5.5.3 that this hypothesis can be verified even for polynomial rates of convergence ($r_k = k^d$ with $d > 2$ in that example). This hypothesis can be seen as a compromise in the choice of $q$ and $s$ between the rate of convergence $r_k$ and the regularity of $H_\theta$. The assumption $s \geq p$ is necessary to control the $V$-norm by the $(\phi \circ V^p)^{1/s}$-norm.*

We then need hypotheses on the regularity of $H_\theta$ and $P_\theta$. Two of them are similar to the ones presented in Andrieu et al. (2005) while the first one will help us to conclude on the validity of Eq. (5.6).

**(H2)** For any compact $\mathcal{K}$, there exists a constant $\beta \in [0, 1]$ such that

**(i)** there exist $T_{\theta,\theta'} \in \mathbb{N}^*$ and $\alpha \in (0, 1)$ such that

$$\sup_{\theta,\theta' \in \mathcal{K}} T_{\theta,\theta'}||\theta - \theta'||^{\beta-\alpha} + ||\theta - \theta'||^{-\alpha} \sum_{k \geq T_{\theta,\theta'}} \frac{1}{r_k^{1/q}} < \infty \,.$$

**(ii)** there exists $C$ such that for all $x \in \mathcal{X}$,

$$\sup_{\theta,\theta' \in \mathcal{K}} ||\theta - \theta'||^{-\beta}|H_\theta(x) - H_{\theta'}(x)| \leq CV^p(x)$$

**(iii)** there exists $C$ such that for all $\theta, \theta' \in \mathcal{K}$,

$$||P_\theta g - P_{\theta'} g||_{(\phi \circ V^p)^{1/s}} \leq C||g||_{(\phi \circ V^p)^{1/s}}||\theta - \theta'||^\beta \qquad \forall g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}} \,.$$

**Remark 5.3.4.** *In the condition (H2-i), $T_{\theta,\theta'}$ is a positive integer. It implies in particular $\beta \geq \alpha$. This condition can be easily verified for $r_k^{1/q} = k^d$ with $d > 1$. Indeed, we know that $\sum_{k=T}^{\infty} \frac{1}{k^d} \sim \frac{1}{(d-1)T^{d-1}}$. Hence, if $0 < \alpha < 1$, we choose $T_{\theta,\theta'} = 1 \vee \left\lfloor ||\theta - \theta'||^{-\frac{\alpha}{d-1}} \right\rfloor$ and we have:*

$$||\theta - \theta'||^{-\alpha} \sum_{k=T_{\theta,\theta'}}^{\infty} \frac{1}{k^d} \sim_{\theta \to \theta'} \frac{1}{d-1} \,.$$

*Moreover, if* $||\theta-\theta'|| \leq 1$, $T_{\theta,\theta'}||\theta-\theta'||^{\beta-\alpha} = ||\theta-\theta'||^{\beta-\alpha-\frac{\alpha}{d-1}}$. *Choosing* $\alpha$ *such that* $\beta-\alpha-\frac{\alpha}{d-1} > 0$ *i.e.* $\alpha < \beta\frac{d-1}{d}$ *allows us to conclude.*

Finally, due to the subgeometric ergodicity, we are unable to iterate the drift condition without making divergent quantities appear. This iteration was however one of the keys of the proof of the condition 5.8. To overcome this problem, we add one last hypothesis on the behaviour of $\phi$ on the petite set $\mathcal{C}$ defined by assumption (DRI):

**(H3)** there exists $\delta > 0$ such that, $\forall x \in \mathcal{C}$,

$$\phi \circ V^p(x) \geq \delta V^p(x)\,.$$

**Remark 5.3.5.** *It is interesting to remark that asking for this condition on the whole set* $\mathcal{X}$ *implies the geometric ergodicity of the chain. However, we only ask it on the petite set* $\mathcal{C}$ *on which we have some freedom. In fact, in most cases, this condition will be easy to verify. Indeed, according to the theorem 16.1.9. of Douc et al. (2018), we can choose* $\mathcal{C} = \{V^p \leq d\}$ *with* $d > 0$. *Hence, if this set is compact (true if* $V$ *is continuous and* $V(x) \longrightarrow_{x\to\infty} \infty$*) and if* $(\phi \circ V^p)^{1/s}/V^p$ *is continuous, (H3) is verified.*

We can now state our major theorem:

**Theorem 5.3.1** *Assume (DRI) and (H1)-(H3). Then, the condition (A3) is verified. In particular, if (A1), (A2) and (A4) are also verified we can apply the theorem 5.2.1 to conclude that* $lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$

## 5.4 Proof of the theorem 5.3.1

### 5.4.1 Sketch of proof

The proof follows the principal ideas of Andrieu et al. (2005). However, due to the fact that our Markov chain is no longer supposed to be geometric ergodic, we need several new arguments. In particular, the behaviour of $\phi$ on the petite set $\mathcal{C}$ and the hypotheses on the rate of convergence $(r_k)_{k\in\mathbb{N}}$ will be of the upmost importance.

The first important result is the fact that we are able to dominate the $V$-norm by the $(\phi \circ V^p)^{1/s}$-norm under the hypothesis (H1). This is particularly important as we need to choose $W = V$ in (A3) to be able to find an upper bound of the expectation of $W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon \geq k}$ (see Eq. (5.8)). Hence, we use this control of the $V$-norm to control the different quantities in Eq. (5.4), (5.5) and (5.6) using the rate of convergence given by Eq (5.11). This control is given by the lemma 5.4.1.

Using this lemma, we can control the norm of the solution of the Poisson equation using the subgeometric ergodicity. This is explained lemma 5.4.2.

We then want to prove the condition (5.6) (lemma 5.4.5). Using once again a decomposition of the solution of the Poisson equation, we see that we need regularity conditions on $\theta \mapsto P_\theta$ and $h$. The regularity of $\theta \mapsto P_\theta$ is given by the condition (H2) while we prove the Hölder continuity of $h$ in lemma 5.4.4.

Finally, while the condition (5.7) is easily proved by iterating the drift condition, we still need to prove the condition (5.8). In Andrieu et al. (2005), the authors prove it using the same argument which does not hold anymore for us as this iteration can make appear divergent quantities. That is why we need to state the condition (H3). It is under this final condition that we are able to iterate an upper bound of the drift and prove (5.8) in lemma 5.4.6.

After this final step, we have all the tools necessary to prove the theorem 5.3.1.

We will now present and prove with details the different lemmas introduced above and implying each of the conditions in (A3) before proving the theorem 5.3.1.

### 5.4.2 Proof of Eq. (5.5)

First, using (H1), we show that we can control the $V$-norm using the $(\phi \circ V^p)^{1/s}$-norm:

**Lemma 5.4.1.** *Assume (H1). Then, there exists $C > 0$ such that, for all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$,*

$$||g||_V \leq C||g||_{(\phi \circ V^p)^{1/s}} .$$

*Proof.* $\phi$ is concave and increasing so, $\forall v \geq 1$, $\phi(v) \leq \phi'(1)(v-1) + \phi(1) \leq cv$ with $c$ a positive constant. Hence, for all $x \in \mathcal{X}$, since $s \geq p$ and $V(x) \geq 1$,

$$(\phi \circ V^p)^{1/s}(x) \leq c^{1/s} V^{p/s}(x) \leq c^{1/q} V(x)$$

which allows us to verify the announced inequality. □

We can now prove the equation (5.5).

**Lemma 5.4.2.** *Suppose (DRI). Then, the Poisson equation $g - P_\theta g = H_\theta - h(\theta)$ has a solution $g_\theta$. Moreover, under (H1),*

$$\sup_{\theta \in \mathcal{K}} ||g_\theta||_V < \infty \qquad \textit{and} \qquad \sup_{\theta \in \mathcal{K}} ||P_\theta g_\theta||_V < \infty .$$

*Proof.* The proposition [21.2.4] of Douc et al. (2018) states the existence of a solution $g_\theta$ of the Poisson equation under the subgeometric ergodicity conditions (DRI) verifying:

$$g_\theta(x) = \sum_{k \geq 0} \left( P_\theta^k H_\theta(x) - h(\theta) \right) .$$

Moreover, we know that for any compact $\mathcal{K}$, there exist a constant $C$ and a convergence rate $(r_k)_{k \in \mathbb{N}}$ independent of $\theta \in \mathcal{K}$ such that, for all $f \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$, for all $\theta \in \mathcal{K}$,

$$r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi \circ V^p)^{1/s}} \leq C||f||_{(\phi \circ V^p)^{1/s}} .$$

Hence, using lemma 5.4.1,

$$\begin{aligned} r_k^{1/q}||P_\theta^k f - \pi_\theta(f)||_V &\leq r_k^{1/q} C||P_\theta^k f - \pi_\theta(f)||_{(\phi \circ V^p)^{1/s}} \\ &\leq C||f||_{(\phi \circ V^p)^{1/s}} . \end{aligned}$$

Since $h(\theta) = \pi_\theta(H_\theta)$ and using (H1), we have that:

$$||g_\theta||_V \leq \sum_{k \geq 0} ||P_\theta^k H_\theta - h(\theta)||_V \leq C||H_\theta||_{(\phi \circ V^p)^{1/s}} \sum_{k \geq 0} \frac{1}{r_k^{1/q}} < \infty .$$

Finally, we can use the same argument for $P_\theta g_\theta$ to prove that $\sup_{\theta \in \mathcal{K}} ||P_\theta g_\theta||_V < \infty$. □

### 5.4.3 Proof of Eq. (5.6)

We now want to prove the condition given by Eq. (5.6). In particular, we need the hypotheses on the regularity in $\theta$ of $H_\theta$ and $P_\theta$ presented in condition (H2). We begin by proving two lemmas implying the Hölder continuity of $h$.

**Lemma 5.4.3.** *Assume (DRI), (H1) and (H2). Then, there exists a constant $C$ such that, for all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$ and any $k \geq 0$,*

$$\sup_{\theta, \theta' \in \mathcal{K}} ||\theta - \theta'||^{-\beta} ||P_\theta^k g - P_{\theta'}^k g||_{(\phi \circ V^p)^{1/s}} \leq C||g||_{(\phi \circ V^p)^{1/s}} .$$

*Proof.* This result is a consequence of (H2-iii). Indeed, we can write, for all $\theta$, $\theta'$ in $\mathcal{K}$, all $k \in \mathbb{N}$ and all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$,

$$P_\theta^k g - P_{\theta'}^k g = \sum_{j=0}^{k-1} P_\theta^j (P_\theta - P_{\theta'})(P_{\theta'}^{k-j-1} g(x) - \pi_{\theta'}(g)) .$$

But, using Eq. (5.11), we know that, for any $l \geq 0$,

$$\sup_{\theta \in \mathcal{K}} ||P_\theta^l - \pi_\theta||_{(\phi \circ V^p)^{1/s}} \leq \frac{C}{r_l^{1/q}} .$$

Hence, $\sup_{l \in \mathbb{N}, \theta \in \mathcal{K}} ||P_\theta^l||_{(\phi \circ V^p)^{1/s}} < \infty$.
Finally, using this result and (H2-iii),

$$\begin{aligned} ||P_\theta^k g - P_{\theta'}^k g||_{(\phi \circ V^p)^{1/s}} &\leq C||\theta - \theta'||^\beta \sum_{j=0}^{k-1} ||P_{\theta'}^{k-j-1} g(x) - \pi_{\theta'}(g)||_{(\phi \circ V^p)^{1/s}} \\ &\leq C||\theta - \theta'||^\beta ||g||_{(\phi \circ V^p)^{1/s}} \sum_{j=0}^{k-1} \frac{1}{r_{k-j-1}^{1/q}} . \end{aligned}$$

We obtain the result using the convergence of the sum of the $1/r_j^{1/q}$. □

We now prove that $h$ is $\beta$-Hölder. We will use this property to finally be able to prove (5.6).

**Lemma 5.4.4.** *Assume (DRI), (H1) and (H2). Then,*

$$\sup_{\theta, \theta' \in \mathcal{K}} ||\theta - \theta'||^{-\beta} |h(\theta) - h(\theta')| < \infty .$$

*Proof.* We use the following decomposition of $|h(\theta) - h(\theta')|$ for $x_0 \in \mathcal{X}$, $(\theta, \theta') \in \mathcal{K}^2$ and $k \in \mathbb{N}$:

$$|h(\theta) - h(\theta')| = |A(\theta, \theta') + B(\theta, \theta') + C(\theta, \theta')|$$

with:

$$\begin{aligned} A(\theta, \theta') &= h(\theta) - P_\theta^k H_\theta(x_0) + P_{\theta'}^k H_{\theta'}(x_0) - h(\theta') \\ B(\theta, \theta') &= P_\theta^k H_\theta(x_0) - P_{\theta'}^k H_\theta(x_0) \\ C(\theta, \theta') &= P_{\theta'}^k H_\theta(x_0) - P_{\theta'}^k H_{\theta'}(x_0)\,. \end{aligned}$$

From lemma 5.4.3, hypotheses (H2-ii) and (DRI), we obtain the following inequalities:

$$|A(\theta, \theta')| \leq \frac{C}{r_k^{1/q}} \sup_{\theta \in \mathcal{K}} ||H_\theta||_{(\phi \circ V^p)^{1/s}} (\phi \circ V^p)^{1/s}(x_0)$$

$$|B(\theta, \theta')| \leq C||H_\theta||_{(\phi \circ V^p)^{1/s}} ||\theta - \theta'||^\beta (\phi \circ V^p)^{1/s}(x_0)$$

$$\begin{aligned} |C(\theta, \theta')| &\leq \int_{\mathcal{X}} P_{\theta'}^k(x_0, dy) |H_\theta(y) - H_{\theta'}(y)| \\ &\leq C||\theta - \theta'||^\beta \int_{\mathcal{X}} P_{\theta'}^k(x_0, dy) V^p(y) \\ &\leq C||\theta - \theta'||^\beta V^p(x_0)\,. \end{aligned}$$

Hence, using the fact that $\sup_{\theta \in \mathcal{K}} ||H_\theta||_{(\phi \circ V^p)^{1/s}} < \infty$ and $(\phi \circ V^p)^{1/s} \leq c V^p$, we find

$$|h(\theta) - h(\theta')| \leq C V^p(x_0) \left( ||\theta - \theta'||^\beta + \frac{1}{r_k^{1/q}} \right).$$

Finally, because $\frac{1}{r_k^{1/q}} \to 0$, there exists $k \in \mathbb{N}$ such that $\frac{1}{r_k^{1/q}} < ||\theta - \theta'||^\beta$ which concludes the proof. □

Finally, we can state the condition (5.6).

**Lemma 5.4.5.** *Assume (DRI), (H1) and (H2). Then,*

$$\sup_{\theta, \theta' \in \mathcal{K}} ||\theta - \theta'||^{-\alpha} \left( ||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W \right) < \infty\,.$$

*Proof.* Using (H2-iii), lemmas 5.4.3 and 5.4.4, we have that, for $x \in \mathcal{X}$, $k \in \mathbb{N}$ and $\theta, \theta' \in \mathcal{K}$,

$$\begin{aligned} D_k(x, \theta, \theta') &:= |P_\theta^k H_\theta(x) - h(\theta) - P_{\theta'}^k H_{\theta'}(x) + h(\theta')| \\ &\leq |P_\theta^k H_\theta(x) - P_\theta^k H_{\theta'}(x)| + |P_{\theta'}^k H_{\theta'}(x) - P_\theta^k H_{\theta'}(x)| + |h(\theta) - h(\theta')| \\ &\leq C||\theta - \theta'||^\beta (\phi \circ V^p)^{1/s}(x) \end{aligned}$$

where we have used the fact that $(\phi \circ V^p)^{1/s}(x) \geq \phi(1) > 0$.

On the other hand, using the ergodicity of the Markov Chain (5.11) and (H1), there exists $c > 0$ such that

$$D_k(x, \theta, \theta') \leq \frac{c}{r_k^{1/q}} (\phi \circ V^p)^{1/s}(x)\,.$$

Hence for $t = 0$ or $1$ and any $T \geq t$ by splitting the sum at $k = T$ and using the two upper bounds found above, we have:

$$\begin{aligned}||\theta - \theta'||^{-\alpha}||P_\theta^t g_\theta - P_{\theta'}^t g_{\theta'}||_V &\leq C||\theta - \theta'||^{-\alpha}||P_\theta^t g_\theta - P_{\theta'}^t g_{\theta'}||_{(\phi\circ V^p)^{1/s}} \\ &\leq C||\theta - \theta'||^{-\alpha}\sum_{k\geq t}||D_k(.,\theta,\theta')||_{(\phi\circ V^p)^{1/s}} \\ &\leq C\left((T-t)||\theta-\theta'||^{\beta-\alpha} + ||\theta-\theta'||^{-\alpha}\sum_{k\geq T}\frac{1}{r_k^{1/q}}\right).\end{aligned}$$

Hence, we can use (H2-i) to conclude the proof. □

**Remark 5.4.1.** *Here, we have in fact proved that, under the hypotheses (DRI), (H1) and (H2), the solution of the Poisson equation is $\alpha$-Hölder.*

Finally, under (DRI), (H1) and (H2), we are able to prove the first item of (A3). We still have to prove the second and third item. The second item is easily proved using the drift condition:

$$\begin{aligned}\mathbb{E}_{x,\theta}^{\Delta}(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\geq k}) &\leq \mathbb{E}_{x,\theta}^{\Delta}\left[\mathbb{E}_{x,\theta}^{\Delta}(PV^p(X_{k-1})|\mathcal{F}_{k-1})\right] \\ &\leq \mathbb{E}_{x,\theta}^{\Delta}(V^p(X_{k-1})) + b \leq V^p(x) + kb\end{aligned}$$

and we conclude using the fact that for any $x \in \mathcal{X}$, $V^p(x) \geq 1$.

Hence, we only need to prove the last item of (A3).

### 5.4.4 Proof of Eq. (5.8)

Under geometrical ergodicity, iterating the drift condition is enough to prove the necessary inequality. However, in the subgeometric case, this iteration can make appear a divergent sum. To overcome this difficulty, we will use the condition (H3).

**Lemma 5.4.6.** *Assume (DRI) and (H3). Then, there exist a sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ and a constant $C$ such that for any sequence $\Delta$ and for any $x \in \mathcal{X}$,*

$$\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^{\Delta}[V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon\geq k}] \leq CV^p(X)\,.$$

*Proof.* Using (DRI) and (H3), we have that, for all $x \in \mathcal{X}$,

$$PV^p(X) \leq V^p(x) - \phi\circ V^p(x) + b\mathbb{1}_{\mathcal{C}}(x)\,.$$

Hence, if $x \notin \mathcal{C}$, $PV^p(x) \leq V^p(x)$ and, if $x \in \mathcal{C}$, $PV^p(x) \leq (1-\delta)V^p(x) + b$.

We first consider the case $\delta \geq 1$. In that case, if $x \in \mathcal{C}$, $PV^p(x) \leq b$. Hence, by induction, $\mathbb{E}_{x,\theta}^{\Delta}\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\right) \leq V^p(x) + b$.

If $\delta < 1$, we note $\tau_k = Card(X_i|X_i \in \mathcal{C}$ for $1 \leq i \leq k)$ the number of elements $(X_i)_{1\leq i\leq k}$ belonging to $\mathcal{C}$. Then,

$$\begin{aligned}\mathbb{E}_{x,\theta}^{\Delta}\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\right) &= \mathbb{E}_{x,\theta}^{\Delta}\left(\mathbb{E}_{x,\theta}^{\Delta}\left(PV^p(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\middle|\mathcal{F}_{k-1}\right)\right) \\ &\leq \mathbb{E}_{x,\theta}^{\Delta}\left((1-\delta\mathbb{1}_{X_{k-1}\in\mathcal{C}})V^p(X_{k-1}) + b\mathbb{1}_{X_{k-1}\in\mathcal{C}}\right)\end{aligned}$$

Hence, at each iteration $i \leq k-1$, if $X_i \in \mathcal{C}$, we multiply the expression by $(1-\delta)$ and add $b$. Such a case happens $\tau_{k-1}$ times. Otherwise, we keep the same expression as before, but at the rank $i-1$. By iterating, we have:

$$\begin{aligned}\mathbb{E}_{x,\theta}^{\Delta}\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\right) &\leq \mathbb{E}_{x,\theta}^{\Delta}\left((1-\delta)^{\tau_{k-1}}V^p(x) + b\sum_{i=0}^{\tau_{k-1}-1}(1-\delta)^i\right)\\ &\leq V^p(x) + \frac{b}{\delta}\,.\end{aligned}$$

Since $V^p(x) \geq 1$, we can conclude the proof. □

### 5.4.5 Proof of Theorem 5.3.1

We can now finalize this section by proving the theorem 5.3.1 using the lemmas previously presented.

*Proof.* Using lemma 5.4.1 and hypothesis (H1), we immediately obtain the first inequality in hypothesis (A3-i). The next two conditions are given respectively by 5.4.2 and 5.4.5. The last conditions are a consequence of lemma 5.4.6. □

## 5.5 Example: Symmetric Random Walk Metropolis Hastings (SRWMH)

### 5.5.1 Presentation of the algorithm

The SRWMH is a popular algorithm allowing for sampling from a distribution $\pi$. It consists in simulating a Markov Chain $(X_n)_{n\in\mathbb{N}}$ whose stationary distribution is $\pi$. The user chooses a symmetric proposal distribution $q$. At each step, if the chain is currently at $x$, a candidate $y$ for $X_{n+1}$ is proposed using $q(x-.)$. This candidate is then accepted with probability:

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} & \text{if } \pi(x) \neq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{5.12}$$

If the candidate is rejected, the chain stays at its current location $x$. The transition kernel of this Markov Chain is: $\forall x \in \mathcal{X}, \forall A \in \mathcal{B}(\mathcal{X})$,

$$P(x,A) = \int_A \alpha(x,y)q(x-y)\lambda^{Leb}(dy) + \mathbb{1}_A(x)\int_X (1-\alpha(x,y))q(x-y)\lambda^{Leb}(dy)\,. \tag{5.13}$$

The choice of the proposal distribution $q$ is of crucial importance. In particular, proposal distributions with a too small or too large covariance matrix lead to a highly correlated Markov Chain. To overcome this difficulty, the authors of Haario et al. (2001) have proposed to learn the covariance matrix while sampling the Markov Chain leading to adaptive MCMC samplers. We note $\theta = (\mu, \Gamma)$ and we suppose that we can choose $q_\theta$ such that $Var(q_\theta) = \Gamma$. For instance, if we choose to work with Gaussian distributions, $q_\theta$ is the density of the distribution $\mathcal{N}(0,\Gamma)$. We then write $P_\theta$ the kernel of the SRWMH when the proposal is $q_\theta$.

We can then adapt the value of $\Gamma$ using the following algorithm:

$$\begin{cases} \mu_{n+1} = \mu_n + \Delta_{n+1}(X_{n+1} - \mu_n) \\ \Gamma_{n+1} = \Gamma_n + \Delta_{n+1}\left((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)^T - \Gamma_n\right) \end{cases} \tag{5.14}$$

with $X_{n+1} \sim P_{\theta_n}(X_n, .)$ where $\theta_n = (\mu_n, \Gamma_n)$ and with $(\Delta_n)_{n\in\mathbb{N}}$ a non-increasing sequence of step sizes such that $\sum_{n=1}^{\infty} \Delta_n = \infty$ and, for some $b > 0$, $\sum_{n=1}^{\infty} \Delta_n^{1+b} < \infty$.

This procedure is in fact a stochastic approximation:

$$\theta_{n+1} = \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1})$$

with

$$H_\theta(x) = (x - \mu, (x - \mu)(x - \mu)^T - \Gamma)\,. \tag{5.15}$$

Moreover, assuming that $\int_{\mathcal{X}} x^2 \pi(dx) < \infty$, one can verify that:

$$h(\theta) = \left(\mu_\pi - \mu, (\mu_\pi - \mu)(\mu_\pi - \mu)^T + \Gamma_\pi - \Gamma\right)$$

with $\mu_\pi$ and $\Gamma_\pi$ respectively the mean and variance of $\pi$.

This algorithm has already been studied in Andrieu et al. (2005). In that paper, the authors make a hypothesis on the tail properties of the target distribution that implies the geometric ergodicity of the Markov Chain $P_\theta$. Under this hypothesis, the authors prove that the conditions (A1)-(A4) are verified and so prove the convergence of the algorithm.
Within our framework, we are able to loosen the hypothesis on $\pi$ to give conditions under which we have a subgeometric ergodicity of the Markov Chain $P_\theta$ while still guaranteeing convergence of the algorithm.

In Andrieu et al. (2005), the verification of the condition (A1) does not use the behaviour of the tail of $\pi$. Hence, it will stay true in our case and we can state it here:

**Proposition 5.5.1** *Let*

$$w(\mu, \Gamma) = -\int_{\mathcal{X}} \log\left(\frac{\pi(x)}{\phi_{\mu,\Gamma}(x)}\right) \pi(dx)$$

*where $\phi_{\mu,\Gamma}$ is the normal density of mean $\mu$ and variance $\Gamma$. Then, $w$ verifies (A1). Furthermore, $\mathcal{L}$ is reduced to a single point $\theta_\pi := (\mu_\pi, \Gamma_\pi)$.*

To prove (A3), we need some hypotheses on the behaviour of $\pi$. In particular, we will verify that we can apply the theorem 5.3.1 under two different sets of hypotheses. The first contains among others the Weibull distributions while the second one includes the Pareto distributions. Those two sets of hypotheses as well as the proof of the condition (A3) are detailed in the following subsections.

### 5.5.2 First family of distributions (including the Weibull one) satisfying our assumptions

In Douc et al. (2004) and Fort and Moulines (2003), the authors present a set of hypotheses on the target and proposal distributions that imply the subgeometric ergodicity of the Markov Chain. The first hypothesis concerns the target distribution:

**(E1)** The target density $\pi$ is continuous and positive on $\mathbb{R}^d$ and there exist $m \in (0,1)$, $r \in (0,1)$, positive constants $d_i$, $D_i$, $i = 0,1,2$ and $R_0 < \infty$ such that, if $|x| \geq R_0$, $x \mapsto \pi(x)$ is twice continuously differentiable and

$$\left\langle \frac{\nabla \pi(x)}{|\nabla \pi(x)|}, \frac{x}{|x|} \right\rangle \leq -r$$

$$d_0|x|^m \leq -\ln \pi(x) \leq D_0|x|^m$$
$$d_1|x|^{m-1} \leq |\nabla \ln \pi(x)| \leq D_1|x|^{m-1}$$
$$d_2|x|^{m-2} \leq |\nabla^2 \ln \pi(x)| \leq D_2|x|^{m-2} .$$

Among others, the Weibull distribution on $\mathbb{R}_+$ $\pi : x \mapsto \beta\eta x^{\eta-1}\exp(-\beta x^\eta)$ with $\beta > 0$ and $\eta \in (0,1)$ verifies those conditions.
We also need some conditions on the proposal distribution:

**(E2)** There exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. Moreover, $q_\theta$ is symmetric, bounded away from zero in a neighborhood of zero, and is compactly supported. We also assume that there exist $C > 0$ and $\beta \in (0,1)$ such that for all $(\theta, \theta') \in \Theta^2$,

$$\int_X |q_\theta(z) - q_{\theta'}(z)| \lambda^{Leb}(dz) \leq C|\theta - \theta'|^\beta .$$

**Remark 5.5.1.** *The compactly supported condition could be relaxed with appropriate moment conditions.*

We can now prove the following theorem:

**Theorem 5.5.1** *Let $\pi$ and $q_\theta$ be distributions satisfying (E1) and (E2) and consider the process defined in* (5.14) *with $\varepsilon$ and $\Delta$ two sequences verifying (A4). Then, (A1), (A2) and (A3) are verified. Moreover, $\theta_n \to \theta_\pi$ w.p. 1 where $\theta_\pi := (\mu_\pi, \Gamma_\pi)$ is the unique stationary point of $(\theta_n)_{n\in\mathbb{N}}$.*

*Proof.* According to the theorem 3.1 of Douc et al. (2004), if (E1) and (E2) are satisfied, there exists $\xi_0$ such that for all $\xi \leq \xi_0$, there exist $c > 0$, $W := \pi^{-\xi}$ and $\phi : x \mapsto cx(1 + \ln(x))^{-2\frac{1-m}{m}}$ verifying:

$$PW + \phi \circ W \leq W + b\mathbb{1}_C .$$

Hence, we have a subgeometric drift condition. It is then possible to compute the associated rate of convergence: $r_k = \exp(ck^{\frac{m}{2-m}})$.

As stated in proposition 5.5.1, the condition (A1) is verified and (A2) is satisfied using the theorem 2.2 of Roberts and Tweedie (1996).
We will prove (A3) using the theorem 5.3.1.

First, the condition (DRI) is verified with $V^2 = \pi^{-\xi}$ and $p = 2$. Indeed, the drift condition is given above while the existence of small sets is ensured given the continuity of $\pi$ and hypothesis (E2) (see Theorem 2.2 of Roberts and Tweedie (1996)).

We then verify the hypothesis (H1). Given the value of $r_k$, the sum of the $r_k^{1/q}$ will be finite for any $q > 0$. Moreover, $\sup_{\theta\in\Theta} ||H_\theta||_{(\phi\circ V^2)^{1/s}} < \infty$ if and only if $x^2\pi^{\xi/s}(x)(1-\xi\ln\pi(x))^{\frac{2(1-m)}{sm}} < \infty$.

This will be true for any $s > 0$ as $\pi(x) \leq \exp(-D_0 x^m)$.

Concerning (H2), as discussed in remark 5.3.4, (H2-i) is verified for polynomial rates of convergence $k^d$ with $d > q$. Using the fact that $r_k^{1/q} > k^d$ for $k$ big enough, we can conclude that (H2-i) is verified in this case.

To verify (H2-ii), we remark that

$$|H_\theta(x) - H_{\theta'}(x)| \leq |\mu - \mu'|(1 + |\mu + \mu'| + 2|x|) + |\Gamma - \Gamma'|\,.$$

Since $||x||_{V^2} < \infty$, we obtain the inequality (H2-ii) for any $\beta \leq 1$.
We now interest ourselves in (H2-iii). Using the definition of the kernel $P_\theta$, we have that

$$\begin{aligned}
|P_\theta g(x) - P_{\theta'} g(x)| &\leq \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|g(x+z)\lambda^{Leb}(dz) \\
&\quad + g(x)\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz) \\
&\leq ||g||_{(\phi\circ V^2)^{1/s}}(\phi \circ V^2)^{1/s}(x)\Big(\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\frac{(\phi \circ V^2)^{1/s}(x+z)}{(\phi \circ V^2)^{1/s}(x)}\lambda^{Leb}(dz) \\
&\quad + \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\Big)\,.
\end{aligned}$$

Hence, writing $\Psi := (\phi \circ V^2)^{1/s}$, we need to study:

$$\alpha(x, x+z)\frac{\Psi(x+z)}{\Psi(x)} = \left(1 \wedge \frac{\pi(x+z)}{\pi(x)}\right)\frac{\pi^{-\xi}(x+z)(1 - \xi \ln \pi(x+z))^{-\frac{2(1-m)}{m}}}{\pi^{-\xi}(x)(1 - \xi \ln \pi(x))^{-\frac{2(1-m)}{m}}}\,.$$

But, if $\pi(x+z) \geq \pi(x)$, this function is always less than $1$.

If $\pi(x+z) \leq \pi(x)$, we use the growth of the function $\Phi(u) = u^{1-\xi}(1 - \xi \ln(u))^{-\frac{2(1-m)}{m}}$ for $u$ in a compact and $\xi$ small enough. Hence, we deduce once again that the function is less than $1$.

Finally,

$$|P_\theta g(x) - P_{\theta'} g(x)| \leq 2||g||_{(\phi\circ V^2)^{1/s}}(\phi \circ V^2)^{1/s}(x)\int_X |q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\,.$$

Hence, the hypothesis (E2) allows us to conclude on the validity of (H2-iii).

Finally, we just have the hypothesis (H3) to prove. According to the theorem 16.1.9 of Douc et al. (2018), $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. But, $V^2$ converges towards infinity at infinity and is continuous so, $\mathcal{C}$ is compact. Hence, because $\frac{\phi \circ V^2}{V^2}$ is continuous, there exists a lower bound of $\frac{\phi \circ V^2}{V^2}$ on $\mathcal{C}$ and (H3) is verified.

All the hypotheses of the theorem 5.3.1 are thus verified and we can apply it to conclude. □

Hence, we have proven the convergence of the Metropolis Hastings algorithm under a subgeometric ergodicity condition. In the next subsection we will interest ourselves in the case where the rate of convergence is not only subgeometric but polynomial and, once again, prove the convergence of a stochastic approximation.

### 5.5.3 Second usual family (including the Pareto distribution) covered by our framework

In Fort and Moulines (2003), the authors give other conditions on the target density for the SRWMH kernel to be subgeometric ergodic when we work in $\mathbb{R}$:

**(E3)** $\pi$ is continuous on $\mathbb{R}$ and there exist some finite constants $\alpha > 1$, $M > 0$, $C > 0$ and a function $\rho : \mathbb{R} \to [0, \infty)$ verifying $\lim_{x\to\infty} \rho(x) = 0$ such that for all $|x| > M$, $\pi$ is strictly decreasing and, for all $y \in \{z \in \mathbb{R} \,|\, \pi(x+z) \leq \pi(x)\}$,

$$\left| \frac{\pi(x+y)}{\pi(x)} - 1 + \alpha y x^{-1} \right| \leq C|x|^{-1}\rho(x)y^2 \,.$$

This class of distributions contains in particular the Pareto distributions ($\pi(x) \propto x^{-\alpha}$) as well as many heavy tail distributions. We also need some hypotheses on our proposal:

**(E4)** There exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. Moreover, $q_\theta$ is symmetric and there exists $\xi \geq 1$ such that $\int |y|^{\xi+3} q_\theta(y)dy < \infty$.

Under those conditions, we can state the following proposition, proved in Fort and Moulines (2003).

**Proposition 5.5.2** *Assume (E3) and (E4). Set $u = \xi \wedge \alpha + 1$ and $W : x \mapsto 1 + |x|^u$. Then, there exist $c > 0$ and a small set $\mathcal{C}$ such that, if we set $\phi : x \mapsto cx^{1-2/u}$,*

$$P_\theta W(x) + \phi \circ W(x) \leq W(x) + b\mathbb{1}_{\mathcal{C}} \,.$$

Under such a drift condition, we are able to deduce the rate of convergence using the value of $\phi$ (Douc et al., 2004): for all $k \in \mathbb{N}$, $r_k \propto k^{u/2-1}$.

**Theorem 5.5.2** *Let $\pi$ and $q_\theta$ be distributions on $\mathbb{R}$ satisfying (E3) and (E4) with $\xi \wedge \alpha > 5$ and consider the model defined in* (5.14) *with $\varepsilon$ and $\Delta$ two sequences verifying (A4). Assume also that (H2-iii) is verified. Then, (A1), (A2) and (A3) are verified. Moreover, $\theta_n \to \theta_\pi$ w.p. 1 where $\theta_\pi := (\mu_\pi, \Gamma_\pi)$ is the unique stationary point of $(\theta_n)_{n\in\mathbb{N}}$.*

**Remark 5.5.2.** *In this theorem, we suppose that (H2-iii) is verified. This condition depends on the function $\pi$. Given the functions $V$ and $\phi$ chosen here, we need, $\forall x, z \in \mathbb{R}$,*

$$\begin{cases} \pi(x+z) \leq \pi(x) & \implies \quad \frac{\pi(x+z)}{\pi(x)} \left( \frac{1+|x+z|^u}{1+|x|^u} \right)^{\frac{u-2}{us}} \leq C \\ \pi(x+z) \geq \pi(x) & \implies \quad |x+z| \leq C|x| \,. \end{cases} \tag{5.16}$$

*Other conditions can appear if $V$ or $\phi$ have another form. It was the case in the previous subsection where we have been able to prove this condition under the hypotheses (E1) and (E2). We prove this particular condition (H2-iii) in the next section in the case of the Pareto distribution.*

*Proof.* (A1) is stated in proposition 5.5.1.
Under (E3) and (E4), $P_\theta$ is $\psi$-irreducible (see theorem 2.2 of Roberts and Tweedie (1996)).

Hence, we have existence and uniqueness of the invariant distribution $\pi_\theta$. Moreover, $H$ is measurable. Hence, (A2) is verified.
We still need to verify (A3). To do so, we will use the theorem 5.3.1 and prove the hypotheses (DRI) and (H1)-(H3).
The proposition 5.5.2 and the theorem 2.2 of Roberts and Tweedie (1996) give us the validity of (DRI) with $p = 2$ and $W = V^2$.

We now prove (H1). First, $\sum_{k\geq 0} \frac{1}{r_k^{1/q}}$ is finite for any $q < \frac{u-2}{2}$. Moreover, recalling that $1/s + 1/q = 1$, that $(\phi \circ V^p)^{1/s} = (1 + |x|^u)^{\frac{u-2}{us}}$ and that $H_\theta$ is quadratic, for any $\mathcal{K}$ compact of $\mathbb{R} \times \mathbb{R}_+^*$, $\sup_{\theta\in\mathcal{K}} ||H_\theta||_{(\phi\circ V^2)^{1/s}} < \infty$ if and only if $q > \frac{u-2}{u-4}$. Hence, we need to choose $q$ such that:

$$\frac{u-2}{u-4} < q < \frac{u-2}{2}\,. \tag{5.17}$$

Since $u > 6$, such a $q$ exists. Moreover, because $\frac{u-2}{2} > 2 = p$, we can also choose $s > p$. Hence, the condition (H1) is verified.

Concerning (H2), as discussed in remark 5.3.4, (H2-i) is verified if $\frac{u/2-1}{q} > 1$ which is true given Eq. (5.17).
Concerning (H2-ii), we have that

$$|H_\theta(x) - H_{\theta'}(x)| \leq |\mu - \mu'|(1 + |\mu + \mu'| + 2|x|) + |\Gamma - \Gamma'|\,.$$

Since $||x||_{V^2} < \infty$ because $u \geq 1$, we obtain the inequality (H2-ii) for any $\beta \leq 1$.

Hence, we only have to prove (H3) to conclude. According to the theorem 16.1.9 of Douc et al. (2018), $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. In particular, since $V^2(x) = 1 + |x|^u$, there exists $d_1 > 0$ such that $\{V \leq d\} = [0, d_1]$. But, $x \mapsto \frac{(\phi\circ V^2)^{1/s}(x)}{V^2(x)}$ is continuous hence, bounded on the compact $[0, d_1]$. Thus, (H3) is verified.

□

We have proved the convergence of the Metropolis Hastings algorithm under a set of hypotheses implying a polynomial rate of convergence. In the next section, we show that those hypotheses are verified for the Pareto distribution with a scale parameter more than 5.

### 5.5.4 Application to the Pareto distribution

In this application, we choose to study the case where the target distribution $\pi$ is a Pareto distribution and the proposal $q_\theta$ is a normal distribution $\mathcal{N}(0, \Gamma)$. As showed in Fort and Moulines (2003), the Pareto distribution $\pi(x) \propto |x|^{-\alpha}$ verifies the condition (E3). Moreover, (E4) is satisfied for any $\xi > 0$. Hence, when applying the theorem 5.5.2, we only need to assume $\alpha \wedge \xi > 5$ i.e. $\alpha > 5$.

We now show that the Pareto distribution verifies the condition (H2-iii):

**Lemma 5.5.1.** *Suppose that $\pi$ is a Pareto distribution with shape $\alpha > 5$ and, for $\theta = (\mu, \Gamma)$, $q_\theta$ is the normal distribution $\mathcal{N}(0, \Gamma)$. Then, if $P_\theta$ is the kernel defined in* (5.13) *and $\mathcal{K}$ is a compact of $\mathbb{R}_+^*$, there exists $C$ such that for all $\theta, \theta' \in \mathcal{K}$ and for all $g \in \mathcal{L}_{(\phi\circ V^p)^{1/s}}$*

$$||P_\theta g - P_{\theta'} g||_{(\phi\circ V^p)^{1/s}} \leq C||g||_{(\phi\circ V^p)^{1/s}}|\theta - \theta'|^\beta\,.$$

*Proof.* As done in the proof of the theorem 5.5.1, writing $\Psi = (\phi \circ V^p)^{1/s}$, we need to find an upper bound to:

$$\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\frac{\Psi(x+z)}{\Psi(x)}\lambda^{Leb}(dz)$$
$$= \int_X \left(1 \wedge \frac{|x|^\alpha}{|x+z|^\alpha}\right) \frac{(1+|x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1+|x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\,.$$

But, if $|x+z|^\alpha \leq |x|^\alpha$,

$$\frac{(1+|x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1+|x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}} \leq 1\,.$$

Similarly, if $|x+z|^\alpha \geq |x|^\alpha$, using Eq. (5.17), we have that $s > 1 \geq \frac{\alpha-1}{\alpha}$. Hence,

$$\frac{|x|^\alpha}{|x+z|^\alpha}\frac{(1+|x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1+|x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}} \leq \left|1+\frac{z}{x}\right|^{-\alpha}\left(1+\left|1+\frac{z}{x}\right|^{\alpha+1}\right)^{\frac{\alpha-1}{s(\alpha+1)}}$$

is bounded since $u \mapsto u^{-\alpha}(1+u^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}$ is bounded on $[1, +\infty)$.

Finally, there exists $C > 0$ such that:

$$|P_\theta g(x) - P_{\theta'} g(x)| \leq C||g||_{(\phi\circ V^p)^{1/s}}(\phi \circ V^p)^{1/s}(x)\int_X |q_\theta(z) - q_{\theta'}(z)|dz\,.$$

But it has already been proved in Andrieu et al. (2005) that, if $q_\theta$ is the normal distribution of variance $\Gamma$ then, for any $\Gamma, \Gamma'$ in a compact subset $\mathcal{K}$ of $\mathbb{R}_+^*$,

$$\int_{\mathbb{R}} |q_\theta(z) - q_{\theta'}(z)|dz \leq \frac{1}{\Gamma_{\min}}|\Gamma - \Gamma'|$$

where $\Gamma_{\min}$ is the minimum value of $\mathcal{K}$ which allows us to conclude for any $\beta \leq 1$.

□

**Theorem 5.5.3** *Suppose that $\pi$ is a Pareto distribution with shape $\alpha > 5$ and, for $\theta = (\mu, \Gamma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$, $q_\theta$ is the normal distribution $\mathcal{N}(0, \Gamma)$. Let $(Z_n)_{n\in\mathbb{N}}$ be the Markov chain as described in 5.3 with $P_\theta$ defined in* (5.13) *and $H$ defined in* (5.15)*. Suppose that $(\Delta_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ are two sequences verifying (A4). Then, $\theta_n \to \theta_\pi = (\mu_\pi, \theta_\pi)$ w.p. 1.*

*Proof.* It is a consequence of the theorem 5.5.2 and lemma 5.5.1. All the conditions have already been proved. □

Thus, we have been able to prove the convergence of an adaptive MCMC algorithm targeting distributions for which the theorem proved in Andrieu et al. (2005) was not enough to conclude.

## 5.6 Application to Independent Component Analysis

Independent component analysis (ICA) is a method which aims at representing a data set of random vectors as linear combinations of a fixed family of vectors with independent random weights. ICA follows somehow the same goal as the Principal Component Analysis (PCA). However, PCA imposes orthogonality between principal components which amounts to supposing that the observed vectors follow a Normal distribution. As for the ICA, it assumes a more general statistical model where the observations are decomposed on components weighted by independent random coefficients. It is sometimes called source separation. ICA has a large range of applications in medical image analysis (Calhoun et al., 2001b,a), computer vision (Bartlett et al., 2002; Bell and Sejnowski, 1995; Liu and Wechsler, 2003), computational biology (Liebermeister, 2002; Makeig et al., 1997), etc.. This method is also used to map the data set onto a smaller space (not orthogonal) as one can choose the number of components in the linear combination.

This method writes an observation $X \in \mathbb{R}^d$ as:

$$X = \sum_{j=1}^{p} \beta_j a_j + \varepsilon = A\beta + \varepsilon \,, \tag{5.18}$$

where $A := (a_1, ..., a_p) \in \mathbb{R}^{d \times p}$ is a parameter, $(\beta_1, ..., \beta_p)$ are independent scalars whose law $q_m$ must be specified and $\varepsilon$ is the additive noise.

In a lot of cases, $\varepsilon$ is supposed to follow a normal distribution. This approximation enables to develop easily many estimation algorithms. However, numerical images are rather affected by a positive valued noise (MRI images for instance). Moreover, the Gaussian assumption reduces the study to very rapidly decreasing noise. In this example, to take into account these two bottlenecks of the Gaussian noise, we choose to model our data with a positive noise with heavy tail: the Weibull distribution.

We suppose that each coordinate of $\varepsilon$ satisfies: $\varepsilon_j \sim \mathcal{W}(\lambda_0, \eta_0)$ with $\lambda_0 \in \mathbb{R}_+^*$ and $\eta_0 \in (0, 1)$.

To estimate $A$, we will use a Monte Carlo Markov Chain - Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm introduced in Kuhn and Lavielle (2004). For this algorithm to converge, we need our joint distribution to belong to the curved exponential family i.e. to be of the following form:

$$q(X, \beta, A) = \phi(A) + \langle S(X, \beta), \psi(A) \rangle \,,$$

where $S(x, \beta)$ is called the sufficient statistic of the model.

However, it can be seen that the joint likelihood does not verify this hypothesis here. A usual work around, first introduced in Kuhn and Lavielle (2005), is to consider that all vectors of $A$: $(a_j)_{1 \leq j \leq p}$ are random vectors following a Gaussian prior. The goal is then to estimate the mean of this prior. This writes, for each vector $a_j$: $a_j \sim \mathcal{N}(a_{0,j}, \sigma_A^2 Id)$.

If $X$ is a data set of $n$ observations $(X^1, ..., X^n)$, we finally have, writing $A_0 = (a_{0,1}, ..., a_{0,p}) \in \mathbb{R}^{d \times p}$:

$$\begin{aligned}\log q(X,\beta,A,A_0) = &\sum_{i=1}^{n}\sum_{j=1}^{p}\left((\eta_0-1)\log(X_j^i-(A\beta^i)_j)-\left(\frac{X_j^i-(A\beta^i)_j}{\lambda_0}\right)^{\eta_0}\right)\\ &+\sum_{i=1}^{n}q_m(\beta^i)-\frac{||A-A_0||^2}{\sigma_A^2}+C\end{aligned} \tag{5.19}$$

where $||A-A_0||^2=\sum_{j=1}^{p}||a_j-a_{0,j}||_2^2$.

The joint distribution now belongs to the curved exponential family. Indeed, it can be written as:

$$\log q(X,\beta,A,A_0)=\phi(A_0)+<S(X,\beta,A),\psi(A_0)>+\tilde{S}(X,\beta,A)\tilde{\phi}(A_0)$$

with:

$$\begin{cases}\phi(A_0)=\frac{||A_0||^2}{\sigma_A^2}+C\\ S(X,\beta,A)=A\\ \psi(A_0)=-2A_0\\ \tilde{S}(X,\beta,A)=\sum_{i=1}^{n}\sum_{j=1}^{p}\left((\eta_0-1)\log(X_j^i-(A\beta^i)_j)-\left(\frac{X_j^i-(A\beta^i)_j}{\lambda_0}\right)^{\eta_0}\right)\\ \qquad\qquad +\sum_{i=1}^{n}q_m(\beta^i)+\frac{||A||^2}{\sigma_A^2}\\ \tilde{\psi}(A_0)=1\end{cases}$$

The maximum of the log-likelihood can then be expressed as a function of the sufficient statistics: the maximum of $q(X,\beta,A,\theta)$ is reached for $A_0=\hat{\theta}(S(X,\beta,A))=A$.

Then, the MCMC-SAEM algorithm consists in the following steps:
(i) Simulation of $\beta, A$ using a Metropolis Hastings algorithm targeting the conditional distribution $q(\beta,A|X,\theta_{k-1})$.
(ii) Stochastic approximation of the sufficient statistics:

$$S_k=S_{k-1}+\Delta_k(A-S_{k-1})\,.$$

(iii) Maximization of the conditional distribution using the sufficient statistics: $\theta_k=\hat{\theta}(S_k)$.

**Remark 5.6.1.** *A fourth step not indicated above for clarity is the truncation process executed as described in section 5.2.2 and allowing our parameters to stay on compact sets.*

We can easily see that the described procedure is a particular case of the theorem 5.2.1 with $P_s$ the kernel of the Metropolis Hastings algorithm targeting $q(\beta,A|X,\hat{\theta}(s))$ and with

$$H_s(\beta,A)=S(X,\beta,A)-s\,.$$

This problem has been tackled for instance in Allassonniere et al. (2012). In that paper, the authors propose several distributions for $\beta$ leading to geometrically ergodic Markov Chains.

Using theorem 5.3.1, we are now able to tackle distributions leading to subgeometric ergodic chains which enables to introduce models with higher variability. We provide here an example of such a chain and prove convergence of the associated ICA parameters.

In the following, we suppose that all coordinates of $\beta$ follow a Weibull distribution: $\forall i \in [|1,n|], \forall j \in [|1,p|], \beta_j^i \sim \mathcal{W}(\lambda_1, \eta_1)$ with $\lambda_1 \in \mathbb{R}_+^*$ and $\eta_1 \in (0,1)$. Other distributions with heavy tails such as the Pareto distribution would yield to similar results.

**Theorem 5.6.1** *Assume (A4), (A1i) and that the proposal distribution in the Metropolis Hastings algorithm verifies (E2). Define* $l(\theta) = \log \int q(X, \beta, A, \theta) d\beta dA$ *and* $\mathcal{L}' = \{\theta \in \hat{\theta}(\mathcal{S}) | \partial_\theta l(\theta)) = 0\}$. *We then have* $d(\theta_k, \mathcal{L}') \to 0$.

**Remark 5.6.2.** *Most of the work has in fact already been done in section 5.5.2. Indeed, the proof of the hypothesis (A3) follows the exact same steps as in 5.5.2 and thus will not be detailed here. Note that Condition (A1i) remains an assumption of the theorem as in many cases.*

*Proof.* We first check the conditions (A1) (ii), (iii) and (iv). Let $w(s) = -l(\hat{\theta}(s))$. As showed in Delyon et al. (1999), this function verifies (A1) (iii) and (iv). Moreover, the authors prove that $\mathcal{L} = \mathcal{L}'$.
It is then easy to verify (A1)(ii) by remarking that $w(s) \to_{||s|| \to \infty} \infty$. Since $w$ is continuous, (A1)(ii) is verified for any $M_1 > 0$.

Concerning (A2), the theorem 2.2 of Roberts and Tweedie (1996) gives the $\psi$-irreducibility of the Markov Chain and thus the existence of the unique stationary distribution $\pi_\theta$. The measurability of $H_\theta$ is immediate.

We can easily verify that (E1) is true for $m = \eta_0 \vee \eta_1$. Hence, we can follow the exact same proof as in theorem 5.5.1 to prove that (H1), (H2) and (H3) are verified and thus the condition (A3) by theorem 5.3.1.

(A4) being supposed, we can apply the theorem 5.2.1 to conclude the proof.

□

Hence, this simple example shows that the algorithm can be applied not only on simulation algorithms but also on optimization algorithms such as Expectation Maximization or stochastic gradient which are involved in many machine learning and deep learning methods.

## 5.7 Conclusion

In this paper, we relaxed the condition of geometric ergodicity previously needed to ensure the convergence of stochastic approximations with Markovian dynamics. We provide therefore theoretical guarantees for a wider class of algorithms that are used in practice.

Our main result proves the convergence of these stochastic approximations for Markov Chains which are only subgeometric ergodic assuming hypotheses on the rate of convergence and the drift condition. A corollary is the convergence of a Metropolis Hastings algorithm with adapted variance, first in the case of the Weibull distribution with a shape parameter between $0$ and $1$ and then in the case of the Pareto distribution with a shape parameter more than $5$.

Another corollary applies to the convergence of a Stochastic Approximation Expectation Maximization algorithm when subgeometric Markov Chains appear. These results suggest that the main theorem could be used to show the convergence of a broader range of algorithms for which the geometric ergodicity is not verified.

# CHAPTER 6

# On the curved exponential family in the Stochastic Approximation Expectation Maximization Algorithm

The Expectation-Maximization Algorithm (EM) is a widely used method allowing to estimate the maximum likelihood of models involving latent variables. When the Expectation step cannot be computed easily, one can use stochastic versions of the EM such as the Stochastic Approximation EM.
In the chapter 3, we used this particular algorithm to estimate the parameters of our model. To do so, we had to take into account that the SAEM convergence is only ensured when the model is curved exponential. It forced us to rewrite the model by considering the population parameters $z_{pop}$ as random variables: we supposed they follow a Gaussian distribution with variance $\sigma^2$ centered on a new parameter $\bar{z}_{pop}$ to estimate. Indeed, without this change, first introduced in Kuhn and Lavielle (2005), the initial model would not be curved exponential. In practice, there is however no guarentee that this change of model and the choice of $\sigma$ will not influence the estimation of the parameters.
In this chapter, we show that this transformation of the model can indeed introduce a bias, that we quantify, in the final estimation of the parameters. In particular, we show that a trade-off must be made between the speed of convergence and the tolerated error. Finally, we propose a new algorithm achieving a better estimation of the maximum of likelihood of the initial model in a reasonable computation time.

This work has been submitted.

## Contents

## 6.1 Introduction

With the increase of data, parametric statistical models have become a crucial tool for data analysis and understanding. To be able to describe complex natural phenomena (epidemiology, ecology, finance, disease evolution, etc.), the models have an increasing complexity. Some of them are based on observed features or data which are assumed to be generated from a latent random effect. A usual example is the family of mixed effects models which have been used in pharmacokinetic, pharmacodynamic, shape analysis, etc. In such a context, one aims at optimizing the model parameter to maximize the likelihood of the observed dataset. This likelihood is also called the incomplete one as the latent variables are unknown.
Formally, this writes as follow: let $y \in \mathbb{R}^n$ be the observation and $\theta \in \Theta$ the model parameter. We call g the incomplete likelihood:

$$g(y, \theta) = \int_Z f(y, z, \theta)\, dz\,.$$

In that case, $z$ is the latent or missing variable and $f$ is the joint likelihood of the observations and latent variables, depending on a parameter $\theta \in \Theta$.
The Expectation Maximization (EM) algorithm provides a numerical process to answer this problem by computing iteratively a sequence of estimates $(\theta_n)_{n\in\mathbb{N}}$ which, under several conditions (see Dempster et al. (1977); Wu (1983)), converges towards the maximum likelihood estimate. It proceeds in two steps for each iteration $k$. First, in the Expectation step (E), the function

$$Q(\theta|\theta_{k-1}) = \int_Z \log(f(y, z, \theta))p(y, z, \theta_{k-1})\, dz$$

is computed where $p$ is the conditional distribution of $z$ given the observations: $p(y, z, \theta) = f(y, z, \theta)/g(y, \theta)$. $\theta_k$ is then updated in the Maximization step (M) as the argument of the maximum of the function $Q(.|\theta_{k-1})$.

The EM algorithm has been first introduced in Dempster et al. (1977). Its properties have then been studied in numerous papers, see Balakrishnan et al. (2017); Chrétien and Hero (2008); Ma et al. (2000); Meng et al. (1994); Redner and Walker (1984); Tseng (2004); Wu (1983) among many other works.

In many cases, the (E) step is in fact intractable as we have no closed form for $Q$. Different algorithms, both deterministic and stochastic, have been introduced in the literature to overcome this problem. The Monte-Carlo EM (Wei and Tanner (1990)) replaces the (E) step by computing a Monte Carlo approximation of $Q$ using a large amount of simulated missing data $z$. Another possibility, more computationally efficient, is to use a Stochastic Approximation (SA) of the function $Q$. This SAEM algorithm has been introduced in Delyon et al. (1999) and the authors proved the convergence towards a local maximum of the incomplete likelihood with probability 1 under several hypotheses. It has later on been generalized in Kuhn and Lavielle (2004) in the case where we are not able to easily sample $z$. This new algorithm, called the SAEM Monte Carlo Markov Chain (SAEM-MCMC) replaces the sampling of $z$ by one step of a Markov Chain targeting the conditional distribution $p$. Those two algorithms have then been applied in lots of different contexts: deformable models (Allassonnière et al., 2010; Bône et al., 2018a; Debavelaere et al., 2020; Schiratti et al., 2015), Independent Component Analysis (Allassonniere et al., 2012) and in many medical problems (see Benzekry et al. (2014); Guedj and Perelson (2011); Lavielle and Mentré (2007); Sissoko et al. (2016) among many others).

Among the hypotheses ensuring the convergence of most of these algorithms, and in particular our focus, the SAEM algorithm, one of the most restrictive is the necessity for the joint likelihood to belong to the curved exponential family. This writes:

$$f(y,z,\theta) = \exp\left(-\Psi(\theta) + \langle S(y,z), \Phi(\theta)\rangle\right) , \tag{6.1}$$

where $S$ is called a sufficient statistic of the model and $\Phi$ and $\Psi$ are two functions on $\Theta$. Similarly, the different extensions to the SAEM algorithm, and some of the EM algorithm, carry the same assumption (Kuhn et al., 2020; Lartigue et al., 2020; Panhard and Samson, 2009; Samson et al., 2006).

However, this hypothesis can in fact be a bottleneck in lots of situations. For example, it is not verified for heteroscedastic models (Dubois et al., 2011; Kuhn and Lavielle, 2005) nor with more complex models (Bône et al., 2018a; Debavelaere et al., 2020; Lindsten, 2013; Meza et al., 2012; Schiratti et al., 2015; Wang, 2007). Most of the authors then choose to compute the maximization step using a gradient descent. However, in that case, there is no theoretical guarantee of convergence. Moreover, the computational complexity increases. One needs to compute the gradient descent steps and compute the stochastic approximation of the complete likelihood while this function may not have a simple form. To solve this problem, Kuhn and Lavielle (2005) propose to transform the initial model to make it curved exponential.

Their solution consists in considering the parameter $\theta$ as a realization of a Gaussian vector of mean $\bar{\theta}$ and fixed variance $\sigma^2$. $\theta$ then becomes an additional latent variable and the new parameter to estimate is $\bar{\theta}$. We call this new model the exponentialized model. It now belongs to the curved exponential family. However, as the likelihood of this exponentialized model is different, the function to maximize has also been modified. In particular, there is no guarantee that the new parameter to estimate is close to the initial one. Nevertheless, this trick has been successfully used in different situations (Ajmal et al. (2019); Bône et al. (2018a); Debavelaere et al. (2020); Schiratti et al. (2015) among others).

In this paper, we will study the maximum likelihood of this new exponentialized model and measure its distance to one of the maxima of the initial likelihood. More precisely, we will show that this distance goes to $0$ as the variance $\sigma^2$ of the exponentialized model tends to $0$. We will also provide an upper bound to this error when $\sigma$ is small enough. Finally, we will verify those results on an example. This example will show us that a compromise must be done in the choice of $\sigma$. Indeed, if $\sigma$ is too big, a substantial error is made in the estimation. However, for $\sigma$ too small, despite the theoretical guarantees, the numerical convergence is difficult to obtain. To overcome this problem, we will present a new algorithm allowing a better estimation of the initial parameter $\theta$ in a reasonable computation time.

## 6.2 Presentation of the Stochastic Approximation Expectation Maximization (SAEM) Algorithm

In this section, we recall the Stochastic Approximation Expectation Maximization (SAEM) algorithm, first presented in Delyon et al. (1999) and recall the hypotheses ensuring convergence. In the following, we suppose that the observation $y$ belongs to $\mathbb{R}^n$, the latent variable $z$ to $\mathbb{R}^l$ and that the parameter space $\Theta$ is an open subset of $\mathbb{R}^p$ with $n, l, p \in \mathbb{N}^*$.

### 6.2.1 Expectation Maximization (EM) Algorithm

The original EM algorithm proposes to maximize a function defined via:

$$g(y,\theta) = \int_{\mathbb{R}^l} f(y,z,\theta)\mu(\mathrm{d}z)$$

with $f$ the joint likelihood of the model and $\mu$ is a $\sigma$-finite measure on $\mathbb{R}^l$.

This situation is of interest to estimate the parameters of a statistical model using maximum likelihood estimates where the model depends on unobserved latent variables.

The Expectation-Maximization consists of iterations which guarantee an increase in $g(\theta_k)$ at each step. Starting from $\theta_0$, the algorithm iterates:

- *Expectation.* Compute $Q_k(\theta) = \int_{\mathbb{R}^l} \log\left(f(y,z,\theta)\right) p(y,z,\theta_k)\mathrm{d}z$.
- *Maximization.* Set $\theta_{k+1} =$ argmax $Q_k(\theta)$.

where $p$ is the conditional distribution of $z$ given the observations:

$$p(y,z,\theta) = \begin{cases} f(y,z,\theta)/g(y,\theta) & \text{if } g(y,\theta) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

### 6.2.2 SAEM Algorithm

Because the expectation with respect to the conditional distribution $p(y,z,\theta)$ is often intractable in practice, a different approach suggests replacing the E-step by a stochastic approximation on $Q$, starting from $\theta_0$ and $Q_0 = 0$. This gives us the following algorithm:

- *Simulation.* Generate $z_k$, a realization of the hidden variable under the conditional density $p(y,z,\theta_k)$.
- *Approximation.* Update

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\log f(y,z_k,\theta) - Q_{k-1}(\theta))\,. \tag{6.2}$$

- *Maximization.* Set $\theta_{k+1} \in$ argmax $Q_k(\theta)$.

Convergence of this procedure is shown under the following hypotheses:

(M1) The parameter space $\Theta$ is an open subset of $\mathbb{R}^p$, and $f$ can write:

$$f(y,z,\theta) = \exp\left(-\Psi(\theta) + \langle S(y,z), \Phi(\theta)\rangle\right)\,, \tag{6.3}$$

where $S(\cdot)$ is a Borel function taking its value in $\mathcal{S}$, an open subset of $\mathbb{R}^{n_s}$. In that case, we say that $f$ belongs to the curved exponential family.
Moreover, the convex hull of $S(\mathbb{R}^l)$ is included in $\mathcal{S}$ and, for all $\theta \in \Theta$,

$$\int_{\mathbb{R}^l} |S(y,z)| p(y,z,\theta)\mu(dz) < \infty\,.$$

(M2) The functions $\Psi$ and $\Phi$ are twice continuously differentiable on $\Theta$.

(M3) The function $s : \Theta \to \mathcal{S}$ defined as:

$$s(\theta) = \int_{\mathbb{R}^l} S(y, z)p(y, z, \theta)\mu(dz)$$

is continuously differentiable on $\Theta$.

(M4) The observed log-likelihood $l(\theta) := \log g(y, \theta)$ is continuously differentiable on $\Theta$ and

$$\partial_\theta g(y, \theta) = \int_{\mathbb{R}^l} \partial_\theta f(y, z, \theta)\mu(dz)\,.$$

(M5) There exists a function $\hat{\theta} : \mathcal{S} \to \Theta$ such that $\forall \theta \in \Theta, \forall s \in \mathcal{S}, L(s, \hat{\theta}(s)) \geq L(s, \theta)$, with $L(s, \theta) = -\Psi(\theta) + \langle s, \Phi(\theta)\rangle$.
Moreover, $\hat{\theta}$ is continuously differentiable on $\mathcal{S}$.

(SAEM1) For all $k \geq 0$, $0 \leq \gamma_k \leq 1$, $\sum_{i=1}^{\infty} \gamma_k = \infty$ and $\sum_{i=1}^{\infty} \gamma_k^2 < \infty$.

(SAEM2) $\hat{\theta} : \mathcal{S} \to \Theta$ and the observed-data log likelihood $l : \theta \to \mathbb{R}$ are $n_s$ times differentiable.

(SAEM3) For all positive Borel function $\phi$:

$$\mathbb{E}(\phi(z_{k+1})|\mathcal{F}_k) = \int_{\mathbb{R}^l} \phi(z)p(z, \theta_k)\mu(\mathrm{d}z)\,,$$

where $z_k$ is the missing value simulated at step $k$ under the conditional density $p(y, z, \theta_{k-1})$ and $\mathcal{F}_n$ is the family of $\sigma$-algebra generated by the random variables $S_0, z_1, \ldots, z_n$.

(SAEM4) For all $\theta \in \Theta$, $\int_{\mathbb{R}^l} ||S(y, z)||^2 p(y, z, \theta)\mu(dz) < \infty$ and $\Gamma(\theta) := \mathrm{Cov}_\theta(S(y, z))$ is continuous with respect to $\theta$.

With the hypothesis (M1) specifying the form of the complete likelihood and (M5) giving us the existence of a maximizer $\hat{\theta}$, the algorithm can take a simpler form. Indeed, using the fact that $Q$ is fully defined by a sufficient statistic $S$, we remark, by linearity, that the stochastic approximation (6.2) is only applied on this sufficient statistic. Similarly, the maximization step can be rewritten using only the sufficient statistic and $\hat{\theta}$. This gives the following algorithm:

- *Simulation.* Generate $z_k$, a realization of the hidden variable under the conditional density $p(y, z, \theta_k)$.
- *Approximation.* Update $S_k = S_{k-1} + \gamma_k(S(y, z) - S_{k-1})$
- *Maximization.* Set $\theta_{k+1} = \hat{\theta}(S_k)$.

We finally assume the following hypothesis:

(A) With probability 1, clos$((S_k)_{k\geq 1})$ is a compact subset of $\mathcal{S}$.

**Remark 6.2.1.** *The assumption (A) can be relaxed by projecting the sequence $(S_k)_{k\in\mathbb{N}}$ on increasing compacts. See Andrieu et al. (2005) for more details.*

Under the hypotheses (M1)-(M5), (SAEM1)-(SAEM4) and (A), it was shown in Andrieu et al. (2005) that the distance between the sequence generated by the SAEM and the set of stationary point of the observed likelihood $g$ converges almost surely towards $0$.

However, in numerous cases, even quite simple (Dubois et al., 2011; Kuhn and Lavielle, 2005), the joint likelihood $f$ does not verify the hypothesis (M1) as it does not belong to the curved exponential family. In the next section, we will present a trick allowing us to approximate the maximum likelihood when (M1) is not verified.
In the following, to simplify the notations, we no longer write the variable $y$ in the different expressions.

### 6.2.3 Exponentialization process

We now denote by $(\theta, \psi)$ the parameters of $g$ where $\theta \in \Theta$, $\psi \in \Psi = \mathbb{R}^m$, and we tackle the case where the model cannot be written under the curved exponential form (6.1) because of the parameter $\psi$. In that case, the log-likelihood can only be written as:

$$f(z, \theta, \psi) = \exp\left(-\Psi(\theta) + \langle S(z), \Phi(\theta)\rangle\right) h(z, \psi) \tag{6.4}$$

and $f$ does not belong to the curved exponential family.

Here, some parameters $\theta$ are separable from the latent variables $z$ and do not require further transformation. Other variables $\psi$ are at the source of the computational problem and the exponentialization process will only apply on those parameters. It must be noticed that, in some cases, $\theta$ can be empty.

The trick proposed in Kuhn and Lavielle (2005) is to consider $\psi$ as a Gaussian random variable $\psi \sim \otimes\mathcal{N}(\bar{\psi}, \sigma^2)$, where the notation $\otimes\mathcal{N}(.,.)$ denotes a multivariate Gaussian distribution with diagonal covariance matrix. Hence, in this augmented model, $\psi$ is no longer a parameter but becomes an additional latent variable while a new parameter $\bar{\psi}$ appears.

The resulting perturbed statistical model is curved exponential, with augmented parameters $\hat{\theta} = (\theta, \bar{\psi})$ and augmented random latent variables $\hat{z} = (z, \psi)$.

The variance $\sigma^2$ is chosen by the user, and should be reasonably small in order to minimally perturb the original model. In practice, this variance should at the same time be chosen reasonably large in order to speed-up the parameter estimation (see experiments in section 6.4).

The complete log-likelihood of this exponentialized model then writes:

$$\log f_\sigma(y, \hat{z}, \hat{\theta}) = -\Psi(\theta) + \langle S(z), \Phi(\theta)\rangle + \log(h(z, \psi)) - \frac{||\psi - \bar{\psi}||^2}{2\sigma^2}\,. \tag{6.5}$$

It is easy to check that the complete log-likelihood now belongs to the curved exponential family with sufficient statistics: $(S(z), \psi)$. Concerning the parameter $\theta$, the maximization is done as usual: $\theta_{k+1} = \hat{\theta}(S_k)$ with $S_k$ the stochastic approximation of the $(S(z_i))_{i\leq k}$. The update of the parameter $\bar{\psi}$ can for its part be written as:

$$\bar{\psi}_{k+1} = \bar{\psi}_k + \gamma_k(\psi_k - \bar{\psi}_k)\,. \tag{6.6}$$

If we suppose that this augmented model satisfies the hypotheses (M1)-(M5), (SAEM1)-(SAEM4) and (A), we know, using the theorem proved in Andrieu et al. (2005), that it will converge towards a critical point of its incomplete likelihood. However, if this process is used in several applications (Lavielle, 2014), there is in fact no guarantee that the algorithm will converge towards a critical point of the incomplete log-likelihood of the initial model.

In the following section, we show that, in general, the parameter returned by the SAEM on the exponentialized model is indeed not a maximum likelihood of the initial model. However, when $\sigma$ goes to 0, it converges towards a critical point of the incomplete log likelihood of the initial model. We also give an upper bound of the error made by this process for $\sigma$ small.

It is interesting to notice that, even if this proof is done in the context of the SAEM algorithm, the same results can be obtained for the MCMC-SAEM (Kuhn and Lavielle, 2004) as well as for the Approximate SAEM (Allassonnière and Chevallier, 2019).

## 6.3 Distance between the limit point and the nearest critical point

In this section, we first present an equation satisfied by the limit of the sequence of estimated parameters of the SAEM algorithm for the exponentialized model. Using this equation, we will then give an upper bound on the distance between this limit point and the nearest critical point of the incomplete likelihood of the non-exponential model. This upper bound will in particular show us that this distance tends to 0 when $\sigma$ goes to 0.

### 6.3.1 Equation verified by the limit

We now state a theorem giving us an equation satisfied by the limit parameter estimated by the SAEM algorithm applied on the exponentialized model. It is important to remark that, if the set of the critical points of $l$ is finite then, the SAEM algorithm converges almost surely towards one of them (and not only towards a point at zero distance). Hence, we can study the parameters returned by the SAEM on the exponential model: $\bar{\psi}_\sigma$ and look at their behaviour when $\sigma$ goes to 0.

**Theorem 6.3.1** *Assume that the exponentialized model with variance $\sigma$ verifies the hypotheses (M1)-(M5), (SAEM1)-(SAEM4), (A) and that $\Psi = \mathbb{R}^m$. Assume also that, for all $\sigma > 0$,*

$$\mathcal{L}_\sigma := \left\{(\theta, \bar{\psi}) \in \Theta \times \Psi \,|\, \partial_{\theta,\bar{\psi}} l_\sigma(\theta, \bar{\psi}) = 0\right\}$$

*is finite where $l_\sigma$ refers to the observed log-likelihood of the exponentialized model of variance $\sigma$. Then, the sequence returned by the SAEM algorithm converges almost surely towards $(\theta_\infty, \bar{\psi}_\sigma)$, solutions of the following set of equations: $\forall 1 \leq k \leq m$,*

$$\int_{\mathbb{R}^m} v_k g(\theta_\infty, \bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv = 0\,, \tag{6.7}$$

*where $v_k$ is the $k$-th coordinate of $v \in \mathbb{R}^m$ and $g(\theta, \psi) = \int_Z f(z, \theta, \psi)\, \mu(dz)$.*

**Remark 6.3.1.** *Here, we suppose $\Psi = \mathbb{R}^m$ to be able to define a Gaussian distribution on $\Psi$. The following proofs would be adaptable as long as one can define such a gaussian distribution, necessary for applying the exponentialization trick.*

*Proof.* The update of $S_k, \psi_k$ can easily be seen as a Robbins Monro update:

$$\begin{cases} \bar{\psi}_{k+1} = \bar{\psi}_k + \gamma_k v_k \\ S_{k+1} = S_k + \gamma_k (S(z_k) - S_{k-1}) \end{cases}$$

where $z_k, v_k$ are sampled following the conditional law $f_\sigma(y, z, \psi + \bar{\psi}_k | \theta_k, \bar{\psi}_k)$.

Under the hypotheses explained section 6.2.2, it has been shown that this Robbin Monroe approximation verifies $\lim_{k\to\infty} d((\theta_k, \bar{\psi}_k), \mathcal{L}_\sigma) = 0$. Moreover, because $\mathcal{L}_\sigma$ is finite, Delyon et al. (1999) show that the sequence $(\theta_k, \bar{\psi}_k)_{k\geq 0}$ converges almost surely towards a point $(\theta_\infty, \bar{\psi}_\sigma) \in \Theta \times \Psi$ (theorem 6). Using the regularity of $l_\sigma$, we deduce that those parameters verify $\partial_{\theta,\bar{\psi}} l_\sigma(\theta_\infty, \bar{\psi}_\sigma) = 0$.
By replacing $l_\sigma$ by its value in this equation and using the assumption (M4), we find, for all $1 \leq k \leq m$:

$$\int_{\mathbb{R}^m} (v_k - \bar{\psi}_\sigma) q(y, z, \theta_\infty, v) \exp(-\frac{||v - \bar{\psi}_\sigma||^2}{2\sigma^2}) dv dz = 0$$

Using a change of variable, we finally find the expected result.

□

**Proposition 6.3.1** *Suppose that a point $\psi \in \mathbb{R}^m$ verifies:*

$$\forall v, \theta \in \mathbb{R}^m \times \Theta,\ g(\theta, \psi + v) = g(\theta, \psi - v)\,. \tag{6.8}$$

*Then, $\psi$ is solution to Eq.* (6.7) *and can be the parameter returned by the exponential model.*

**Remark 6.3.2.** *It is not necessarily the only possibility of returned parameter. Several values of $\psi$ could be solutions of Eq.* (6.7).

In particular, a solution of (6.8) is a critical point of $g$. However, a critical point of $g$ is not always solution of such an equation and will not always be a solution of Eq. (6.7). It is in fact easy to find cases where the maximum is not a solution to Eq. (6.7) and hence where the exponentialized model introduces a bias in the estimation of maximum likelihood.
We introduce next subsection a function $g$ presenting such a behaviour and we explain the heuristics behind Theorem 6.3.1.

### 6.3.2 Heuristics

We want to compare the solution of equation (6.7) to a maximum of the function $g$. Because of the form of $f$ supposed in equation (6.4), we see that we can maximize $g$ in $\theta$ and $\psi$ independently of the other. In particular, we still immediately have $\theta_\infty \in \text{argmax}_\theta g(\theta, \psi)$ (independent of $\psi$ as can be seen in equation (6.4)).

To explain equation (6.7), we introduce the function $g : v \mapsto \frac{1}{v}\exp(-\frac{1}{v^2})$ presented Figure 6.1. This function has a maximum for $\psi = \sqrt{2}$ but is not symmetric around it. We will look at the integral:

$$\int_{\Psi} vg(\psi + v)\exp\left(-\frac{v^2}{2\sigma^2}\right) dv$$

for different values of $\psi$ and $\sigma$. $\psi$ is a solution of equation (6.7) if and only if this integral is null. It is interesting to remark that one can consider this integral as an expectation if normalized.

First we look at the case $\psi = \sqrt{2}$, the argmax of $g$, and $\sigma = 1$ on Figure (6.2a). In that case, because $g$ is not symmetric around its maximum, $v \mapsto g(\sqrt{2} + v)\exp\left(-\frac{v^2}{2\sigma^2}\right)$ is not symmetric either. In particular, it means that $\sqrt{2}$ is not a solution of equation (6.7) as the integral is strictly positive.

We then reduce the value of $\sigma$ by taking $\sigma = 0.1$ on Figure (6.2b). The function $v \mapsto g(\sqrt{2} + v)\exp\left(-\frac{v^2}{2\sigma^2}\right)$ is still not symmetric around $0$. Even if the value of the integral is smaller, $\sqrt{2}$ is still not a solution of equation (6.7).

We now interest ourselves in the case where $\psi$ is not the argmax of $g$ by taking $\psi = 1$ and $\sigma = 1$ on Figure 6.3. This time, $g$ is strictly increasing from 1 to $\sqrt{2}$. Hence, even if we multiply by the exponential, $v \mapsto g(1 + v)\exp\left(-\frac{v^2}{2\sigma^2}\right)$ is still increasing at 0. As $g$ decreases slower than it increases, $1$ cannot be a solution of equation (6.7). The integral is indeed strictly positive. The same behaviour would be observed for any point before $\sqrt{2}$.

We now look at a value bigger than the maximum: $\psi = 4$ and $\sigma = 5$ Figure 6.4a. This time, as $g$ decreases at $v = 4$, $v \mapsto g(4 + v)\exp\left(-\frac{v^2}{2\sigma^2}\right)$ still decreases at 0. But $g$ decreases slower than it increases. Hence, it is possible to compensate this difference of variation by taking $\psi > \sqrt{2}$ as a solution of equation (6.7).

Let us now take a smaller value of $\sigma$ as in Figure 6.4b. This time, the integral is negative. Indeed, $v \mapsto g(4 + v)\exp\left(-\frac{v^2}{2\sigma^2}\right)$ still decreases at 0. However, due to the multiplication by the exponential, the difference of variation before and after the maximum is now way smaller. In particular, it is this time too small to compensate the decrease at 0 and the integral will be negative. To have a solution of equation (6.7) for this value of $\sigma$, we would need to choose a value of $\psi$ smaller. This suggests that, as $\sigma$ goes to 0, the solution of equation (6.7) is closer to $\sqrt{2}$.

From these examples, we can deduce two things. First, the argmax of $g$ is not always solution of the equation (6.7) even for small values of $\sigma$ because there is a difference in the speed of variation before and after this maximum. However, it is possible to compensate this difference of variation by choosing a value different than the argmax and thus to find a solution of (6.7) different than the argmax. Moreover, when $\sigma$ goes to 0, the difference of variation obtained by multiplying by the exponential is smaller and smaller. It means that a parameter closer and closer to the argmax will be solution of equation (6.7).
We illustrate this behaviour by plotting the exact value of the solution of equation (6.7) as a function of $\sigma$ in Figure 6.5.

In the following, we write $\psi_M$ the critical point of $g(\theta, \psi)$ minimizing the distance to $\bar{\psi}_\sigma$. Using the heuristics presented above, we will, in the next section, state the theorem giving us an upper bound on the distance to the nearest critical point of $g$. We will then prove it in the case $\Psi = \mathbb{R}$. A more general proof in $\mathbb{R}^m$ for $m \geq 2$ is given in the annex.

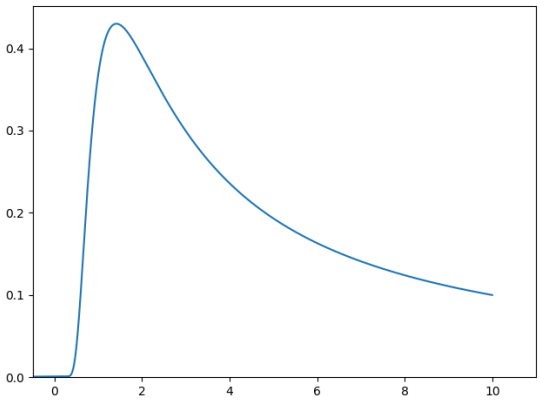
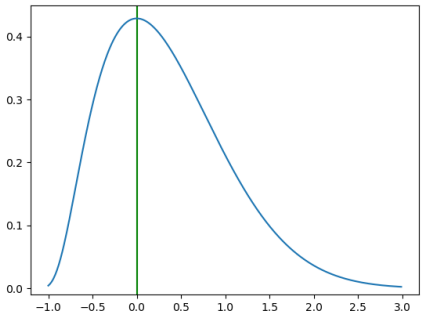
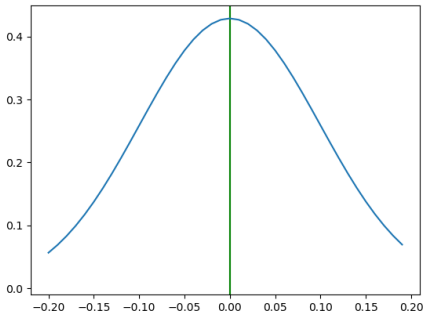



**Figure 6.1** Function $g$ studied subsection 6.3.2, with a maximum reached at $\sqrt{2}$.




**(a)** $v \mapsto g(\theta, \sqrt{2} + v) \exp(\frac{-v^2}{2})$




**(b)** $v \mapsto g(\theta, \sqrt{2} + v) \exp(\frac{-v^2}{2*0.1^2})$

**Figure 6.2** Plot of $v \mapsto g(\theta, \sqrt{2} + v) \exp(\frac{-v^2}{2\sigma^2})$ for different values of $\sigma$. In all cases, we can see that $\sqrt{2}$ is not a solution of Eq. (6.7).

### 6.3.3 Upper bound on the distance between $\bar{\psi}_\sigma$ and the nearest critical point of g

**Theorem 6.3.2**

1. *Assume that the exponential model verifies the hypotheses (M1)-(M5), (SAEM1)-(SAEM4) and (A). Assume also that, for all $\sigma > 0$, $\mathcal{L}_\sigma$ is finite, that $\mathcal{L} := \{\psi \in \mathbb{R}^m | \partial_\psi g(\theta, \psi) = 0\}$ is*

*compact and that there exists $K$ compact such that, $\forall \sigma > 0$, $\bar{\psi}_\sigma \in K$. Then,*

$$d(\bar{\psi}_\sigma, \mathcal{L}) \xrightarrow[\sigma \to 0]{} 0 \,.$$

2. *Assume also that $\mathcal{L}$ is finite and that, for all $\psi_M \in \mathcal{L}$, there exists an integer $l_M$ such that $g$ is $l_M$-times continuously differentiable and such that*

$$\forall k \leq m, \exists i \leq l_M \text{ with } \frac{\partial^i g}{\partial \psi_k^i}(\psi_M) \neq 0 \,.$$

*We write $l = max_{\psi_M \in \mathcal{L}} l_M$.*
*Then, there exists $c > 0$ such that, for $\sigma$ small enough,*

$$d(\bar{\psi}_\sigma, \mathcal{L}) \leq c\sigma^{\frac{2}{l+2}} \,. \tag{6.9}$$

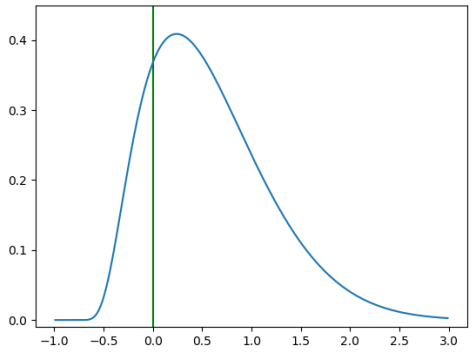
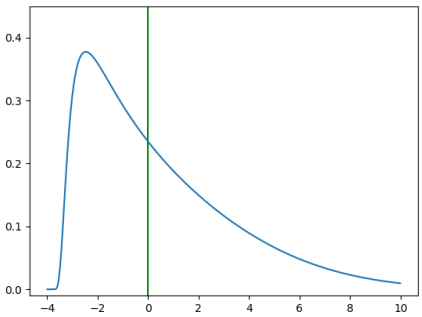
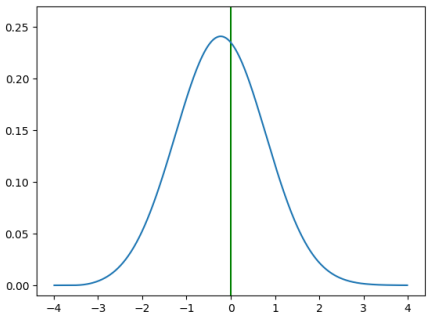



**Figure 6.3** Plot of $v \mapsto g(\theta, 1+v)\exp(\frac{-v^2}{2\sigma^2})$. Since $g$ is increasing at 1 and increases quicker than it decreases, $1$ is not solution of (6.7).




**(a)** $v \mapsto g(\theta, 4+v)\exp(\frac{-v^2}{2*5^2})$




**(b)** $v \mapsto g(\theta, 4+v)\exp(\frac{-v^2}{2})$

**Figure 6.4** Plot of $v \mapsto g(\theta, 4+v)\exp(\frac{-v^2}{2\sigma^2})$ for different values of $\sigma$.

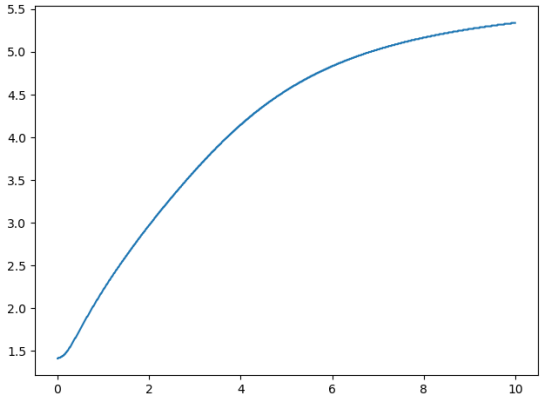



**Figure 6.5** Solution of Eq. (6.7) as a function of $\sigma$ for the function $g$ studied subsection 6.3.2.

3. *Suppose that $v \mapsto g(\theta_\infty, v)$ and $v \mapsto v_k g(\theta_\infty, v)$ are integrable for all $k$ between $1$ and $m$. Then, we have the following approximation of $\bar{\psi}_\sigma$ when $\sigma$ goes to infinity: for all $1 \leq k \leq m$,*

$$(\bar{\psi}_\sigma)_k \xrightarrow[\sigma \to \infty]{} \frac{\int_{\mathbb{R}^m} v_k g(\theta_\infty, v) dv}{\int_{\mathbb{R}^m} g(\theta_\infty, v) dv} .$$

**Remark 6.3.3.** *When $m = 1$, $l_M$ is the smallest integer such that the $l_M$-th derivative of $g(\theta_\infty, .)$ at $\psi_M$ is not 0. The inequality* (6.9) *indicates that the convergence will be slower when the function to maximize behaves as a flat curve around the maximum, which was expected.*
*If such a $l_M$ does not exist, it means that $g$ is constant in at least one direction around $\psi_M$.*

**Remark 6.3.4.** *We need to take a maximum over all $\psi_M$ in $\mathcal{L}$ since, from one $\sigma$ to another, $\psi_\sigma$ can approach a different maximum $\psi_M \in \mathcal{L}$. It is also why the upper bound depends on this maximum $l$. It is constrained by the critical point for which the convergence is the slowest.*

**Remark 6.3.5.** *For $m = 1$, we have the exact value of the constant in* (6.9)*:*

$$c = \max_{\psi_M \in \mathcal{L}} \left[ \left( 12(l_M - 1)! \frac{||g||_\infty}{|\partial_\psi^{l_M} g(\theta_\infty, \psi_M)|} \right)^{\frac{1}{l_M + 2}} \right] .$$

**Remark 6.3.6.** *The results presented here would still be true in the case of the SAEM-MCMC algorithm. Indeed, the equation* (6.7) *verified by the limit parameter of the SAEM would still be verified in the case of the SAEM-MCMC and the same reasoning could then be done.*

In the following, we will present the proof in the case $m = 1$. The proof in the multi-dimensional case follows the same ideas than in dimension one but is more technical. It is presented in the annex.

*Proof.* We present the proof in the case $m = 1$. As the maximum does not depend of $\theta_\infty$ and to simplify notations, we will forget the variable $\theta$ in $g$ and use $g(\psi) = g(\theta_\infty, \psi)$.

**First step:** $d(\bar{\psi}_\sigma, \mathcal{L}) \xrightarrow[\sigma \to 0]{} 0$

We suppose that $\bar{\psi}_\sigma$ is never a critical point of $g$ for $\sigma$ small enough. Otherwise, we directly have the result. The equation (6.7) writes:

$$\int_{\mathbb{R}} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv = 0\,.$$

The first step is to show that $d(\bar{\psi}_\sigma, \mathcal{L}) \xrightarrow[\sigma \to 0]{} 0$. By contradiction, even if it means extracting, we can suppose that there exists $c > 0$ such that $\forall \sigma > 0$, $d(\bar{\psi}_\sigma, \mathcal{L}) > 3c$.

Because there is no critical point between $\bar{\psi}_\sigma - c$ and $\bar{\psi}_\sigma + c$, $g$ is either increasing or decreasing on $[\bar{\psi}_\sigma - c, \bar{\psi}_\sigma + c]$. We first suppose it is increasing. In particular, $K_0 := K \setminus \{y \,|\, d(y, \mathcal{L}) < c\}$ is compact and thus $c_0 := \inf\{g'(y) \,|\, y \in K_0,\, g'(y) \geq 0\} > 0$. According to equation (6.7), the integral on $[-c, c]$ must have the same absolute value as the integral on $[-c, c]^c$. However, we will show that, when $\sigma$ goes to zero, the first one converges towards 0 much more slowly than the second one. Indeed,

$$\begin{aligned}
\int_{|v| \geq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv &\geq \int_{v \leq -c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv \\
&\geq ||g||_\infty \int_{v \leq -c} v \exp(-\frac{v^2}{2\sigma^2}) dv \\
&\geq -\sigma^2 ||g||_\infty \exp(-\frac{c^2}{2\sigma^2})\,.
\end{aligned}$$

On the other hand, we have:

$$\int_{|v| \leq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv = \int_{0 \leq v \leq c} v \left(g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma - v)\right) \exp(-\frac{v^2}{2\sigma^2}) dv\,.$$

Using the mean value theorem, for all $0 \leq v \leq c$, there exists $\tilde{\psi}_v \in [\bar{\psi}_\sigma - v, \bar{\psi}_\sigma + v] \subset K_0$ such that $g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma - v) = 2v g'(\tilde{\psi}_v) \geq 2c_0 v$. Hence, we find:

$$\int_{|v| \leq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv \geq 2c_0 \int_{0 \leq v \leq c} v^2 \exp(-\frac{v^2}{2\sigma^2}) dv\,.$$

But, using an integration per part and defining

$$\mathsf{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt\,,$$

we have:

$$\int_{|v| \leq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv \geq 2c_0 \sigma^2 \left[-c \exp(-\frac{c^2}{2\sigma^2}) + \sigma \frac{\sqrt{\pi}}{2} \mathsf{erf}(\frac{c}{\sqrt{2}\sigma})\right]\,.$$

Hence, because

$$\int_{|v| \leq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv = -\int_{|v| \geq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv\,,$$

we have:

$$||g||_\infty \exp(-\frac{c^2}{2\sigma^2}) \geq 2c_0 \left[-c \exp(-\frac{c^2}{2\sigma^2}) + \sigma \frac{\sqrt{\pi}}{2} \mathsf{erf}(\frac{c}{\sqrt{2}\sigma})\right]\,.$$

It is easy to find the same inequality if $g$ is decreasing on $[\bar{\psi}_\sigma - c, \bar{\psi}_\sigma + c]$. Indeed, in that case, by integrating only on $\{v \geq c\}$, we first show that

$$\int_{|v|\geq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv \leq \sigma^2 ||g||_\infty \exp(-\frac{c^2}{2\sigma^2}) \,.$$

Then, by considering this time $c_1 := \sup\{g'(y) \,|y \in K_0,\, g'(y) \leq 0\} < 0$, we find:

$$\int_{|v|\leq c} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv \leq 2c_1\sigma^2 \left[-c\exp(-\frac{c^2}{2\sigma^2}) + \sigma\frac{\sqrt{\pi}}{2}\mathsf{erf}(\frac{c}{\sqrt{2}\sigma})\right] .$$

Hence, for all $\sigma > 0$, there exists $C := 2\max(c_0, -c_1) > 0$ such that:

$$\frac{||g||_\infty}{\sigma}\exp(-\frac{c^2}{2\sigma^2}) \geq C\left[-\frac{c}{\sigma}\exp(-\frac{c^2}{2\sigma^2}) + \frac{\sqrt{\pi}}{2}\mathsf{erf}(\frac{c}{\sqrt{2}\sigma})\right] .$$

By taking $\sigma$ to 0 and using the fact that $\mathsf{erf}(x) \xrightarrow[x\to\infty]{} 1$, we find $C \leq 0$ which is a contradiction. Hence, we have proved that

$$d(\bar{\psi}_\sigma, \mathcal{L}) \xrightarrow[\sigma\to 0]{} 0 \,.$$

The next step is to find an upper bound on $d(\bar{\psi}_\sigma, \mathcal{L})$.

**Second step: Search of the upper bound**

In the following, we will suppose that the critical point towards which $\bar{\psi}_\sigma$ converges is a maximum. In practice, it will always be the case as any other critical point would be unstable numerically. Theoretically, a set of conditions (LOC1)-(LOC3) are given in Delyon et al. (1999) insuring the convergence towards a local maximum.

We write $\psi_M$ the closest critical point to $\bar{\psi}_\sigma$ and $\alpha_\sigma = |\bar{\psi}_\sigma - \psi_M|$. We also write $l_M$ the smallest integer such that $g^{(l_M)}(\psi_M) \neq 0$. Moreover, as explained above, we assume that $\psi_M$ is a maximum. It must be remarked that $\psi_M$ depends on $\sigma$. However, as $\mathcal{L}$ is finite, we will be able to consider maxima at the end of the proof. Since we assume $\psi_M$ maximum, $l_M$ is even and, for $\sigma$ small enough, since $g^{(l_M)}$ is continuous, $\forall v \in [\bar{\psi}_\sigma - \alpha_\sigma, \bar{\psi}_\sigma + \alpha_\sigma]$,

$$g^{(l)}(v) \leq \frac{1}{2}g^{(l_M)}(\psi_M) := -c_M < 0 \,.$$

As before, we will split up the integral (6.7) in two parts: $\{v||v| < \alpha_\sigma\}$ and $\{v||v| > \alpha_\sigma\}$. The idea behind the computations is that $\alpha_\sigma$ cannot be too big without making the absolute value of the integral on $\{v||v| < \alpha_\sigma\}$ strictly superior than the one on $\{v||v| > \alpha_\sigma\}$.

On $\{v\,|\,|v| > \alpha_\sigma\}$ we can use the same upper and lower bounds as before to find:

$$\left|\int_{|v|\geq\alpha_\sigma} v g(\bar{\psi}_\sigma + v)\exp(-\frac{v^2}{2\sigma^2})dv\right| \leq \sigma^2||g||_\infty\exp(-\frac{\alpha_\sigma^2}{2\sigma^2}) \,. \tag{6.10}$$

On $\{v\,|\,|v| < \alpha_\sigma\}$, we use twice the mean value theorem to find, for any $v \in [0, \alpha_\sigma]$, there exist $\tilde{\psi}_v^0 \in [\bar{\psi}_\sigma - v, \bar{\psi}_\sigma + v]$ and $\tilde{\psi}_v^1 \in [\bar{\psi}_\sigma - \alpha_\sigma, \bar{\psi}_\sigma + \alpha_\sigma]$ such that:

$$\begin{aligned} g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma - v) &= 2vg'(\tilde{\psi}_v^0) = 2v(g'(\tilde{\psi}_v^0) - g'(\psi_M)) \\ &= 2v(\tilde{\psi}_v^0 - \psi_M)^{l_M-1} g^{(l_M)}(\tilde{\psi}_v^1)/(l_M-1)! \\ &\geq 2v(\bar{\psi}_\sigma + v - \psi_M)^{l_M-1} g^{(l_M)}(\tilde{\psi}_v^1)/(l_M-1)! \,. \end{aligned}$$

We first suppose that $g$ is increasing on $[\bar{\psi}_\sigma - \alpha_\sigma, \bar{\psi}_\sigma + \alpha_\sigma]$. Then, $\alpha_\sigma = \psi_M - \bar{\psi}_\sigma$ and:

$$g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma - v) \geq \frac{2c_M}{(l_M - 1)!} v(\alpha_\sigma - v)^{l_M - 1} .$$

Hence, computing the integral (6.7) on $\{v \,|\, |v| < \alpha_\sigma\}$, we find:

$$\begin{aligned}
\int_{|v| \leq \alpha_\sigma} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv &\geq \frac{2c_M}{(l_M - 1)!} \int_0^{\alpha_\sigma} v^2 (\alpha_\sigma - v)^{l_M - 1} \exp(-\frac{v^2}{2\sigma^2}) dv \\
&\geq \frac{2c_M}{(l_M - 1)!} \alpha_\sigma^{l_M + 2} \int_0^1 v^2 (1 - v)^{l_M - 1} \exp(-\frac{\alpha_\sigma^2 v^2}{2\sigma^2}) dv \\
&\geq \frac{2c_M}{(l_M - 1)!} \alpha_\sigma^{l_M + 2} \exp(-\frac{\alpha_\sigma^2}{2\sigma^2}) \int_0^1 v^2 (1 - v) dv .
\end{aligned}$$

Finally, by combining this inequality and (6.10), we find:

$$\sigma^2 ||g||_\infty \exp(-\frac{\alpha_\sigma^2}{2\sigma^2}) \geq \frac{c_M}{6(l_M - 1)!} \alpha_\sigma^{l_M + 2} \exp(-\frac{\alpha_\sigma^2}{2\sigma^2}) .$$

Hence, if $\sigma \leq 1$,

$$\begin{aligned}
\alpha_\sigma &\leq \left( 6(l_M - 1)! \frac{||g||_\infty}{c_M} \right)^{1/(l_M + 2)} \sigma^{\frac{2}{l_M + 2}} \\
&\leq \max_{\psi_M \in \mathcal{L}} \left( \left( 6(l_M - 1)! \frac{||g||_\infty}{c_M} \right)^{1/(l_M + 2)} \right) \sigma^{\frac{2}{l + 2}} .
\end{aligned}$$

Because $\mathcal{L}$ is finite, we indeed have a maximum which is strictly positive.

In the case where g in decreasing on $[\bar{\psi}_\sigma - \alpha_\sigma, \bar{\psi}_\sigma + \alpha_\sigma]$, we have $\alpha_\sigma = \bar{\psi}_\sigma - \psi_M$ and it is easy to show that we have this time

$$g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma - v) \leq -2c_M v(\alpha_\sigma - v)^{l_M - 1}/(l_M - 1)! .$$

Hence, we can use the same inequalities as before to find again:

$$\alpha_\sigma \leq \max_{\psi_M \in \mathcal{L}} \left( \left( 6(l_M - 1)! \frac{||g||_\infty}{c_M} \right)^{1/(l_M + 2)} \right) \sigma^{\frac{2}{l + 2}} .$$

**Third step: Approximation when $\sigma$ goes to infinity**

We use again the equation (6.7). For all $\sigma > 0$,

$$\int_{\mathbb{R}} v g(\bar{\psi}_\sigma + v) \exp(-\frac{v^2}{2\sigma^2}) dv = 0 .$$

Using the change of variable $\bar{\psi}_\sigma + v$, we find:

$$\bar{\psi}_\sigma = \frac{\int_{\mathbb{R}} v g(v) \exp\left( -\frac{(v - \bar{\psi}_\sigma)^2}{2\sigma^2} \right) dv}{\int_{\mathbb{R}} g(v) \exp\left( -\frac{(v - \bar{\psi}_\sigma)^2}{2\sigma^2} \right) dv} .$$

But $\bar{\psi}_\sigma$ is supposed to stay in a compact so, $\forall v \in \mathbb{R}$, $\exp\left( -\frac{(v - \bar{\psi}_\sigma)^2}{2\sigma^2} \right) \xrightarrow[\sigma \to \infty]{} 1$. Using the integrability of $g$ and $v \mapsto v g(v)$, it is easy to conclude using the dominated convergence theorem.

□

## 6.4 Simulation of a counter example

In this section, we demonstrate that the maximum likelihood of $g$ is indeed not reached by the SAEM algorithm on the exponentialized model on a concrete situation.

### 6.4.1 Application of the SAEM algorithm to the exponentialized model

We choose to study a heteroscedastic model where the variance depends on the observation. This model has been used in Kuhn and Lavielle (2005) in order to analyze the growth of orange trees. The parameters to estimate are the age $\beta_1$ at half asymptotic trunk circumference $\psi_i$ and the grow scale $\beta_2$ of $n$ orange trees according to the measurement of their circumference $y_{i,j}$ at $m$ different ages $x_j$.

We suppose that our observation $y_{i,j}$ verifies, for $i$ between $1$ and $n$ and $j$ between $1$ and $k_i$:

$$y_{i,j} = \frac{\phi_i}{1 + \exp\left(-\frac{x_j - \beta_1}{\beta_2}\right)}(1 + \varepsilon_{i,j}),$$

where $\varepsilon_{i,j}$ are independent noises of distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$ of variance $\sigma_\varepsilon^2$ supposed to be known. $\phi_i$ is treated as a random effect and is supposed to follow a Gaussian distribution of mean $\mu$ to estimate and known variance $\tau^2$.

Such a model cannot be written in an exponential form due to the parameters $\beta_1$ and $\beta_2$ and we will hence consider an exponentialized model where $\beta_1$ and $\beta_2$ are considered as random effects with $\beta_1 \sim \mathcal{N}(\bar{\beta}_1, \sigma^2)$ and $\beta_2 \sim \mathcal{N}(\bar{\beta}_2, \sigma^2)$.

Writing

$$h(\phi, \beta_1, \beta_2, x) = \frac{\phi}{1 + \exp\left(-\frac{x - \beta_1}{\beta_2}\right)},$$

the complete likelihood of the exponentialized model can then be written as:

$$f(y, \phi, \beta_1, \beta_2, \theta) = 2\pi\sigma^2(2\pi\sigma_\varepsilon^2)^{-nm/2}(2\pi\tau^2)^{-n/2} \cdot \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i,j}\left(\frac{y_{i,j}}{h(\phi_i, \beta_1, \beta_2, x_j)} - 1\right)\right.$$
$$\left. - \sum_{i,j}\log(h(\phi_i, \beta_1, \beta_2, x_j)) - \sum_i \frac{(\phi_i - \mu)^2}{2\tau^2} - \frac{(\beta_1 - \bar{\beta}_1)^2}{2\sigma^2} - \frac{(\beta_2 - \bar{\beta}_2)^2}{2\sigma^2}\right],$$

where $\theta = (\mu, \bar{\beta}_1, \bar{\beta}_2)$ are the exponentialized model parameters to estimate.

**Remark 6.4.1.** *It would be easy to suppose $\tau$ and $\sigma_\varepsilon^2$ unknown and estimate them using the SAEM algorithm. Those parameters would leave the joint distribution curved exponential and it would not be necessary to further exponentialize the model. To simplify, we assume them known here.*

It is then easy to show that this likelihood belongs to the curved exponential family with sufficient statistics being:

$$\begin{cases} S_1(\phi) = \sum_i \phi_i \,, \\ S_2(\beta_1) = \beta_1 \,, \\ S_3(\beta_2) = \beta_2 \,. \end{cases}$$

The maximum likelihood estimator can then be expressed as a function of $S_1(\phi)$, $S_2(\beta_1)$ and $S_3(\beta_2)$ as follows:

$$\begin{cases} \hat{\mu} = S_1(\phi)/n \,, \\ \hat{\hat{\beta}}_1 = S_2(\beta_1) \,, \\ \hat{\hat{\beta}}_1 = S_3(\beta_2) \,. \end{cases}$$

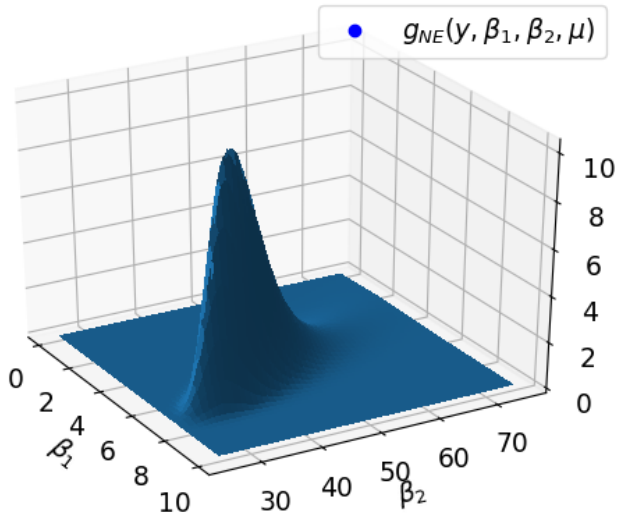
Because we cannot easily sample $(\phi, \beta_1, \beta_2)$ from the conditional distribution, we will not directly use the SAEM algorithm but the SAEM-MCMC algorithm. We replace the sampling step by one iteration of a Metropolis Hastings algorithm targeting the posterior distribution. Under hypotheses presented in Kuhn and Lavielle (2004), this process converges towards the same limit as the SAEM algorithm. In particular, it has been proved in Kuhn and Lavielle (2005) that those conditions are indeed verified here and thus that the algorithm converges. Moreover, as the limit is the same than the one given by the SAEM, our theorem 6.3.2 still applies.

We then create a synthetic dataset of a thousand observations following this model (100 subjects observed at 10 different ages). Knowing the exact value of $\mu$, we plot the incomplete likelihood of the non-exponentialized model $g_{NE}$ as a function of $(\beta_1, \beta_2)$ figure 6.6a. We also plot its behaviour around the maximum and along the axes $\beta_1$ and $\beta_2$ figures 6.6b and 6.6c.
As we can see, the function is not symmetric around the maximum. Hence, there should be a bias while estimating the maximum likelihood using the exponentialized model. More precisely, we can see the error in $\beta_2$ should be larger than the one in $\beta_1$ as the function is less symmetric along the $y$ axis than along the $x$ axis.

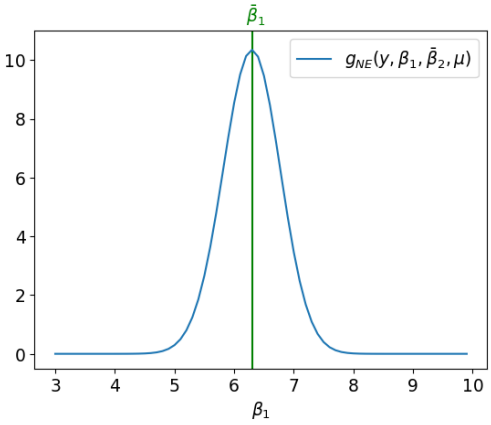
To verify this heuristic, we use the SAEM-MCMC algorithm and launch our algorithm a hundred times for different values of $\sigma$. We then compare the results given by the SAEM-MCMC algorithm to the exact value of the maximum likelihood of the initial model. Because we know the exact parameters from which the dataset has been simulated, we are also able to compute numerically the solution of the equation (6.7) as a function of $\sigma$. The results are presented figure 6.7.

For $\sigma \geq 1$, the results of the simulation follow our theory with the estimated parameters estimated close to the solution of the equation (6.7). Moreover, as expected, the error is bigger in the estimation of $\beta_2$ than in the estimation of $\beta_1$ (see axis scale).

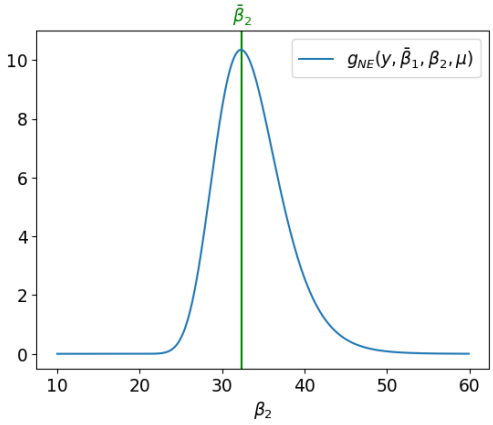
However, for a small $\sigma$, the algorithm does not converge. Indeed, in that case, the variance of the conditional distribution is really small as it is proportional to $\exp\left(-\frac{(\beta-\bar{\beta})^2}{2\sigma^2}\right)$. In particular, it means that the algorithm will be extremely long to converge and, in practice, will stay near the initial value ($\beta_1 = 6$, $\beta_2 = 34$ here).




**(a)** Plot of the incomplete likelihood of the initial model as a function of $(\beta_1, \beta_2)$.




**(b)** Plot of the incomplete likelihood of the initial model as a function of $\beta_1$ for $\tilde{\beta}_2$ the argmax of likelihood.

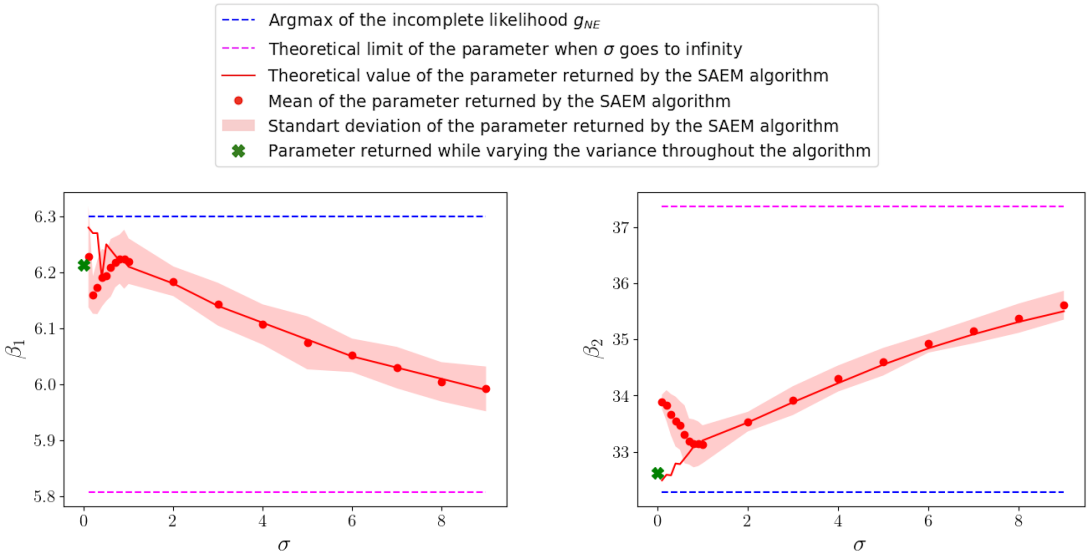



**(c)** Plot of the incomplete likelihood of the initial model as a function of $\beta_2$ for $\tilde{\beta}_1$ the argmax of likelihood.

**Figure 6.6** Plot of the incomplete likelihood of the initial model as a function of $(\beta_1, \beta_2)$ along different sections for $\mu = 5$.

### 6.4.2 Proposition of a new algorithm

To prevent this phenomenon, we now propose a new process that will allow a better estimation of the real maximum of the non-exponentialized likelihood. We will still use the exponential trick but using an adaptive $\sigma$ along the iterations. The goal is to allow the estimate to escape from its




**(a)** Results for $\beta_1$

**(b)** Results for $\beta_2$

**Figure 6.7** The red line represents the theoretical value towards which the algorithm is supposed to converge. The red points are the means of the parameters estimated over 100 iterations with their standart deviations represented by the red zone. In dotted blue is the maximum likelihood of the initial model. In magenta, the theoretical limit towards which the parameter converges when $\sigma$ goes to infinity. Finally, the green cross represents the value returned while varying the variance of the exponentialized model throughout the algorithm.

initial value while converging towards a point closer to the true maximum.

We propose to first run the algorithm with $\sigma = 1$ for a certain number $m$ of iterations and then reduce the value of $\sigma$ by multiplying it by $0.9$. We iterate this process every $m$ iterations until the difference between successive parameters estimated is sufficiently small. We then let the algorithm converge with this small value of $\sigma$. This may be seen as launching the algorithm several times with an initialization closer and closer to the true maximum likelihood. While the algorithm will not converge towards the real maximum likelihood estimate as $\sigma$ is still positive during the last iterations, the error should be smaller than before as $\sigma$ has been significantly reduced.

To test this new algorithm, we launch this process a hundred times. We present the means and variances of the estimated parameters in Table 6.1 and as green crosses in figure 6.7. If we do not reach the maximum likelihood of the initial model, the error for $\beta_2$ is now smaller: $1.04\%$ while it was at least $2.6\%$ without reducing the variance throughout the algorithm. As for $\beta_1$, the error is of the same order as before.

**Remark 6.4.2.** *In Kuhn and Lavielle (2005), the authors use this model and algorithm on a real dataset for different values of $\sigma$. They conclude that the estimation of $(\beta_1, \beta_2)$ does not seem to depend on the choice of $\sigma$. In fact, for the particular values of this real dataset, the likelihood is*

| Mean of $\bar{\beta}_1$ | Variance of $\bar{\beta}_1$ | Mean of $\bar{\beta}_2$ | Variance of $\bar{\beta}_2$ |
|---|---|---|---|
| $6.21$ | $0.10$ | $32.62$ | $0.26$ |

**Table 6.1** Mean and variance of the parameters estimated while reducing the variance throughout the algorithm. To be compared with the maximum likelihood of the non-exponentialized model reached for $\beta_1 = 6.3$ and $\beta_2 = 32.28$.

*practically symmetric around its maximum. Hence, the error made in that case is indeed small for any $\sigma$.*

# Conclusion

In this paper, we have proved that the exponentialization process does not converge in general towards the maximum likelihood of the initial model using the SAEM or SAEM-MCMC algorithm. If the error converges towards $0$ when $\sigma$ goes to $0$, it is numerically impossible to take $\sigma$ too small as the algorithm is numerically never able to converge. To overcome this problem, we propose a new numerical scheme consisting in launching the algorithm several times while making the variance of the exponentialized model decrease. Thanks to out theoretical results, we show that this new process converges towards a better estimation of the maximum of likelihood of the initial model, as verified by the numerical simulations. Hence, we are able to approach the exact maximum likelihood even in the case where our likelihood does not belong to the curved exponential family.

# Annex: Proof of theorem 6.3.2 for $m \geq 2$

*Proof.* We prove the theorem 6.3.2 in the case $m \geq 2$ and $l = 2$. For $l \geq 3$, the proof could be obtained using Taylor Lagrange formula at a higher order.

We recall the following equation verified by the limit (theorem 1):

$$\int_{\mathbb{R}^m} v_k g(\theta_\infty, \bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv = 0\,, \tag{6.11}$$

# First step: $d(\bar{\psi}_\sigma, \mathcal{L}) \underset{\sigma \to 0}{\longrightarrow} 0$

We suppose that $d(\bar{\psi}_\sigma, \mathcal{L})$ does not converge towards 0. Even if it means extracting, we can suppose that, $\exists c > 0, \forall \sigma > 0, d(\bar{\psi}_\sigma, \mathcal{L}) > 3c$. As for the one-dimensional proof, we forget the $\theta$ in $g$ and write $g(\psi) = g(\theta_\infty, \psi)$. We also set $K_0 = K \setminus \{y \,|\, d(y, \mathcal{L}) < c\}$.

We want to show that

$$\exists c_0 > 0, \exists c_1 > 0, \forall y \in K_0, \exists 1 \leq i \leq m, \forall x \text{ verifying } ||x - y|| \leq c_1, |\frac{\partial g}{\partial \psi_i}(x)| > c_0\,. \tag{6.12}$$

By contradiction, we can take $c_0 = 1/n$ and extract a converging subsequence in the compact $K_0$ to find:

$$\forall c_1 > 0, \exists y \in K_0, \forall 1 \leq i \leq m, \exists x \text{ verifying } ||x - y|| \leq c_1, |\frac{\partial g}{\partial \psi_i}(x)| = 0\,.$$

However, because $y \notin \mathcal{L}$, there exists $1 \leq j \leq m$ such that $|\frac{\partial g}{\partial \psi_j}(y)| \neq 0$. If we take $c_1$ small enough then, for all $x$ such that $||y - x|| \leq c_1$, $|\frac{\partial g}{\partial \psi_j}(x)| \neq 0$ and we find a contradiction. Hence, the condition (6.12) is verified.

Hence,

$$\exists c_0 > 0, \exists c_1 > 0, \forall \sigma > 0, \exists k \in [|1, m|], \forall v \text{ verifying } ||v - \bar{\psi}_\sigma|| \leq c_1, |\frac{\partial g}{\partial \psi_k}(v)| > c_0\,.$$

As for the proof in dimension 1, we split up our integral in two parts: $I_1 = \{v \,|\, \forall i \in [|1, n|], v_i \leq c_2\}$ and $I_2 = \{v \,|\, \exists i \in [|1, n|], v_i \geq c_2\}$ where $c_2$ is chosen such that $\{v \,|\, \forall i \in [|1, n|], v_i \leq c_2\} \subset \{v \,|\, ||v|| \leq c_1\}$.

First, on $I_2$,

$$\begin{aligned}
\int_{I_2} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv &\geq \int_{I_2, v_k \leq 0} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv \\
&\geq ||g||_\infty \left( \int_{\mathbb{R}^m, v_k \leq 0} v_k \exp(-\frac{||v||^2}{2\sigma^2}) dv - \int_{I_1, v_k \leq 0} v_k \exp(-\frac{||v||^2}{2\sigma^2}) dv \right) \\
&\geq -\sigma^2 (\sqrt{2\pi}\sigma)^{m-1} ||g||_\infty \left( 1 - \left( 1 - \exp(-\frac{c_2^2}{2\sigma^2}) \right) \text{erf} \left( \frac{c_2}{\sqrt{2}\sigma} \right)^{m-1} \right)
\end{aligned}$$

where erf is the error function defined by $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$.

We now integrate on $I_1$. We write $v^{-k}$ the vector such that, $\forall i \neq k, (v^{-k})_i = v_i$ and $(v^{-k})_k = -v_k$. Then, using the mean value theorem, we have

$$\begin{aligned}\int_{I_1} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv &= \int_{I_1, v_k \leq 0} v_k (g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma + v^{-k})) \exp(-\frac{||v||^2}{2\sigma^2}) dv \\ &= \int_{I_1} 2v_k^2 \frac{\partial g}{\partial \psi_k}(\tilde{\psi}_v) \exp(-\frac{||v||^2}{2\sigma^2}) dv\end{aligned}$$

where, for $i \neq k$, $(\tilde{\psi}_v)_i = (\bar{\psi}_\sigma)_i + v_i$ and $(\tilde{\psi}_v)_k \in [(\bar{\psi}_\sigma)_k - v_k, (\bar{\psi}_\sigma)_k + v_k]$. But we know that $\frac{\partial g}{\partial \psi_k}$ does not cancel on $I_1$. Hence, it is either positive or negative. If it is positive, we find:

$$\begin{aligned}\int_{I_1} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv &\geq 2c_0 \int_{I_1, v_k \leq 0} v_k^2 \exp(-\frac{||v||^2}{2\sigma^2}) dv \\ &\geq 2c_0 (\sqrt{2\pi}\sigma)^{m-1} \sigma^2 \left[ -c_2 \exp(-\frac{c_2^2}{2\sigma^2}) + \sigma \frac{\sqrt{\pi}}{2} \text{erf}(\frac{c_2}{\sqrt{2}\sigma}) \right] \text{erf}\left(\frac{c_2}{\sqrt{2}\sigma}\right)^{m-1}\end{aligned}$$

Finally, using (6.7), we have:

$$\begin{aligned}&2c_0 \left[ -\frac{c_2}{\sigma} \exp(-\frac{c_2^2}{2\sigma^2}) + \frac{\sqrt{\pi}}{2} \text{erf}(\frac{c_2}{\sqrt{2}\sigma}) \right] \text{erf}\left(\frac{c_2}{\sqrt{2}\sigma}\right)^{m-1} \\ &\qquad \leq |g||_\infty \left[ \frac{1 - \text{erf}\left(\frac{c_2}{\sqrt{2}\sigma}\right)^{m-1}}{\sigma} + \frac{1}{\sigma} \exp(-\frac{c_2^2}{2\sigma^2}) \text{erf}\left(\frac{c_2}{\sqrt{2}\sigma}\right)^{m-1} \right]\end{aligned} \tag{6.13}$$

But,

$$\text{erf}(x) =_{x \to \infty} 1 - \frac{\exp(-x^2)}{\sqrt{\pi}x} + o\left(\frac{\exp(-x^2)}{x}\right).$$

Hence, when $\sigma$ goes to $0$, the left-hand side of the inequality goes to $\sqrt{\pi}c_0$ while the right-hand side goes to 0. We thus find a contradiction.

Finally, if $\frac{\partial g}{\partial \psi_k}$ is not positive on $I_1$ (as supposed here) but negative, we can use the same method to find an upper bound on the integral on $I_2$ and on the integral on $I_1$. We would then find the same inequality as in (6.13).

Hence, in all cases, we have proved that $d(\bar{\psi}_\sigma, \mathcal{L}) \xrightarrow[\sigma \to 0]{} 0$.

# Second step: Choice of the basis

The upper bound using second derivatives is more complex to obtain for $m > 1$ as crossed partial derivatives appear that can be either positive or negative. To control those parts, the choice of the compact is more complex. We will first show that we can express our vector $v$ and our function $g$ in any orthonormal basis and still have the equation (6.7).

Indeed, let $P$ be a change-of-basis matrix. Then, because the equation (6.7) is linear on $v_k$ and true for all $k \in [|1, m|]$, we still have:

$$\int_{\mathbb{R}^m} (Pv)_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv = 0\,.$$

Using the change of variable $u = Pv$, we then find, for any $k \in [|1, m|]$:

$$\int_{\mathbb{R}^m} u_k g(P^{-1}(P\bar{\psi}_\sigma + u)) \exp(-\frac{||u||^2}{2\sigma^2}) dv = 0\,.$$

We write $h : u \mapsto g(P^{-1}u)$. Hence, $h$ verifies the equation (6.7).

We can thus choose to express our function $g$ in any base. In particular, we write $\psi_M$ the nearest maximum of $\bar{\psi}_\sigma$. Then, the Hessian of $g$ at $\psi_M$ is a negative symmetric matrix. Hence, it is diagonal in an orthonormal basis. We choose to express $g$ in that basis. With a change of notation, we can hence assume that the hessian of $g$ at $\psi_M$ is diagonal. In particular, for all $i \neq j \in [|1, m|]$,

$$\frac{\partial^2 g}{\partial \psi_i^2}(\psi_M) < 0 \text{ and } \frac{\partial^2 g}{\partial \psi_i \partial \psi_j}(\psi_M) = 0\,.$$

In particular, we are now able to impose a condition between the second derivatives of $g$ on a compact centered around $\psi_M$. There exists $K_0$ compact such that,

$$\begin{aligned} -\sup_{K_0} \frac{\partial^2 g}{\partial \psi_k^2} >& (m-1) \sup \left\{ \frac{\partial^2 g}{\partial \psi_k \partial \psi_j}(v) \,\middle|\, v \in K_0, j \neq k, \frac{\partial^2 g}{\partial \psi_k \partial \psi_j}(v) > 0 \right\} \\ & - \frac{m-1}{2} \inf \left\{ \frac{\partial^2 g}{\partial \psi_k \partial \psi_j}(v) \,\middle|\, v \in K_0, j \neq k, \frac{\partial^2 g}{\partial \psi_k \partial \psi_j}(v) < 0 \right\} \end{aligned} \tag{6.14}$$

## Third step: Search of the upper bound

As for the proof in 1D, we will split our integral into two parts and say that neither can be too big for the complete integral to be equal to 0. More precisely, for $\sigma > 0$, let $k$ be the coordinate such that $|(\bar{\psi}_\sigma)_k - (\psi_M)_k| = \max |(\bar{\psi}_\sigma)_i - (\psi_M)_i|$. We write for $i \in [|1, m|]$, $(\alpha_\sigma)_i = |(\bar{\psi}_\sigma)_i - (\psi_M)_i|$. The goal is to show that $(\alpha_\sigma)_k$ goes to 0 when $\sigma$ goes to 0.

Let $c > 0$ and $\sigma$ small enough such that

$$I_1 := \left\{ v \in \mathbb{R}^m \,|\, v_k \in [-(\alpha_\sigma)_k, (\alpha_\sigma)_k] \text{ and, for } i \neq k, v_i \in [-c, c] \right\} \subset K_0\,.$$

On $I_1^c$, we use the same upper bounds as in the first step to find:

$$\begin{aligned} \int_{I_1^c} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv \geq & -\sigma^2 (\sqrt{2\pi}\sigma)^{m-1} ||g||_\infty \\ & \cdot \left( 1 - \left( 1 - \exp(-\frac{(\alpha_\sigma)_k^2}{2\sigma^2}) \right) \text{erf} \left( \frac{c}{\sqrt{2}\sigma} \right)^{m-1} \right) \end{aligned} \tag{6.15}$$

On $I_1$ we will use once again the mean value theorem, first between $\bar{\psi}_\sigma + v^{-k}$ and $\bar{\psi}_\sigma + v$ to find $\tilde{\psi}_v \in K_0$ such that, for $i \neq k$, $(\tilde{\psi}_v)_i = (\bar{\psi}_\sigma + v)_i$, $(\tilde{\psi}_v)_k \in [(\bar{\psi}_\sigma - v)_k, (\bar{\psi}_\sigma + v)_k]$ and

$$g(\bar{\psi}_\sigma + v) - g(\bar{\psi}_\sigma + v^{-k}) = 2v_k \frac{\partial g}{\partial \psi_k}(\tilde{\psi}_v)$$

and then between $\tilde{\psi}_v$ and $\psi_M$ to find $\tilde{\psi}_v^1 \in K_0$ such that:

$$\frac{\partial g}{\partial \psi_k}(\tilde{\psi}_v) = \sum_{i=1}^{m} (\tilde{\psi}_v - \psi_M)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1)\,.$$

Even if it means changing basis, we can assume that, $\forall i \in [|1,m|], (\alpha_\sigma)_i = |(\bar{\psi}_\sigma)_i - (\psi_M)_i| = (\psi_M)_i - (\bar{\psi}_\sigma)_i$ without modifying the hypothesis (6.14).

The difficulty to find upper bounds is that $\frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1)$ and also $(\tilde{\psi}_v - \psi_M)_i$ can be either positive or negative.

Using those previous equalities and the facts that $\frac{\partial^2 g}{\partial \psi_k^2} < 0$ on $K_0$ and $(\tilde{\psi}_v - \psi_M)_k \leq (\bar{\psi}_\sigma + v - \psi_M)_k = (v - \alpha_\sigma)_k$, we have:

$$\int_{I_1} v_k g(\bar{\psi}_\sigma + v) \exp(-\frac{||v||^2}{2\sigma^2}) dv \geq 2 \int_{I_1, v_k \geq 0} v_k^2 \sum_{i=1}^{m} (v - \alpha_\sigma)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) dv$$

We will study the different terms of the sum differently according to $i = k$, or $i \neq k$.

First, for $i = k$, using the fact that $(v - \alpha_\sigma)_k \leq 0$ on $I_1$, we can compute the integral using integration per part and the function erf defined above to find:

$$\begin{aligned}\int_{I_1, v_k \geq 0} 2v_k^2 (v - \alpha_\sigma)_k \frac{\partial^2 g}{\partial \psi_k^2}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) dv &\geq \sup_{K_0} \left(\frac{\partial^2 g}{\partial \psi_k^2}\right) \int_{I_1, v_k \geq 0} 2v_k^2 (v - \alpha_\sigma)_k \exp(-\frac{||v||^2}{2\sigma^2}) dv \\ = -2 \sup_{K_0} \left(\frac{\partial^2 g}{\partial \psi_k^2}\right) (\sqrt{2\pi}\sigma)^{m-1} \text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1} (\alpha_\sigma)_k^4 \Bigg[ &\sqrt{\frac{\pi}{2}} \left(\frac{\sigma}{(\alpha_\sigma)_k}\right)^3 \text{erf}\left(\frac{(\alpha_\sigma)_k}{\sqrt{2}\sigma}\right) \\ -2 \left(\frac{d}{(\alpha_\sigma)_k}\right)^4 &\left(1 - \exp(-\frac{(\alpha_\sigma)_k}{2\sigma^2})\right)\Bigg]\end{aligned}$$

For $i \neq k$, we do similar computations remarking that:

- if $\frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) > 0$ and $(v - \alpha_\sigma)_i > 0$,

$$2v_k^2 (v - \alpha_\sigma)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) > 0$$

- if $\frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) > 0$ and $(v - \alpha_\sigma)_i < 0$, with $K_0^+ = \left\{v \in K_0 | \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(v) > 0\right\}$,

$$2v_k^2 (v - \alpha_\sigma)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) > 2v_k^2 (v - \alpha_\sigma)_i \exp(-\frac{||v||^2}{2\sigma^2}) \sup_{K_0^+, i} \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}$$

- if $\frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) < 0$ and $(v - \alpha_\sigma)_i < 0$

$$2v_k^2 (v - \alpha_\sigma)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) > 0$$

- if $\frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) < 0$ and $(v - \alpha_\sigma)_i > 0$, with $K_0^- = \left\{v \in K_0 | \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(v) < 0\right\}$,

$$2v_k^2 (v - \alpha_\sigma)_i \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}(\tilde{\psi}_v^1) \exp(-\frac{||v||^2}{2\sigma^2}) > 2v_k^2 (v - \alpha_\sigma)_i \exp(-\frac{||v||^2}{2\sigma^2}) \inf_{K_0^-, i} \frac{\partial^2 g}{\partial \psi_k \partial \psi_i}$$

Hence, for $i \neq k$, we write $I_1^- = \{v \in I_1 \,|\, v_k \geq 0, v_i \leq (\alpha_\sigma)_i\}$ and:

$$\begin{aligned}
\int_{I_1^-} 2v_k^2(v-\alpha_\sigma)_i \frac{\partial^2 g}{\partial\psi_k\partial\psi_i}(\tilde{\psi}_v^1)\exp(-\frac{||v||^2}{2\sigma^2})dv &\geq \sup_{K_0^+,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}\int_{I_1^-} 2v_k^2(v-\alpha_\sigma)_i\exp(-\frac{||v||^2}{2\sigma^2})dv \\
&= 2\sup_{K_0^+,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}(\sqrt{2\pi}\sigma)^{m-1}\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}(\alpha_\sigma)_k^4 \\
&\quad\cdot\left[\sqrt{\frac{\pi}{2}}\left(\frac{\sigma}{(\alpha_\sigma)_k}\right)^4\text{erf}\left(\frac{(\alpha_\sigma)_k}{\sqrt{2}\sigma}\right)-\left(\frac{\sigma}{(\alpha_\sigma)_k}\right)^3\exp\left(-\frac{(\alpha_\sigma)_k^2}{2\sigma^2}\right)\right] \\
&\quad\cdot\left[\frac{\exp\left(\frac{-c^2}{2\sigma^2}\right)-\exp\left(\frac{-(\alpha_\sigma)_j^2}{2\sigma^2}\right)}{\sqrt{2\pi}\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)}-\frac{1}{2}\frac{(\alpha_\sigma)_j}{d}\left(1+\frac{\text{erf}\left(\frac{(\alpha_\sigma)_j}{\sqrt{2}\sigma}\right)}{\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)}\right)\right]
\end{aligned}$$

Similarly, with $I_1^+ = \{v \in I_1 \,|\, v_k \geq 0, v_i \geq (\alpha_\sigma)_i\}$,

$$\begin{aligned}
\int_{I_1^+} 2v_k^2(v-\alpha_\sigma)_i \frac{\partial^2 g}{\partial\psi_k\partial\psi_i}(\tilde{\psi}_v^1)\exp(-\frac{||v||^2}{2\sigma^2})dv &\geq \inf_{K_0^-,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}\int_{I_1^+} 2v_k^2(v-\alpha_\sigma)_i\exp(-\frac{||v||^2}{2\sigma^2})dv \\
&= 2\inf_{K_0^-,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}(\sqrt{2\pi}\sigma)^{m-1}\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}(\alpha_\sigma)_k^4 \\
&\quad\cdot\left[\sqrt{\frac{\pi}{2}}\left(\frac{\sigma}{(\alpha_\sigma)_k}\right)^4\text{erf}\left(\frac{(\alpha_\sigma)_k}{\sqrt{2}\sigma}\right)-\left(\frac{\sigma}{(\alpha_\sigma)_k}\right)^3\exp\left(-\frac{(\alpha_\sigma)_k^2}{2\sigma^2}\right)\right] \\
&\quad\cdot\left[\frac{\exp\left(\frac{-(\alpha_\sigma)_j^2}{2\sigma^2}\right)-\exp\left(\frac{-c^2}{2\sigma^2}\right)}{\sqrt{2\pi}\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)}-\frac{1}{2}\frac{(\alpha_\sigma)_j}{d}\left(1-\frac{\text{erf}\left(\frac{(\alpha_\sigma)_j}{\sqrt{2}\sigma}\right)}{\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)}\right)\right]
\end{aligned}$$

Those three upper bounds are quite complex, but we can remark that they can be written as $d^{m-1}\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}(\alpha_\sigma)_k^4 h\left(\frac{(\alpha_\sigma)_k}{\sigma}\right)$.
We will now see that this function $h$ is strictly positive at infinity.

Indeed, using all the previous upper bounds presented previously, equation (6.7) and using the fact that $\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right) \geq 1/2$ for $\sigma$ small enough, we can write:

$$(\alpha_\sigma)_k^4\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}\left(h\left(\frac{\alpha_k}{\sigma}\right)-\frac{1-\text{erf}\left(c/\sqrt{2}\sigma\right)}{\sigma^2\text{erf}\left(c/\sqrt{2}\sigma\right)}\right) \leq \frac{||g||_\infty}{2}\sigma^2$$

with:

$$\begin{aligned}
h(x) = \frac{e^{x^2/2}}{x^4}&\left[-\sup_{K_0}\frac{\partial^2 g}{\partial\psi_k^2}\left(\frac{\sqrt{\pi}}{2}x\text{erf}(x/\sqrt{2})-2(1-e^{-x^2/2})\right)\right. \\
&+\left(\frac{\sqrt{\pi}}{2}x\text{erf}(x/\sqrt{2})-x^2e^{-x^2/2}\right)\cdot\left[(m-1)(\inf_{K_0^-,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}-\sup_{K_0^+,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i})(\frac{2}{\sqrt{2\pi}x}+\frac{1}{2})\right. \\
&\left.\left.-\frac{m-1}{2}\sup_{K_0^+,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}\right]\right]
\end{aligned}$$

is a function independent of $\alpha_\sigma$ and $\sigma$.

In particular, when $x$ goes to infinity, $h(x)$ is equivalent to:

$$\frac{e^{x^2/2}}{x^4}\frac{\sqrt{\pi}}{2}x\left[-\sup_{K_0}\frac{\partial^2 g}{\partial\psi_k^2}+\frac{m-1}{2}\inf_{K_0^-,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}-(m-1)\sup_{K_0^+,i}\frac{\partial^2 g}{\partial\psi_k\partial\psi_i}\right]$$

But, according to the hypothesis done on the compact $K_0$, this is strictly positive.

Hence, there exist $c_0>0, c_1>0$ such that, if $x\geq c_1$, $h(x)>c_0>0$.
We will now suppose that $(\alpha_\sigma)_k \geq c_1\sigma$. Then, $h\left(\frac{\alpha_k}{\sigma}\right)\geq c_0>0$. Moreover,

$$\frac{1-\text{erf}\left(c/\sqrt{2}\sigma\right)}{\sigma^2\text{erf}\left(c/\sqrt{2}\sigma\right)}\xrightarrow[\sigma\to 0]{}0\,.$$

So, for $\sigma$ small enough, it is smaller than $c_0/2$ and we finally find:

$$(\alpha_\sigma)_k^4\leq c_0||g||_\infty\frac{\sigma^2}{\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}}$$

Using the fact that $\text{erf}\left(\frac{c}{\sqrt{2}\sigma}\right)^{m-1}\xrightarrow[\sigma\to 0]{}1$ gives us finally the existence of a constant $c>0$ such that

$$d(\bar{\psi}_\sigma,\mathcal{L})\leq(c_1\sigma)\vee(c\sqrt{\sigma})$$

which allows us to conclude for $\sigma$ small enough.

**Fourth step: Approximation when $\sigma$ goes to infinity**

The last step follows the exact same steps as for $m=1$. It is copied here.
We use again the equation (6.7). For all $\sigma\in\mathbb{R}$, $\forall 1\leq k\leq m$,

$$\int_{\mathbb{R}^m}v_k g(\bar{\psi}_\sigma+v)\exp(-\frac{||v||^2}{2\sigma^2})dv=0$$

Using the change of variable $\bar{\psi}_\sigma+v$, we find:

$$(\bar{\psi}_\sigma)_k=\frac{\int_{\mathbb{R}^m}v_k g(v)\exp\left(-\frac{||v-\bar{\psi}_\sigma||^2}{2\sigma^2}\right)dv}{\int_{\mathbb{R}^m}g(v)\exp\left(-\frac{||v-\bar{\psi}_\sigma||^2}{2\sigma^2}\right)dv}$$

But $\bar{\psi}_\sigma$ is supposed to stay in a compact so, $\forall v\in\mathbb{R}^m$, $\exp\left(-\frac{||v-\bar{\psi}_\sigma||^2}{2\sigma^2}\right)\xrightarrow[\sigma\to\infty]{}1$. Using the integrability of $g$ and $v\mapsto v_k g(v)$, it is easy to conclude using the dominated convergence theorem.

□

# Part IV

# Conclusion and perspectives

In this thesis, we focused on different aspects of shape analysis and statistical optimization. In this final part, we will review the contributions we have done in those fields and the questions that have not yet been closed.

## Analysis of longitudinal Riemannian manifold valued data with multiple dynamics.

In the chapter 3, we generalized the model proposed in Schiratti et al. (2015) in the case where the subjects can have different dynamics on different time intervals. Building on the work of Chevallier et al. (2017) done in dimension 1, we proposed to model the representative trajectory as a piecewise geodesic by introducing rupture times at which the population goes from one dynamic to another. This allows to represent more complex dynamics such as the behaviour of tumors during chemotherapy or the apparition of a disease after the beginning of the observations. Moreover, it is possible to allow different clusters of the population to merge or branch at the rupture times. The clusters will then have common dynamics on certain time intervals but split up on others. We thus introduced unsupervised clustering and used a scheme of temperature to achieve numerical convergence. This new model allows to considerate deviations of behaviours between subjects along time. For example, one can consider the efficiency of a treatment for certain subjects as a deviation from the usual behaviour of sick patients.

If this model can be applied to a large number of situations, it has the disadvantage to require the knowledge of lots of hyperparameters. One must choose the number of clusters, for each of them, the number of rupture times and if the population merges or branches. The choice of those hyperparameters will then dramatically change the clusters and dynamics estimated. This problem could be solved by selecting models. This would have the advantage to allow to test different assumptions on a particular data set. One could ask itself if there is or not different clusters in a population, if there is a change of dynamic following a particular event, etc.. An attempt has been made chapter 3 by using a Bayesian Information Criterion. However, in this large dimension framework, model selection is not an easy question and would require further investigation.

Another question arising from chapter 3 is the convergence of a tempered model under a scheme of temperature depending of the current state of the algorithm. Indeed, the particular scheme of temperature used chapter 3 depends on the current state of the Markov Chain to reach an optimal acceptation rate in the Metropolis Hastings algorithm. If the convergence of tempered SAEM has been studied in Allassonnière and Chevallier (2019), their assumptions do not include a state-dependent temperature. Further work would thus be required to prove the theoretical convergence of the algorithm proposed.

## Detection of anomalies using the LDDMM framework.

In the chapter 4, we applied the LDDMM framework to the detection of anomalies. An anomaly is then defined as what cannot be obtained as the diffeomorphic deformation of a control template. This can in particular be applied to the detection of tumors. Indeed, tumors appear as structures with a different grey level on images. Hence, those tumors cannot be retrieved by a diffeomorphic deformation of a control template.
More precisely, in that chapter, we chose to represent the observations as the sum of a diffeomorphic deformation of a template and a sparse matrix, and we estimate both at the same time.

Our method is particularly interesting for its versatility. One does not require any large data sets nor any annotation by doctors. If one only disposes of one sane subject and one sick subject, it is still possible to retrieve anomalies, even if the precision is improved using a template of control subjects.

Improvements could still be applied to the method, particularly on the post-process of the anomaly matrix. Indeed, in addition to the real anomalies, small errors of reconstruction are retrieved, particularly on the border of the considered object. Further investigation should be made to only retrieve the real anomalies without missing some.

One could also try to add a longitudinal layer to the model. The goal would then be to estimate the temporal trajectory of the tumors. This would allow to predict their future, if they will shrink or not and to adapt the treatment according to the prediction. However, as new tumors can appear or disappear along time, a diffeomorphic deformation from an initial time point, as done chapter 3, would not be relevant. At the time of writing of this manuscript, this work is still in progress.

## On the convergence of Stochastic Approximations with a subgeometric Markovian dynamic.

To estimate the parameters of the mixed effect models with incomplete data, we used, in the chapter 3, the Monte Carlo Markov Chain Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm. This variant of the EM uses Stochastic Approximations to compute the Expectation step while using only one realization of the conditional distribution. Based on Stochastic Approximations with Markovian dynamics, the proof of convergence of the MCMC SAEM algorithm (Allassonnière et al., 2010) takes up the assumptions of the proof of convergence of SA (Andrieu et al., 2005). In particular, one needs to verify the geometric ergodicity of the Markov Chain. In practice, this assumption limits the conditional distributions one can consider. In particular, when sampling from a heavy tail distribution using a Metropolis Hastings algorithm, the resulting Markov Chain can be subgeometric.
To overcome this difficulty, we proposed chapter 5 a new set of hypotheses asking only the subgeometric ergodicity of the Markov Chain. Those assumptions express a compromise between the rate of convergence of the Markov Chain and its regularity with respect to a control function. We showed that they are verified in several applications where the initial theorem could not have been applied.

## Creation of a new EM algorithm for non curved exponential distributions.

Using the SAEM algorithm, one needs to compute a stochastic approximation of the sufficient statistics. This, in particular, supposes that such sufficient statistics exist i.e. that our joint distribution belongs to the curved exponential family. However, this hypothesis is not verified for lots of realistic problems. For instance, if one supposes that

$$\begin{cases} y_i \sim \mathcal{N}(f(\psi_i, \alpha), \Sigma_0) \\ \psi_i \sim \mathcal{N}(g(C_i, \beta), \Sigma_1) \end{cases}$$

where $\theta = (\alpha, \beta)$ are fixed effects, $\psi_i$ are random individual parameters, $C_i$ are hyperparameters and $f$ and $g$ are non linear, the model is not curved exponential.

To overcome this difficulty, Kuhn and Lavielle (2005) proposed to transform the model to make it curved exponential. Their method consists in considering the parameters $\theta$ of the initial model as additional latent variables following a Normal distribution centered on a new parameter $\bar{\theta}$ with fixed variance $\sigma^2$. Instead of estimating $\theta$, they estimate the mean $\bar{\theta}$. This new model is now curved exponential and it is this modification we applied chapter 3 to estimate the population parameters. However, there is no guarantee that the new maximum likelihood is close to the one of the initial model. In chapter 6, we exhibit an example where the two maxima are different. We also prove a theorem giving an upper bound on their distance for $\sigma$ small. If the second derivative of the incomplete likelihood at its maxima is not $0$, this distance decreases in $\sqrt{\sigma}$.
By verifying our results on an example, we see that, for $\sigma$ small, the convergence of the SAEM applied to the modified model is not achieved numerically. Hence, we propose a new algorithm, where one decreases the value of $\sigma$ along iterations, achieving a better estimation of the maximum likelihood of the initial model.

If we are now able to give a better estimation of the maximum likelihood of a non curved exponential model, we still do not have an algorithm converging almost surely towards its value. When confronted to the problem, one could choose to transform the SAEM algorithm to apply it to the exact model. An idea could be to replace the stochastic approximation and maximization steps:

$$\begin{cases} Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left( \log p(y, \psi^{(k)}, \theta) - Q_{k-1}(\theta) \right) \\ \theta_k = \text{argmax}_\theta \, Q_k(\theta) \end{cases} \tag{6.16}$$

by:

$$\begin{cases} \hat{\theta}(y, \psi^{(k)}) = \text{argmax}_\theta \, p(y, \psi^{(k)}, \theta) \\ \theta_k = \theta_{k-1} + \gamma_k \left( \hat{\theta}(y, \psi^{(k)}) - \theta_{k-1} \right) . \end{cases} \tag{6.17}$$

While in the first case, one can often find a closed form of the maximization step using the sufficient statistics, it is no longer the case in the second case and one needs to use gradient descent algorithms. However, we are no longer in the curved exponential framework and the question of the convergence of this SAEM algorithm is still open.

# Bibliography


Mohamed F. Abdelkader, Wael Abd-Almageed, Anuj Srivastava, and Rama Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on Riemannian shape manifolds. *Computer Vision and Image Understanding*, 2011. ISSN 10773142. doi: 10.1016/j.cviu.2010.10.006.

Jinane Abounadi, Dimitri P Bertsekas, and Vivek Borkar. Stochastic approximation for nonexpansive maps: Application to q-learning algorithms. *SIAM Journal on Control and Optimization*, 41(1):1–22, 2002.

Oodally Ajmal, Luc Duchateau, and Estelle Kuhn. Convergent stochastic algorithm for parameter estimation in frailty models using integrated partial likelihood. *arXiv preprint arXiv:1909.07056*, 2019.

Stéphanie Allassonnière and Juliette Chevallier. A new class of em algorithms. escaping local minima and handling intractable sampling. 2019.

Stéphanie Allassonnière and Estelle Kuhn. Stochastic algorithm for bayesian mixture effect template estimation. *ESAIM: Probability and Statistics*, 14:382–408, 2010.

Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

Stéphanie Allassonnière, Estelle Kuhn, Alain Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.

Stéphanie Allassonniere, Laurent Younes, et al. A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160, 2012.

Stéphanie Allassonnière, Stanley Durrleman, and Estelle Kuhn. Bayesian mixed effect atlas estimation with a diffeomorphic deformation model. *SIAM Journal on Imaging Sciences*, 8(3): 1367–1395, 2015.

Stéphanie Allassonniere, Juliette Chevallier, and Stephane Oudard. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *Advances in Neural Information Processing Systems*, pages 1152–1160, 2017.

Christophe Andrieu and Christian P Robert. *Controlled MCMC for optimal sampling*. INSEE, 2001.

Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization*, 44(1):283–312, 2005.

Yves Atchadé, Gersende Fort, et al. Limit theorems for some adaptive mcmc algorithms with subgeometric kernels. *Bernoulli*, 16(1):116–154, 2010.

Yves F Atchadé, Gersende Fort, et al. Limit theorems for some adaptive mcmc algorithms with subgeometric kernels: Part ii. *Bernoulli*, 18(3):975–1001, 2012.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Marian Stewart Bartlett, Javier R Movellan, and Terrence J Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on neural networks*, 13(6):1450–1464, 2002.

Douglas M Bates and Donald G Watts. *Nonlinear regression analysis and its applications*, volume 2. Wiley New York, 1988.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Sébastien Benzekry, Clare Lamont, Afshin Beheshti, Amanda Tracz, John ML Ebos, Lynn Hlatky, and Philip Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol*, 10(8):e1003800, 2014.

Rabi Bhattacharya, Vic Patrangenaru, et al. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of statistics*, 31(1):1–29, 2003.

Jonathan Boisvert, Xavier Pennec, Hubert Labelle, Farida Cheriet, and Nicholas Ayache. Principal spine shape deformation modes using riemannian geometry and articulated models. In *International Conference on Articulated Motion and Deformable Objects*, pages 346–355. Springer, 2006.

Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9271–9280, 2018a.

Alexandre Bône, Maxime Louis, Benoît Martin, and Stanley Durrleman. Deformetrica 4: an open-source software for statistical shape analysis. In *International Workshop on Shape in Medical Imaging*, pages 3–13. Springer, 2018b.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Chaim Broit. Optimal registration of deformed images. 1981.

Vince D Calhoun, Tulay Adali, Godfrey D Pearlson, and James J Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001a.

Vince Daniel Calhoun, T Adali, VB McGinty, James J Pekar, TD Watson, and GD Pearlson. fmri activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage*, 14(5):1080–1088, 2001b.

Gilles Celeux. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2:73–82, 1985.

Rudrasis Chakraborty, Vikas Singh, Nagesh Adluru, and Baba C Vemuri. A geometric framework for statistical analysis of trajectories with distinct temporal spans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 172–181, 2017.

Nicolas Charon and Alain Trouvé. The varifold representation of nonoriented shapes for diffeomorphic registration. *SIAM Journal on Imaging Sciences*, 6(4):2547–2580, 2013.

Han-Fu Chen. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2006.

Han-Fu Chen, Lei Guo, and Ai-Jun Gao. Convergence and robustness of the robbins-monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications*, 27:217–231, 1987.

Juliette Chevallier, Stéphane Oudard, and Stéphanie Allassonnière. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *31st Conference on neural information processing systems (NIPS 2017)*, 2017.

Stéphane Chrétien and Alfred O Hero. On em algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

Gary Edward Christensen. Deformable shape models for anatomy. 1994.

Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

W Thompson d'Arcy et al. On growth and form. *On growth and form*, 1917.

Vianney Debavelaere and Stéphanie Allassonnière. On the curved exponential family in the stochatic approximation expectation maximization algorithm. 2021.

Vianney Debavelaere, Alexandre Bône, Stanley Durrleman, and Stéphanie Allassonnière. Clustering of longitudinal shape data sets using mixture of separate or branching trajectories. In *International conference on medical image computing and computer-assisted intervention*, pages 66–74. Springer, 2019.

Vianney Debavelaere, Stanley Durrleman, and Stéphanie Allassonnière. Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories. *International Journal of Computer Vision*, pages 1–16, 2020.

Vianney Debavelaere, Stanley Durrleman, and Stéphanie Allassonnière. On the convergence of stochastic approximations under a subgeometric ergodic markov dynamic. *Electronic Journal of Statistics*, 15(1):1583–1609, 2021.

Bernard Delyon, Marc Lavielle, Eric Moulines, et al. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Loïc Devilliers, Stéphanie Allassonnière, Alain Trouvé, and Xavier Pennec. Inconsistency of template estimation by minimizing of the variance/pre-variance in the quotient space. *Entropy*, 19(6):288, 2017.

Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer, 2017.

Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Raman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5):S400–S410, 2014.

Randal Douc, Gersende Fort, Eric Moulines, Philippe Soulier, et al. Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14(3):1353–1377, 2004.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

Ian L Dryden, Alexey Koloydenko, Diwei Zhou, et al. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3): 1102–1123, 2009.

Anne Dubois, Marc Lavielle, Sandro Gsteiger, Etienne Pigeolet, and France Mentré. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. *Statistics in medicine*, 30(21):2582–2600, 2011.

Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

Paul Dupuis, Ulf Grenander, and Michael I Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of applied mathematics*, pages 587–600, 1998.

Sandy Durrleman, Marcel Prastawa, Guido Gerig, and Sarang Joshi. Optimal data-driven sparse parameterization of diffeomorphisms for population analysis. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 123–134. Springer, 2011a.

Stanley Durrleman, Pierre Fillard, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Registration, atlas estimation and variability analysis of white matter fiber bundles modeled as currents. *NeuroImage*, 55(3):1073–1090, 2011b.

Stanley Durrleman, Stéphanie Allassonnière, and Sarang Joshi. Sparse adaptive parameterization of variability in image ensembles. *International Journal of Computer Vision*, 101(1): 161–183, 2013.

Stanley Durrleman, Marcel Prastawa, Nicolas Charon, Julie R Korenberg, Sarang Joshi, Guido Gerig, and Alain Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 101:35–49, 2014.

P Thomas Fletcher. Geodesic regression and the theory of least squares on riemannian manifolds. *International journal of computer vision*, 105(2):171–185, 2013.

P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8): 995–1005, 2004.

Gersende Fort. Méthodes de monte carlo et cha^ınes de markov pour la simulation. 2009.

Gersende Fort and Eric Moulines. V-subgeometric ergodicity for a hastings–metropolis algorithm. *Statistics & probability letters*, 49(4):401–410, 2000.

Gersende Fort and Eric Moulines. Polynomial ergodicity of markov transition kernels. *Stochastic Processes and their Applications*, 103(1):57–99, 2003.

Gersende Fort, Eric Moulines, et al. Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259, 2003.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

Joan Glaunes, Alain Trouvé, and Laurent Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

Ulf Grenander. *General pattern theory: A mathematical study of regular structures Oxford mathematical monographs*. Oxford University Press: Clarendon, 1993.

Jeremie Guedj and Alan S Perelson. Second-phase hepatitis c virus rna decline during telaprevir-based therapy increases with drug effectiveness: implications for treatment duration. *Hepatology*, 53(6):1801–1808, 2011.

Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Automatic computation of average brain models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–640. Springer, 1998.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Yi Hong, Nikhil Singh, Roland Kwitt, and Marc Niethammer. Group testing for longitudinal data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9123, pages 139–151. Springer Verlag, 2015. doi: 10.1007/978-3-319-19992-4_11.

Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000.

B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, J. L. Prince, et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.

Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

David G. Kendall. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, mar 1984. ISSN 00246093. doi: 10.1112/blms/16.2.81. URL http://doi.wiley.com/10.1112/blms/16.2.81.

Wilfrid S Kendall. Probability, convexity, and harmonic maps with small image i: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(2):371–406, 1990.

Hyunwoo J. Kim, Nagesh Adluru, Heemanshu Suri, Baba C. Vemuri, Sterling C. Johnson, and Vikas Singh. Riemannian nonlinear mixed effects models: Analyzing longitudinal deformations in neuroimaging. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 5777–5786. Institute of Electrical and Electronics Engineers Inc., nov 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.612.

Igor Koval, Jean-Baptiste Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stephanie Allassonniere, and Stanley Durrleman. Spatiotemporal propagation of the cortical atrophy during the course of alzheimer's disease: Population and individual patterns. *Frontiers in Neurology*, 9:235, 2018.

Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

Estelle Kuhn and Marc Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, 49(4):1020–1038, 2005.

Estelle Kuhn, Catherine Matias, and Tabea Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.

Estelle Kuhn, Catherine Matias, and Tabea Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *Statistics and Computing*, 30(6):1725–1739, 2020.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Jacques Lafontaine, Sylvestre Gallot, and Dominique Hulin. Riemannian geometry, 2004.

Kenneth Lange. A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):425–437, 1995.

Thomas Lartigue, Stanley Durrleman, and Stéphanie Allassonnière. Deterministic approximate em algorithm; application to the riemann approximation em and the tempered em. *arXiv preprint arXiv:2003.10126*, 2020.

Marc Lavielle. *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press, 2014.

Marc Lavielle and France Mentré. Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. *Journal of pharmacokinetics and pharmacodynamics*, 34(2):229–249, 2007.

Huiling Le and Alfred Kume. The fréchet mean shape and the shape of the means. *Advances in Applied Probability*, pages 101–113, 2000.

Wolfram Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

Fredrik Lindsten. An efficient stochastic approximation em algorithm using conditional particle filters. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6274–6278. IEEE, 2013.

Roderick JA Little and Donald B Rubin. The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326, 1989.

Chengjun Liu and Harry Wechsler. Independent component analysis of gabor features for face recognition. *IEEE transactions on Neural Networks*, 14(4):919–928, 2003.

Peter Lorenzen, Brad C Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–418. Springer, 2005.

Marco Lorenzi, Nicholas Ayache, and Xavier Pennec. Schild's ladder for the parallel transport of deformations in time series of images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 463–474. Springer, 2011.

Maxime Louis, Alexandre Bône, Benjamin Charlier, Stanley Durrleman, Alzheimer's Disease Neuroimaging Initiative, et al. Parallel transport in shape analysis: a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, 2017.

Jinwen Ma, Lei Xu, and Michael I Jordan. Asymptotic convergence rate of the em algorithm for gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2000.

Jun Ma, Michael I Miller, Alain Trouvé, and Laurent Younes. Bayesian template estimation in computational anatomy. *NeuroImage*, 42(1):252–261, 2008.

Iain L MacDonald and Walter Zucchini. *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press, 1997.

Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Dara Ghahremani, and Terrence J Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94(20):10979–10984, 1997.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Xiao-Li Meng et al. On the rate of convergence of the ecm algorithm. *The Annals of Statistics*, 22(1):326–339, 1994.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Cristian Meza, Felipe Osorio, and Rolando De la Cruz. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, 22(1):121–139, 2012.

Michael I. Miller and Laurent Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1-2):61–84, 2001.

Michael I Miller, Alain Trouvé, and Laurent Younes. On the metrics and euler-lagrange equations of computational anatomy. *Annual review of biomedical engineering*, 4(1):375–405, 2002.

Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006.

Washington Mio, Anuj Srivastava, and Shantanu Joshi. On shape of plane elastic curves. *International Journal of Computer Vision*, 73(3):307–324, 2007.

Prasanna Muralidharan and P Thomas Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1027–1034. IEEE, 2012.

Xavière Panhard and Adeline Samson. Extension of the saem algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10(1):121–135, 2009.

Nick Pawlowski, MC Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. *Medical Imaging with Deep Learning*, 2018.

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

Xavier Pennec et al. Barycentric subspace analysis on manifolds. *Annals of Statistics*, 46(6A): 2711–2746, 2018.

Simon Pernot, Olivier Pellerin, Pascal Artru, Carole Montérymard, Denis Smith, Jean-Luc Raoul, Christelle De La Fouchardière, Laetitia Dahan, Rosine Guimbaud, David Sefrioui, et al. Intra-arterial hepatic beads loaded with irinotecan (debiri) with mfolfox6 in unresectable liver metastases from colorectal cancer: a phase 2 study. *British Journal of Cancer*, 123(4): 518–524, 2020.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied probability*, 7(1):110–120, 1997.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

JS Rosenthal and GO Roberts. Coupling and ergodicity of adaptive mcmc. *Journal of Applied Probablity*, 44:458–475, 2007.

Adeline Samson, Marc Lavielle, and France Mentré. Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model: Application to hiv dynamics model. *Computational Statistics & Data Analysis*, 51(3):1562–1574, 2006.

Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412, 2015.

Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *The Journal of Machine Learning Research*, 18(1):4840–4872, 2017.

Hyunseok Seo, Charles Huang, Maxime Bassenne, Ruoxiu Xiao, and Lei Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE transactions on medical imaging*, 39(5):1316–1325, 2019.

Lewis B Sheiner and Stuart L Beal. Evaluation of methods for estimating population pharmacokinetic parameters. i. michaelis-menten model: routine clinical pharmacokinetic data. *Journal of pharmacokinetics and biopharmaceutics*, 8(6):553–571, 1980.

Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. Hierarchical geodesic models in diffeomorphisms. *International Journal of Computer Vision*, 117(1):70–92, 2016.

Daouda Sissoko, Cedric Laouenan, Elin Folkesson, Abdoul-Bing M'lebing, Abdoul-Habib Beavogui, Sylvain Baize, Alseny-Modet Camara, Piet Maes, Susan Shepherd, Christine Danel, et al. Experimental treatment with favipiravir for ebola virus disease (the jiki trial): a historically controlled, single-arm proof-of-concept trial in guinea. *PLoS medicine*, 13(3):e1001967, 2016.

James C Spall et al. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

Anuj Srivastava, Shantanu H. Joshi, Washington Mio, and Xiuwen Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):590–602, apr 2005. ISSN 01628828. doi: $10.1109/\mathrm{TPAMI}.2005.86$.

Jingyong Su, Sebastian Kurtek, Eric Klassen, Anuj Srivastava, et al. Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552, 2014.

Patrick Therasse, Susan G Arbuck, Elizabeth A Eisenhauer, Jantien Wanders, Richard S Kaplan, Larry Rubinstein, Jaap Verweij, Martine Van Glabbeke, Allan T van Oosterom, Michaele C Christian, et al. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216, 2000.

D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley„ 1985.

Alain Trouvé. An infinite dimensional group approach for physics based models in pattern recognition. *preprint*, 1995.

Paul Tseng. An analysis of the em algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 381–392. Springer, 2005.

Tom Vercauteren, Aymeric Perchant, Xavier Pennec, and Nicholas Ayache. Mosaicing of confocal microscopic in vivo soft tissue video sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 753–760. Springer, 2005.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*, 45(1 Suppl), 2009. ISSN 10959572. doi: $10.1016/\mathrm{j.neuroimage}.2008.10.040$.

Jing Wang. Em algorithms for nonlinear mixed effects models. *Computational statistics & data analysis*, 51(6):3244–3256, 2007.

Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85 (411):699–704, 1990.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

Chao Yang. Recurrent and ergodic properties of adaptive mcmc. *Preprint*, 2008.

L Yin, X Chenand Y Sun, T Worm, and M Reale. A high-resolution 3d dynamic facial expression database, 2008. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, volume 126.

Suhang You, Kerem C Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pages 540–556. PMLR, 2019.

Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer, 2010.

Qian Yu, Yinghuan Shi, Jinquan Sun, Yang Gao, Jianbing Zhu, and Yakang Dai. Crossbar-net: A novel convolutional neural network for kidney tumor segmentation in ct images. *IEEE transactions on image processing*, 28(8):4060–4074, 2019.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre:** Modélisation statistique de données médicales et analyse théorique des algorithmes d'estimation

**Mots clés:** Modélisation statistique, Algorithmes stochastiques, Données médicales, Variétés Riemaniennes

**Résumé:** Dans le domaine médicale, l'usage de caractéristiques extraites d'images est de plus en plus répandu. Ces mesures peuvent être des nombres réels (volume, score cognitifs), des maillages d'organes ou l'image elle-même. Dans ces deux derniers cas, un espace Euclidien ne peut décrire l'espace de mesures et il est nécessaire de se placer sur une variété Riemanienne. En utilisant ce cadre Riemannien et des modèles à effets mixtes, il est alors possible d'estimer un objet représentatif de la population ainsi que la variabilité inter-individuelle.
Dans le cas longitudinal (sujets observés de manière répétée au cours du temps), ces modèles permettent de créer une trajectoire moyenne représentative de l'évolution globale de la population. Dans cette thèse, nous proposons de généraliser ces modèles dans le cas d'un mélange de population. Chaque sous-population peut suivre différentes dynamiques au cours du temps et leur trajectoire représentative peut être la même ou différer d'un intervalle temporel à l'autre. Ce nouveau modèle permet par exemple de modéliser l'apparition d'une maladie comme une déviation par rapport à un vieillissement normal.
Nous nous intéressons également à la détection d'anomalies (par exemple de tumeurs) dans une population. En disposant d'un objet représentant une population contrôle, nous définissons une anomalie comme ce qui ne peut être reconstruit par déformation difféomorphique de cet objet représentatif. Notre méthode à l'avantage de ne nécessiter ni grand jeu de donnée, ni annotation par des médecins et peut être facilement appliquée à tout organe.
Finalement, nous nous intéressons à différentes propriétés théoriques des algorithmes d'estimation utilisés. Dans le cadre des modèles à effets mixtes non linéaires, l'algorithme MCMC-SAEM est utilisé. Nous discuterons de deux limitations théoriques. Premièrement, nous lèverons l'hypothèse d'ergodicité géométrique en la remplaçant par une hypothèse d'ergodicité sous-géométrique. De plus, nous nous intéresserons à une méthode permettant d'appliquer l'algorithme SAEM quand la distribution jointe n'est pas courbe exponentielle. Nous montrerons que cette méthode introduit un biais dans l'estimation que nous mesurerons. Nous proposerons également un nouvel algorithme permettant de le réduire.

**Title:** Statistical modelisation of medical data and theoretical analysis of estimation algorithms

**Keywords:** Statistical models, Stochastic algorithms, Medical data, Riemannian manifolds

**Abstract:** In the medical field, the use of features extracted from images is increasingly common to perform diagnostics or measure the effectiveness of a treatment over time. These measures can for example be real numbers (volume, cognitive scores), meshes of an organ or even the image itself. In the latter two cases, a Euclidean space cannot describe the space of measurements and it is necessary to use Riemannian manifolds. Using this Riemannian framework and mixed effects models, it is then possible to estimate a representative object of the population as well as the inter-individual variability.
In the longitudinal case (subjects observed repeatedly over time), these models allow to create an average trajectory, representative of the global evolution of the population. In this thesis, we propose to generalize these models in the case of a mixture of populations. Each sub-population can follow different dynamics over time and their representative trajectory can branch or join from one time interval to another. This new model allows, for example, to model the onset of a disease as a deviation from a normal aging.
In a second step, we are also interested in the detection of anomalies (e.g. tumours) in a population. Given an object representing a control population, we define an anomaly as a structure that cannot be reconstructed by a diffeomorphic deformation of this representative object. Our method has the advantage of requiring neither a large data set nor annotation by physicians. Moreover, it can be easily applied to any organ.
Finally, we are interested in different theoretical properties of the previously used estimation algorithms. In the context of non-linear mixed effects models, the MCMC-SAEM algorithm is used. In this thesis, we will discuss two theoretical limitations. Firstly, we will lift the geometric ergodicity assumption by replacing it with a sub-geometric ergodicity assumption. Furthermore, we will look at a method, often used in practice, allowing to apply the SAEM algorithm when the joint distribution is not exponentially curved. We will show that this method introduces a bias in the estimation that we will measure. We will also propose a new algorithm to reduce it.

**Institut Polytechnique de Paris**
91120 Palaiseau, France